TARGET DRIFT IN MULTI-CONSTRAINT LAGRANGIAN RL: THEORY AND PRACTICE

Anonymous authors
Paper under double-blind review

ABSTRACT

Lagrangian-based methods are one of the dominant approaches for safe reinforcement learning (RL) in constrained Markov decision processes, commonly used across domains with multiple constraints. While some implementations combine all constraints into a mixed penalty term and others use one estimator per constraint, the fundamental question of which design is theoretically sound has received little scrutiny. We provide the first theoretical analysis showing that the mixed-critic architecture induces a persistent bias due to target drift from evolving Lagrange multipliers. In contrast, dedicated-critic design—separate critics for reward and each constraint—avoids this issue. We also validate our findings in a simulated but realistic power system with multiple physical constraints, where the dedicated-critic method achieves stable learning and consistent constraint satisfaction, while the mixed-critic method fails. Our results offer a principled argument for preferring dedicated-critic architectures in multi-constraint safe RL problems.

1 INTRODUCTION

Safe reinforcement learning (RL) in constrained Markov decision processes (CMDPs) (Altman, 1999) has become increasingly important in real-world applications such as robotics, power systems, autonomous driving, and healthcare (Yan and Xu, 2020; Wang et al., 2020; Calascibetta et al., 2023; Zhang et al., 2020; Shi et al., 2023). Among the most widely adopted frameworks for handling such problems are Lagrangian-based methods, which introduce Lagrange multipliers for constraints and optimize a mixed/augmented objective (Achiam et al., 2017; Ray et al., 2019b). This approach offers appealing theoretical properties: it transforms a constrained problem into an unconstrained one, allowing the use of powerful policy gradient and actor-critic techniques, while enabling principled constraint enforcement via dual variable updates. Theoretically, under suitable assumptions, this leads to saddle-point solutions that jointly maximize reward and satisfy constraints, making it both elegant and scalable for complex, high-dimensional systems (Achiam et al., 2017).

One important but often overlooked reality is that real-world CMDPs rarely involve a single constraint. Instead, agents are typically required to satisfy multiple, interacting safety, resource, or operational constraints during both training and deployment. For instance, robotic systems must avoid unsafe behaviors while simultaneously respecting torque and energy limitations (Liu et al., 2022; Junges et al., 2016); autonomous driving agents must account for safety margins, passenger comfort, and compliance with traffic laws (Zhang et al., 2023; 2021); and power systems must balance supply and demand while maintaining safe voltage and capacity constraints (Wu et al., 2023; Chen et al., 2022). These constraints are rarely independent and often conflict, making multi-constraint settings the norm rather than the exception.

A critical but insufficiently studied aspect of Lagrangian safe RL is the value-critic architecture for multi-constraint problems. Although the CMDP formalism and constraint-aware algorithms such as CPO define per-constraint(dedicated) quantities for policy updates (Achiam et al., 2017), there remains *no theoretical justification* in the literature for why to use this approach in practice. On the other hand, most widely used implementations, including PPO-/TRPO-Lagrangian baselines (Ray et al., 2019b; Stooke et al., 2020; Yang et al., 2020; Bhatnagar et al., 2009; Kim et al., 2023), implicitly collapse all constraints into a single mixed penalty term and estimate it using one cost critic. While simple and computationally efficient, this design sidesteps the unique challenges posed by multi-constraint CMDPs. As shown in Figure 1, practitioners tackling real-world problems are left without clear criteria for choosing between approaches. However, the theoretical validation of these approaches remains unproven, and the optimization bias may induce remains underexplored. This issue is not merely theoretical. Empirical evidence shows that PPO-Lagrangian, despite its widespread use, suffers from instability and inconsistent performance in multi-constraint settings (Stooke et al., 2020; Tessler et al., 2019). These observations further motivate a deeper theoretical analysis of critic design and support our proposal for a dedicated-critic framework as a necessary advancement for stable and scalable safe RL under multiple constraints.

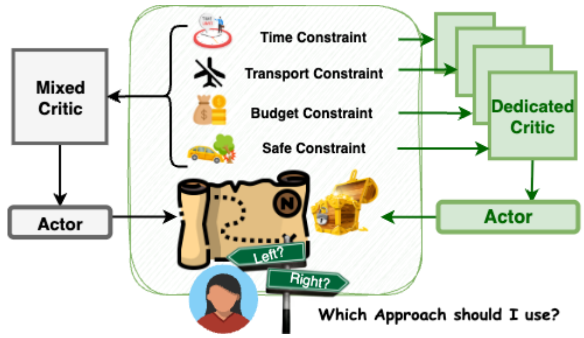


Figure 1: Difference between the two approaches.

Motivated by the widespread use of mixed-critic architectures in safe RL and the increasing demand for multi-constraint decision-making in real-world applications, we aim to close a crucial theoretical gap in constrained reinforcement learning. In this work, we provide the first formal analysis of mixed- versus dedicated-critic designs in Lagrangian-based constrained RL. We show that training a mixed critic on multi-constraint signals introduces a structural bias in the actor update. Specifically, as the Lagrange multipliers evolve during training, the critic's target drifts in a way that violates the stationarity assumption required by temporal-difference learning. This leads to a persistent error in the estimated policy gradient. We prove that the dedicated-critic design, training separate critics for the reward and each constraint signal, eliminates the dual-driven drift altogether. To validate our theoretical results in practice, we implement both mixed- and dedicated-critic methods in a constrained bandit problem and a constrained energy control problem with multiple physical limits. The experiments reveal that the mixed-critic approach frequently violates constraints, whereas the dedicated-critic design achieves stable learning and consistently satisfies all constraints. This work makes three main contributions: 1) We provide the first formal analysis of mixed- vs. dedicated-critic designs in Lagrangian safe RL, showing that mixed critics suffer from dual-induced bias. 2) We prove that dedicated critics yield stationary targets, eliminating drift and enabling stable policy gradient estimation. 3) We validate our theory in constrained MDP tasks and complex energy control task, where dedicated critics achieve stable learning and consistent constraint satisfaction.

2 Related Work

Safe RL aims to train agents that not only maximize long-term performance but also respect safety or risk-related constraints during learning and deployment. This is often formalized through the framework of *Constrained Markov Decision Processes* (CMDPs), where the objective is to maximize expected return while ensuring that expected costs, representing safety violations or resource usage, remain below specified thresholds. (Altman, 1999; Garcıa and Fernández, 2015). This formulation admits a primal–dual view in which constraints are handled by Lagrange multipliers, giving rise to the widely used *Lagrangian (lag-based)* methods: they update policy parameters to ascend a Lagrangian objective and update dual variables toward feasibility. Prominent examples include TRPO-Lagrangian and PPO-Lagrangian (Achiam et al., 2017; Ray et al., 2019b), SAC-Lagrangian variants (Ray et al., 2019b), and Reward-Constrained Policy Optimization (RCPO) (Tessler et al., 2019).

Compared to alternative approaches(Stooke et al., 2020; Liu et al., 2020; Xu et al., 2021; Chow et al., 2018), Lagrangian methods offer several practical advantages that have led to their widespread adoption (Achiam et al., 2017; Schulman et al., 2015; Ray et al., 2019b; Yang et al., 2021a; Kim et al., 2023). They are *plug-and-play* compatible with both on-policy and off-policy learners, and introduce only a small number of hyperparameter. These properties make them highly amenable to integration within standard RL pipelines. Consequently, Lagrangian variants like PPO-Lag and TRPO-Lag have become *de facto baselines* in major Safe RL benchmarks and toolkits. Their accessibility, combined with consistently strong empirical performance, has made them the dominant choice in both robotics and simulated safety-critical control environments.

A critical yet underexplored aspect of Lagrangian based safe reinforcement learning is the architecture of value critics when dealing with multiple constraints. The question of how to estimate constraints returns in deep RL has received far less attention than objective or dual design, but it underlies much of the instability noted in safe RL. In early safe RL and constrained MDP work, classic algorithms (e.g., Constrained Policy Iteration) implicitly worked with per-constraint value functions, but without deeply discussing representation in function approximation settings (Altman, 1999; Achiam et al., 2017). As deep safe RL matured, many practical baselines resorted to collapsing cost signals into an aggregated penalty and training a mixed "cost critic" alongside a reward critic; this pattern is pervasive in benchmark codebases (e.g. Safety Starter Agents, PPO-/TRPO-Lagrangian) (Ray et al., 2019b; Stooke et al., 2020). Some recent methods extend to multiple constraints, but often leave the critic architecture unspecified or adopt ad-hoc shared representations rather than formally treating per-constraint estimation (Kim et al., 2023). Work on stabilizing Lagrangian dual updates—such as PID-Lagrangian, dual clipping, or adaptive multiplier heuristics—addresses the dual dynamics but typically retains the standard two-critic collapse architecture (Stooke et al., 2020; Xu et al., 2021; Liu et al., 2020). In off-policy safe RL methods like SAC-Lagrangian variants or worst-case safety critics (e.g. WCSAC), the separation between reward and cost critics is common, but again usually implemented at the aggregate cost level even when multiple constraints are present (Yang et al., 2021b; Tessler et al., 2019). Across the literature, the critic architecture—whether to collapse or separate constraints—is treated as an afterthought, often chosen for ease or efficiency rather than guided by theoretical insight. This pervasive gap means that many empirical instability observations, violation spikes, slow convergence remain underexplained—pointing to a need for more rigorous analysis of critic structure in multi-constraint safe RL. In this work, we fill this gap by theoretically and empirically analyzing these design choices. Our analysis shows that mixed constraint critics can introduce structural bias in multi-constraint settings, whereas dedicated critics mitigate this bias by isolating constraint signals.

3 Problem Formulation

We begin by formalizing the constrained reinforcement learning problem with multiple constraints. A discounted Constrained Markov Decision Process (CMDP)(Altman, 1999) is specified by the tuple $(\mathcal{S},\ \mathcal{A},\ P,\ \gamma,\ r,\ \{c_i\}_{i=1}^m)$, where: $\mathcal{S}$ is the (possibly infinite) state space; $\mathcal{A}$ is the action space; $P(\cdot|s,a)$ is the transition kernel governing state evolution; $\gamma \in (0,1)$ is the discount factor; $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward signal we aim to maximize; $c_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ are cost signals corresponding to m safety or resource constraints.

For a stochastic policy $\pi_\theta(a|s)$ parameterized by θ, the expected discounted return of a signal $x \in \{r, c_1, \ldots, c_m\}$ is

$$J_x(\pi_\theta) \ = \ \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t\, x(s_t, a_t)\right]. \tag{1}$$

In particular, $J_r(\pi_\theta)$ is the expected reward return, while $J_{c_i}(\pi_\theta)$ is the expected discounted cost associated with constraint i. The returns in equation 1 can be characterized via value functions.

Constraints of the form $J_{c_i}(\pi_\theta) \leq d_i$ can be enforced via a Lagrangian formulation. Introducing multipliers $\lambda = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}_+^m$, we define $\mathcal{L}(\theta, \lambda) \ = \ J_r(\pi_\theta) - \sum_{i=1}^m \lambda_i \big(J_{c_i}(\pi_\theta) - d_i\big)$. Optimization then proceeds in a *primal–dual* fashion: the actor seeks to maximize $\mathcal{L}(\theta, \lambda)$ over θ, while the dual variables λ adaptively adjust to enforce the constraints. Let $\mathcal{E}^x_{\pi_\theta}(\cdot;\omega)$ denote a learned *signal estimator* for $x \in \{r, c_1, \ldots, c_m\}$ parameterized by ω. Its input may be s or (s,a) depending on the method (e.g., value-, advantage-, or return-based); the analysis does not depend on this choice.

By the policy gradient theorem, the gradient of the Lagrangian w.r.t. θ is

$$\nabla_\theta \mathcal{L}(\theta, \lambda) = \mathbb{E}_{s,a\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(a|s) \left(\mathcal{E}^r_{\pi_\theta}(s,a) \ - \ \sum_{i=1}^m \lambda_i\, \mathcal{E}^{c_i}_{\pi_\theta}(s,a)\right)\right], \tag{2}$$

where $\mathcal{E}^r_{\pi_\theta}$ and $\mathcal{E}^{c_i}_{\pi_\theta}$ denote the learned signal estimators for reward and each cost under π_θ. Note that λ merely scales the $\mathcal{E}^{c_i}_{\pi_\theta}$ contributions; the estimators themselves depend only on π_θ.

Mixed-critic (Classic) methods. Constraints are *aggregated* before (or within) estimation, so there is no per-constraint head. Two common variants both qualify as "mixed critic":

$$\textit{(a) Single-Estimator:} \quad \mathcal{E}^{\text{mix}}_{\pi_\theta}(\cdot;\omega) \;\approx\; \mathcal{E}^{r-\sum_{i=1}^m \lambda_i c_i}_{\pi_\theta}(\cdot).$$

$$\textit{(b) Two-Estimator:} \quad \mathcal{E}^{r}_{\pi_\theta}(\cdot;\omega^r), \qquad \mathcal{E}^{\text{cost-agg}}_{\pi_\theta}(\cdot;\omega^c) \;\approx\; \mathcal{E}^{\sum_{i=1}^m \lambda_i c_i}_{\pi_\theta}(\cdot).$$

In both (a) (Altman, 1999) and (b) (Ray et al., 2019b; Stooke and Abbeel, 2020), all constraints are *mixed into a single scalar cost signal*, hence "mixed critic."

Dedicated-critic methods. Maintain one estimator *per signal*, i.e., a *Separate Estimator* for reward and for *each* constraint (Achiam et al., 2017):

$$\big\{\, \mathcal{E}^{x}_{\pi_\theta}(\cdot;\omega^x) \;:\; x \in \{r, c_1, \ldots, c_m\} \big\}.$$

4 Theoretical Analysis: Why we need Dedicated Critics for Multi-Constraint Problems

In this section, we will systematically show that training a *mixed critic* on the multiple signal generally yields actor updates that *do not track* the true Lagrangian gradient $\nabla_\theta \mathcal{L}(\theta, \lambda)$ during learning, unless one imposes stronger timescale separation (critic faster than *both* actor and dual) and near-exact critics. In contrast, a *dedicated-critic* design (one critic per signal) does not suffer from this issue. Our argument is constructive and quantitative.

4.1 Setting

Assumption 4.1 (Stepsizes and timescale separation)**.** Critic, actor, and dual stepsizes $\eta_t, \alpha_t, \beta_t > 0$ satisfy $\sum_{t=0}^\infty \eta_t = \infty$, $\sum_{t=0}^\infty \eta_t^2 < \infty$. $\frac{\alpha_t}{\eta_t} \to 0$, $\frac{\beta_t}{\eta_t} \to 0$.

Assumption 4.2 (Bounded policy score and compact dual domain)**.** At iteration t, states/actions (s_t, a_t) are sampled on-policy under π_{θ_t}. There exists $G < \infty$ such that the log-policy score is uniformly bounded almost surely: $\big\|\nabla_\theta \log \pi_{\theta_t}(a_t \mid s_t)\big\| \le G$. The dual variable sequence $\{\lambda_t\}_{t\ge 0} \subset \mathbb{R}^m_{\ge 0}$ remains in a fixed compact set $\Lambda \subset \mathbb{R}^m_{\ge 0}$ (e.g., via projected updates onto Λ).

Assumption 4.3 (Critic noise regularity)**.** Let the critic update use the population linear form with additive martingale-difference noise: $\omega_{t+1} = \omega_t + \eta_t\big(b_t - A_t\omega_t\big) + \eta_t\,\zeta_{t+1}$, where $A_t := A(\theta_t)$ and $b_t := b(\theta_t)$ are $\mathcal{F}_t$-measurable. The noise $\{\zeta_{t+1}\}_{t\ge 0}$ satisfies $\mathbb{E}[\zeta_{t+1} \mid \mathcal{F}_t] = 0$ and $\mathbb{E}\big[\|\zeta_{t+1}\|^2 \mid \mathcal{F}_t\big] \le \sigma^2 < \infty$ a.s. for all t.

Assumption 4.1 matches standard stochastic-approximation practice: the critic uses diminishing stepsizes and runs faster than the actor and dual. Assumption 4.2 is routine for common policies (softmax, Gaussian with clipped parameters) and for lag-based methods that project/clip λ_t onto a compact box Λ. Assumption 4.3 follows from on-policy sampling with bounded features/signals and mini-batch estimates, which yield martingale-difference noise with bounded conditional variance. These mild conditions are typical in deep RL and suffice to ensure critic contraction and to isolate the dual-induced drift term that motivates dedicated per-signal critics.

For a fixed policy π_θ and signal $x \in \{r, c_1, \ldots, c_m\}$, the per-signal critic in a linear class (Sutton, 1988; Tsitsiklis and Van Roy, 1996a) satisfies the projected Bellman equation (PBE) (Munos, 2003; Tsitsiklis and Van Roy, 1996b), which yields the normal equations $A(\theta)\,\omega^{x,*}(\theta) = b^x(\theta)$ with $A(\theta) = \Phi^\top D_\theta (I - \gamma P_{\pi_\theta})\Phi$ and $b^x(\theta) = \Phi^\top D_\theta r^x$ (state–action and advantage/GAE variants give the same linear template with the appropriate A, b). We work under $A(\theta) \succeq \mu I$ for some $\mu > 0$, and $A(\cdot)$, $b^x(\cdot)$ are locally Lipschitz in θ (See Appendix B and F for details).

4.2 Mixed-Critic in Multi-Constraint CMDPs

When a mixed critic is used for the scalarized signal, $r_{\lambda_t} = r - \sum_{i=1}^m \lambda_{t,i}\, c_i$, The PBE is $A(\theta_t)\,\omega_t^{\text{mix},*} = b_t^{\text{mix}} := b^r(\theta_t) - \sum_{i=1}^m \lambda_{t,i}\, b^{c_i}(\theta_t)$. The stochastic update implements a Robbins–Monro step toward this fixed point using mini-batch estimates of $A(\theta_t)$ and b_t^{mix}. Writing the update in *population form* plus a mean-zero error gives

$$\omega^{\text{mix}}_{t+1} = \omega^{\text{mix}}_t + \eta_t\big(b^{\text{mix}}_t - A(\theta_t)\,\omega^{\text{mix}}_t\big) + \eta_t\,\zeta_{t+1}, \tag{3}$$

where $\eta_t > 0$ is the critic stepsize and ζ_{t+1} is a martingale-difference noise capturing finite-sample and sampling variability. For a fixed (θ, λ), the PBE for the mixed signal is $A(\theta)\,\omega^{\text{mix}}(\theta, \lambda) = b^r(\theta) - \sum_{i=1}^m \lambda_i b^{c_i}(\theta)$. By linearity of the operator, the solution decomposes as $\omega^{\text{mix}}(\theta, \lambda) = \omega^r(\theta) - \sum_{i=1}^m \lambda_i\, \omega^{c_i}(\theta)$, where $\omega^r(\theta)$ and $\omega^{c_i}(\theta)$ are the PBE solutions for the reward and each cost signal individually. Let us denote the instantaneous target at time t by $\omega_t^{\text{mix},\star} = \omega^{\text{mix}}(\theta_t, \lambda_t)$, $e_t = \omega_t^{\text{mix}} - \omega_t^{\text{mix},\star}$. That is, e_t is the critic error relative to the exact PBE solution for the current (θ_t, λ_t). Subtracting $\omega_{t+1}^{\text{mix},\star}$ from both sides of the recursion equation 3 yields the exact error recursion

$$e_{t+1} = (I - \eta_t A(\theta_t))\, e_t + \underbrace{(\omega_t^{\text{mix},\star} - \omega_{t+1}^{\text{mix},\star})}_{\text{target drift}} + \eta_t \zeta_{t+1} + \Delta_t^\theta, \tag{4}$$

where Δ_t^θ collects the small changes in $A(\theta)$, $I \in \mathbb{R}^{d\times d}$ denote the $d \times d$ identity matrix, and $b(\theta)$ induced by $\theta_{t+1} \neq \theta_t$. Using the stationary equivalence $\omega_t^{\text{mix},\star} = \omega^r(\theta_t) - \sum_{i=1}^m \lambda_{t,i}\, \omega^{c_i}(\theta_t)$ and the analogous expression at time $t{+}1$, we have

$$\begin{aligned}\omega_t^{\text{mix},\star} - \omega_{t+1}^{\text{mix},\star} &= \Big[\omega^r(\theta_t) - \sum_{i=1}^m \lambda_{t,i}\, \omega^{c_i}(\theta_t)\Big] \;-\; \Big[\omega^r(\theta_{t+1}) - \sum_{i=1}^m \lambda_{t+1,i}\, \omega^{c_i}(\theta_{t+1})\Big] \\ &= \Big(\omega^r(\theta_t) - \omega^r(\theta_{t+1})\Big) \;-\; \sum_{i=1}^m \Big(\lambda_{t,i}\, \omega^{c_i}(\theta_t) - \lambda_{t+1,i}\, \omega^{c_i}(\theta_{t+1})\Big).\end{aligned} \tag{5}$$

Add and subtract $\lambda_{t,i}\, \omega^{c_i}(\theta_{t+1})$ inside the sum:

$$\lambda_{t,i}\, \omega^{c_i}(\theta_t) - \lambda_{t+1,i}\, \omega^{c_i}(\theta_{t+1}) = \lambda_{t,i}\big(\omega^{c_i}(\theta_t) - \omega^{c_i}(\theta_{t+1})\big) \;+\; (\lambda_{t,i} - \lambda_{t+1,i})\, \omega^{c_i}(\theta_{t+1}). \tag{6}$$

Substituting equation 6 into equation 5 yields

$$\begin{aligned}\omega_t^{\text{mix},\star} - \omega_{t+1}^{\text{mix},\star} = \Big(\omega^r(\theta_t) - \omega^r(\theta_{t+1})\Big) - \sum_{i=1}^m \lambda_{t,i}\Big(\omega^{c_i}(\theta_t) - \omega^{c_i}(\theta_{t+1})\Big) \\ - \sum_{i=1}^m \big(\lambda_{t+1,i} - \lambda_{t,i}\big)\, \omega^{c_i}(\theta_{t+1}).\end{aligned} \tag{7}$$

By Lipschitz continuity, there exist $L_r, L_{c_i} < \infty$ such that $\|\omega^r(\theta_{t+1}) - \omega^r(\theta_t)\| \le L_r \|\theta_{t+1} - \theta_t\|$, $\|\omega^{c_i}(\theta_{t+1}) - \omega^{c_i}(\theta_t)\| \le L_{c_i}\|\theta_{t+1} - \theta_t\|$. With a standard actor update $\theta_{t+1} = \theta_t + \alpha_t\, g_t^{\text{act}}$, where g_t^{act} is a stochastic policy-gradient estimate and we assume $\|g_t^{\text{act}}\| \le C_\theta$, it follows that $\|\theta_{t+1} - \theta_t\| = \alpha_t \|g_t^{\text{act}}\| = O(\alpha_t)$. Therefore,

$$\|\omega^r(\theta_t) - \omega^r(\theta_{t+1})\| = O(\alpha_t), \Big\|\lambda_{t,i}\big(\omega^{c_i}(\theta_t) - \omega^{c_i}(\theta_{t+1})\big)\Big\| \le \|\lambda_t\|_\infty\, L_{c_i} \|\theta_{t+1} - \theta_t\| = O(\alpha_t),$$

using $\lambda_t \in \Lambda$ compact. Thus the first two terms in equation 7 are $O(\alpha_t)$.

Consider the dual-induced part of equation 7: $\sum_{i=1}^m \big(\lambda_{t+1,i} - \lambda_{t,i}\big)\, \omega^{c_i}(\theta_{t+1})$. By Assumption 4.2, $\lambda_t \in \Lambda$ with Λ compact, and local Lipschitzness, the map $\theta \mapsto \omega^{c_i}(\theta)$ is continuous; hence $M := \sup_{i,\theta} \|\omega^{c_i}(\theta)\| < \infty$ (on the on-policy region visited by $\{\theta_t\}$). Therefore

$$\Big\|\sum_{i=1}^m (\lambda_{t+1,i} - \lambda_{t,i})\, \omega^{c_i}(\theta_{t+1})\Big\| \;\le\; \|\lambda_{t+1} - \lambda_t\|_1 \, \max_i \|\omega^{c_i}(\theta_{t+1})\| \;\le\; M\, \|\lambda_{t+1} - \lambda_t\|.$$

For a standard projected dual update, $\lambda_{t+1} = \Pi_\Lambda\big(\lambda_t + \beta_t\, g_t\big)$, with Π_Λ nonexpansive and $\|g_t\| \le C_\lambda$ ($g_t \in \mathbb{R}^m$ denotes a subgradient of the dual objective with respect to λ), we have

$$\|\lambda_{t+1} - \lambda_t\| \;\le\; \big\|\lambda_t + \beta_t g_t - \lambda_t\big\| \;=\; \beta_t\, \|g_t\| \;\le\; \beta_t C_\lambda \;=\; O(\beta_t).$$

Combining the two displays yields

$$\Big\|\sum_{i=1}^m (\lambda_{t+1,i} - \lambda_{t,i})\, \omega^{c_i}(\theta_{t+1})\Big\| \;\le\; M C_\lambda\, \beta_t \;=\; O(\beta_t).$$

The target drift decomposes as $\omega_t^\star - \omega_{t+1}^\star = O(\alpha_t) \;-\; O(\beta_t)$. The $O(\alpha_t)$ term arises from policy updates (present in any actor–critic), with the additional $O(\beta_t)$ term. In this case, the target moves not only because the policy parameters θ evolve (the standard $O(\alpha_t)$ policy-driven drift), but also because the dual variables λ evolve, producing an additional $O(\beta_t)$ dual-driven term. This extra dual-driven drift induces persistent bias in the actor gradient.

Remark 4.4 (Drift for the reward critic + cost critic design). In the two-critic setup, one critic estimates the reward target $\omega^r(\theta)$ and a second critic estimates the aggregated cost target $\omega^{\text{cost}}(\theta, \lambda) = \sum_{i=1}^m \lambda_i \, \omega^{c_i}(\theta)$,(used in PPO-LAG, TRPO-LAG) follow the same process on ω^{cost}, we have

$$\omega_t^{\text{cost},\star} - \omega_{t+1}^{\text{cost},\star} = -\sum_{i=1}^{m} \lambda_{t,i}\Big(\omega^{c_i}(\theta_t) - \omega^{c_i}(\theta_{t+1})\Big) \; - \; \sum_{i=1}^{m} \big(\lambda_{t+1,i} - \lambda_{t,i}\big)\, \omega^{c_i}(\theta_{t+1}). \tag{8}$$

Here the reward term $\omega^r(\theta_t) - \omega^r(\theta_{t+1})$ does not appear because it is handled by the *reward critic's* own recursion; equation 8 isolates the drift of the *aggregated-cost* head, which contains a policy-induced component (through $\theta_{t+1} - \theta_t$) and a dual-induced component (through $\lambda_{t+1} - \lambda_t$).

Lemma 4.5 (Mixed-critic error bound). *If $A(\theta) \succeq \mu I$ uniformly and the stepsizes satisfy $\alpha_t/\eta_t \to 0$, $\beta_t/\eta_t \to 0$, then there exist constants C_λ, C_θ, such that:*

$$\limsup_{t\to\infty} \mathbb{E}\|e_t\| \;\le\; \frac{C_\lambda}{\mu} \limsup_{t\to\infty} \frac{\beta_t}{\eta_t} \;+\; \frac{C_\theta}{\mu} \limsup_{t\to\infty} \frac{\alpha_t}{\eta_t} \;+\; O(1). \tag{9}$$

Proof sketch. From equation 4 and since $A(\theta_t) \succeq \mu I$, $\|(I - \eta_t A(\theta_t))e_t\| \le (1 - \mu\eta_t)\|e_t\|$, using equation 7, we have:

$$\|\omega_t^\star - \omega_{t+1}^\star\| \;\le\; C_\lambda \,\|\lambda_{t+1} - \lambda_t\| + C_\theta \,\|\theta_{t+1} - \theta_t\| \;\le\; C_\lambda \,\beta_t + C_\theta \,\alpha_t.$$

Taking expectations and using bounded MDS noise and apply the standard SA comparison: if $x_{t+1} \le (1 - a_t)x_t + b_t$ with $a_t = \mu\eta_t$, then $\limsup x_t \le \limsup b_t/a_t$. Hence,

$$\limsup_{t\to\infty} \mathbb{E}\|e_t\| \;\le\; \frac{1}{\mu} \limsup_{t\to\infty} \Big(\tfrac{C_\lambda \beta_t}{\eta_t} + \tfrac{C_\theta \alpha_t}{\eta_t} + O(\eta_t)\Big),$$

which yields equation 9 since $\eta_t \to 0$. Please refer to Appendix C for detailed proof. □

Lemma 4.6 (Actor-gradient bias bound). *Let $\widehat{g}_t$ and $g_t^\star$ be the actor's estimated and ideal gradients,*

$$\widehat{g}_t = \mathbb{E}_t\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\, \phi(s_t, a_t)^\top \omega_t\big], \qquad g_t^\star = \mathbb{E}_t\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\, \phi(s_t, a_t)^\top \omega_t^\star\big], \tag{10}$$

with critic error $e_t = \omega_t - \omega_t^\star$. Assume the score and features are bounded as $\|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\| \le G$, $\|\phi(s_t, a_t)\| \le L_\phi$. Then the actor-gradient bias $B_t := \widehat{g}_t - g_t^\star$ satisfies

$$\|B_t\| \;\le\; G L_\phi \,\|e_t\|. \tag{11}$$

Proof. See Appendix D for detailed proof. □

Theorem 4.7 (Bias from a Mixed Critic). *Suppose Assumptions 4.1–4.3 hold and $A(\theta) \succeq \mu I$ uniformly in θ. Then the actor-gradient bias B_t incurred by using a* single mixed critic *satisfies*

$$\limsup_{t\to\infty} \mathbb{E}\|B_t\| \;\le\; G L_\phi \left(\frac{C_\lambda}{\mu} \limsup_{t\to\infty} \frac{\beta_t}{\eta_t} \;+\; \frac{C_\theta}{\mu} \limsup_{t\to\infty} \frac{\alpha_t}{\eta_t} \right). \tag{12}$$

Proof. See Appendix E for detailed proof. □

Mixed-critic design introduces an **additional bias term** of order β_t/η_t, arising from the dependence of the mixed constraints on the dual variables λ. Consequently, the actor's update does not follow the true Lagrangian gradient unless the critic runs much faster than the dual ($\beta_t/\eta_t \to 0$ sufficiently quickly), or is essentially exact.

4.3 DEDICATED-CRITIC IN MULTI-CONSTRAINT CMDPS

For dedicated-critic design, we maintain a separate critic with parameters ω_t^x for each reward and constraints $x \in \{r, c_1, \ldots, c_m\}$, updated by

$$\omega_{t+1}^x \;=\; \omega_t^x + \eta_t \Big(- A(\theta_t)\, \omega_t^x + b^x(\theta_t) + \zeta_{t+1}^x \Big), \qquad x \in \{r, c_1, \ldots, c_m\}. \tag{13}$$

Define the signal-specific fixed point $\omega^{x,\star}(\theta)$ by $A(\theta)\,\omega^{x,\star}(\theta) = b^x(\theta)$ and let the tracking error be $e_t^x := \omega_t^x - \omega_t^{x,\star}$. Subtract $\omega_{t+1}^{x,\star}(\theta_{t+1})$ to equation 13 to obtain

$$\begin{aligned} e_{t+1}^x = \omega_{t+1}^x - \omega_{t+1}^{x,\star}(\theta_{t+1}) &= \Big(\omega_t^x - \eta_t A(\theta_t)\omega_t^x + \eta_t b^x(\theta_t) + \eta_t \zeta_{t+1}^x\Big) - \omega_{t+1}^{x,\star}(\theta_{t+1}) \\ &= \Big(I - \eta_t A(\theta_t)\Big) e_t^x \;+\; \eta_t \zeta_{t+1}^x \;+\; \Delta_t^{\theta,x}. \end{aligned}$$

A dedicated critic answers a fixed question: under the *current policy*, what is the expected cumulative value of one signal—either reward or a single cost? Because that question does not mention the penalty weights, changing λ does not change what the critic is trying to predict; only changing the policy does. To be more specific, each dedicated critic estimates a *signal-specific* fixed point $\omega^{x,\star}(\theta)$ defined by $A(\theta)\,\omega^{x,\star}(\theta) = b^x(\theta)$, where both $A(\theta)$ and $b^x(\theta)$ depend on the policy θ and the single signal $x \in \{r, c_1, \ldots, c_m\}$, but not on the Lagrange multipliers; Therefore, in the one-step error recursion the only drift term comes from policy movement $\theta_t \to \theta_{t+1}$, and no $(\lambda_{t+1} - \lambda_t)$ term appears. By contrast, a mixed-critic's target *is* defined using λ (it blends reward and costs with those weights), so every time λ is updated the target itself shifts, creating the extra $(\lambda_{t+1} - \lambda_t)$ drift.

Lemma 4.8 (Dedicated-critic tracking error)**.** *Suppose assumptions 4.1 and 4.3 hold, then there exists $\widetilde{C}_\theta < \infty$ such that for every $x \in \{r, c_1, \ldots, c_m\}$,*

$$\limsup_{t\to\infty} \mathbb{E}\|e_t^x\| \;\le\; \frac{\widetilde{C}_\theta}{\mu}\; \limsup_{t\to\infty} \frac{\alpha_t}{\eta_t}. \tag{14}$$

Proof. Please refer to Appendix C for detailed proof. □

Theorem 4.9 (Dedicated-critic Bias)**.** *Suppose Assumptions 4.1–4.3 hold, let $\widehat{g}_t$ and $g_t^\star$ be the actor's estimated and ideal gradients, the dedicated-critic actor bias be $B_t^{\text{multi}} := \widehat{g}_t^{\text{multi}} - g_t^\star$. Then*

$$\limsup_{t\to\infty} \mathbb{E}\|B_t^{\text{multi}}\| \;\le\; GL_\phi\, \frac{\widetilde{C}_\theta}{\mu}\; \limsup_{t\to\infty} \frac{\alpha_t}{\eta_t}, \tag{15}$$

Proof. See Appendix E for detailed proof. □

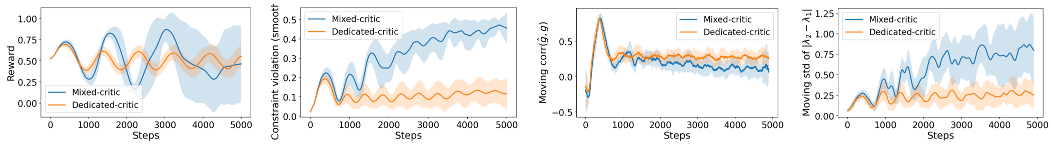
5 Experiments

5.1 Simple CMDP Bandit Problem

We design a minimal yet diagnostic constrained bandit CMDP that cleanly isolates the effect of *single* vs. *dedicated-critic* architectures under Lagrangian updates. All implementation choices below are fixed and reported for full reproducibility. We use a one-state bandit with binary actions $\mathcal{A} = \{a_1, a_2\}$, discount $\gamma = 0$, reward $r(a_1) = 0, r(a_2) = 1$, and two costs $c_1(a_1) = 0,\; c_1(a_2) = 1, c_2(a_1) = 1,\; c_2(a_2) = 0$, with constraints $J_{c_1} \le d_1$ and $J_{c_2} \le d_2$, where $d_1 = d_2 = 0.5$. The policy π_θ is a Bernoulli with a single logit $\theta \in \mathbb{R}$: $\pi_\theta(a_1) = \sigma(\theta), \quad \pi_\theta(a_2) = 1 - \sigma(\theta), \qquad \sigma(\theta) = \frac{1}{1+e^{-\theta}}$. Under this policy, expected costs are $J_{c_1} = 1 - \sigma(\theta)$, $J_{c_2} = \sigma(\theta)$, and expected reward is $J_r = \sigma(\theta)\, r(a_1) + (1 - \sigma(\theta))\, r(a_2) = 1 - \sigma(\theta)$.

We compare two actor–critic variants that share the same actor and dual updates. On each step, we sample $a \sim \pi_\theta$ and apply an update with a (learned) advantage surrogate from the critic(s): $\theta_{t+1} = \theta_t + \alpha\, \widehat{g}_t, \quad \widehat{g}_t := \nabla_\theta \log \pi_{\theta_t}(a_t)\, \widehat{Q}_t(a_t)$, where $\nabla_\theta \log \pi_\theta(a_1) = 1 - \sigma(\theta)$ and $\nabla_\theta \log \pi_\theta(a_2) = -\sigma(\theta)$. We maintain $\lambda = (\lambda_1, \lambda_2) \in \mathbb{R}_+^2$ with projected stochastic ascent: $\lambda_{i,t+1} = \Pi_{[0,\lambda_{\max}]}\Big(\lambda_{i,t} + \beta\,(\widehat{c}_i - d_i)\Big),\; i \in \{1, 2\}$, where $\widehat{c}_i$ is the instantaneous cost sample (0 or 1 in this bandit) and $\lambda_{\max} = 10$. Projection keeps λ_t in a compact set. In this bandit, the true Lagrangian gradient has a closed form. Let $\pi_1 = \sigma(\theta)$ and $f(a) := r(a) - \lambda_1 c_1(a) - \lambda_2 c_2(a), \qquad f(a_1) = r(a_1) - \lambda_2, \quad f(a_2) = r(a_2) - \lambda_1$. Then $g_t := \nabla_\theta \mathcal{L}(\theta_t, \lambda_t) = \pi_1(1 - \pi_1)\,\big(f(a_1) - f(a_2)\big) = \sigma(\theta_t)\big(1 - \sigma(\theta_t)\big)\,\big((r(a_1) - \lambda_{2,t}) - (r(a_2) - \lambda_{1,t})\big)$.

Results and Discussion: Both methods attain similar average returns, but the mixed-critic curve has much higher variance (large confidence band, occasional dips). The dedicated-critic maintains



(a) Average Returns (b) Constraint violation (c) Gradient alignment (d) Dual-oscillation

Figure 2: Performance for CMDP Bandit.

comparable reward with markedly lower variability. The mixed-critic exhibits large and growing volatility, whereas the dedicated-critic remains low and stable. This matches the theory: a mixed critic's target moves with $\Delta\lambda_t$, inducing oscillatory dual dynamics; dedicated critics avoid this λ–coupling.

To measure safety, we compute the violation which quantifies by how much the learned policy exceeds constraint thresholds at each step. Mixed-critic drifts to higher violations, while dedicated-critic settles much lower, implying better safety during training rather than only at convergence.

We also measure how well the actor's update direction matches the true Lagrangian gradient by computing a moving Pearson correlation between the estimated gradient $\widehat{g}_t$ and the true gradient g_t. Based on the results, dedicated-critic sustains a higher correlation than mixed-critic, indicating the actor follows the true Lagrangian gradient more reliably when critics are per-signal.

We quantify the stability of the dual variables by tracking the moving standard deviation of the gap $|\lambda_{2,t} - \lambda_{1,t}|$. This measures whether Lagrange multipliers converge smoothly or oscillate over time. The mixed-critic exhibits large and growing volatility, whereas the dedicated-critic remains low and stable. This matches the theory: a mixed critic's target moves with $\Delta\lambda_t$, inducing oscillatory dual dynamics; dedicated critics avoid this λ–coupling.

5.2 Multi-constraint Power System Application

Rather than relying on standard safe-RL benchmarks (whose constraints are few and stylized), we evaluate in a *complex energy scenario* designed to stress realism and constraint diversity. Compared with standard, the environment couples stochastic demand, renewable generation uncertainty, ramping limits, transmission congestion, reserve requirements, and device-level safety, yielding *multiple, interacting* constraints with heterogeneous timescales. This setting captures (i) tight operational envelopes,(ii) correlated risks across assets, and (iii) nontrivial trade-offs between cost and safety. In this case, we adopt a standard deep PPO configuration with neural-network critics (two-layer MLPs with 256 units per layer), so the empirical results directly evaluate our approach in the deep RL regime rather than in the idealised linear setting used for the theory.

System Overview and Constraints. We consider a radial distribution network with high rooftop PV penetration, where community battery energy storage systems (CBESSs) are coordinated to ensure safe and efficient operation. Each CBESS is subject to power, efficiency, and state-of-charge (SoC) constraints, can transact with the upstream grid under trading limits, and incurs both trading and degradation costs. When storage is saturated, PV curtailment is applied with fairness constraints to avoid disproportionate restrictions across buses. The system is modeled using the LinDistFlow approximation. The central control task is to schedule CBESS actions and PV curtailment to minimize trading cost while maintaining constraint satisfaction across multiple operational and fairness dimensions. To enforce safety and equity, we define five cost terms (constraints) monitored over the scheduling horizon: (1) *Voltage Violation Ratio* penalizes the number of buses breaching voltage limits; (2) *Voltage Deviation Degree* penalizes the severity of such violations; (3) *Line Loading Cost* penalizes thermal overloads on network branches; (4) *Battery Degradation Cost* discourages excessive CBESS cycling; and (5) *PV Curtailment Unfairness* penalizes uneven curtailment across buses. These constraints interact over heterogeneous timescales, capturing the multifaceted trade-offs in real-world power systems. Full modeling details are provided in Appendix J.

Results and Discussion: In this work, we prioritize *constraint satisfaction* as the central performance objective. Under consistent PPO backbones and training configurations, we compare two architectures: (i) the widely used PPO-Lagrangian baseline, which utilizes a single reward critic and a mixed critic for the aggregated cost signal, and (ii) the proposed Dedicated critic setup, which retains a shared reward critic but replaces the single cost critic with multiple per-constraint critics.

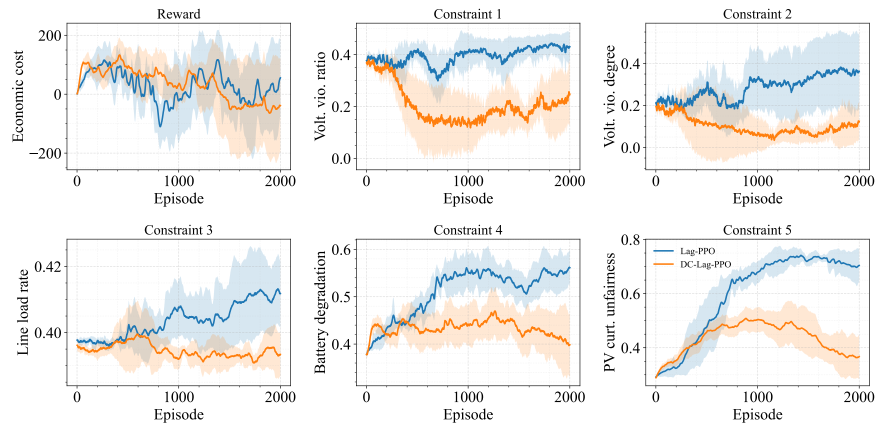


Figure 3: Learning curves for the power system application.

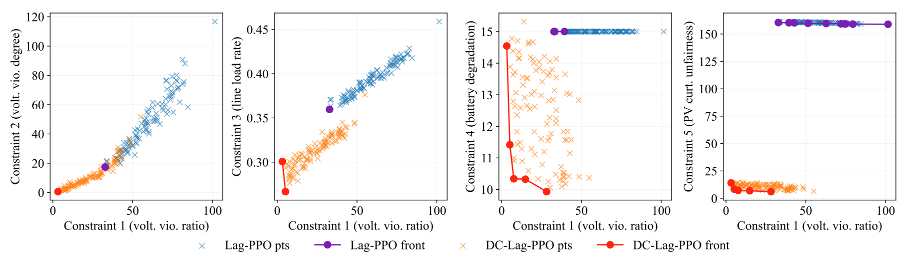


Figure 4: Pareto fronts from the test results of the power system application(the lower the better).

Experimental details are provided in Table 3 in Appendix J. As shown in Fig. 3, the dedicated architecture exhibits significantly more stable training behavior, with smoother learning curves and reduced variance in value estimates, while the baseline often suffers from unstable updates and erratic dual dynamics—particularly when constraint signals conflict. On unseen demand and renewable profiles, the Dedicated model consistently achieves the lowest violation rates and magnitudes across all five constraint dimensions, with noticeably fewer and shorter spikes in unsafe behavior. To quantify trade-offs between return and safety, we construct empirical Pareto fronts using ε-constraint sweeps. As shown in Fig. 4, dedicated policies cluster tightly near the estimated frontiers, consistently outperforming the baseline across a broad range of safety budgets—achieving either lower constraint violations for the same reward, or higher reward at equivalent violation levels.

These results collectively shows that performance gains of the dedicated setup are consistent with the theoretical mechanism identified in our analysis: by estimating each constraint with its own critic, the actor update depends on per-constraint advantages that are independent of the evolving multipliers, thereby avoiding the λ-driven target drift that can destabilise training. Per-constraint critics preserve the relative scale and variance of individual constraint signals, enabling head-wise normalisation and reducing "winner-takes-all" effects where the most active constraint dominates updates. This appears crucial for tracing clean Pareto sets: Dedicated policies concentrate near the frontier across budgets, whereas the aggregated-cost baseline often lies inside the frontier, indicative of optimisation bias introduced by collapsing constraints.

6 Conclusion

This paper examined how critic design shapes stability and safety in Lagrangian (policy-gradient) methods for constrained Markov decision processes. We showed, both theoretically and empirically, that mixing all reward and cost signals into a mixed critic couples the evaluation target to the evolving dual variables, introducing a form of dual-induced nonstationarity that can impair learning stability. In contrast, dedicated per-signal critics yield targets that depend solely on the policy, eliminating this source of drift. Our experiments across both bandit and stylized power system environments confirm these theoretical insights. This paper provides concrete guidance for the design of safe reinforcement learning algorithms under multiple constraints, highlighting the importance of critic architecture in ensuring both stability and constraint satisfaction in Lagrangian-based methods.

Ethics Statement

This work investigates algorithmic design choices for safe reinforcement learning in constrained Markov decision processes. All experiments are conducted exclusively in simulated environments; no human subjects, personal data, or identifiable information are involved at any stage. The complex energy-system environment used in our study is a stylized simulator designed to explore safety–performance trade-offs in high-stakes decision-making. It does not interface with or control any real-world infrastructure. Our aim is to advance the understanding of algorithmic safety in reinforcement learning without posing risks to individuals, communities, or operational systems.

Reproducibility Statement

To support the reproducibility of our results, we provide comprehensive details on training procedures, including step sizes, optimization parameters, and evaluation metrics used throughout the experiments. With the release of our codebase, we will ensure full transparency by including random seeds, environment specifications, dependency versions, and scripts necessary to replicate all experiments and figures. All results presented in the paper can be reproduced using the provided scripts without manual tuning. Where applicable, we will also include pretrained models and logs to facilitate result verification and benchmarking.

References

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International conference on machine learning*, pages 22–31. PMLR, 2017.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Chiara Calascibetta, Luca Biferale, Francesco Borra, Antonio Celani, and Massimo Cencini. Taming lagrangian chaos with multi-objective reinforcement learning. *The European Physical Journal E*, 46(3):9, 2023.

Xin Chen, Guannan Qu, Yujie Tang, Steven Low, and Na Li. Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. *IEEE Transactions on Smart Grid*, 13(4):2935–2958, 2022.

Yinlam Chow, Ofir Nachum, Edgar Duenez-Guzman, and Mohammad Ghavamzadeh. A lyapunov-based approach to safe reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Sebastian Junges, Nils Jansen, Christian Dehnert, Ufuk Topcu, and Joost-Pieter Katoen. Safety-constrained reinforcement learning for mdps. In *International conference on tools and algorithms for the construction and analysis of systems*, pages 130–146. Springer, 2016.

Dohyeong Kim, Kyungjae Lee, and Songhwai Oh. Trust region-based safe distributional reinforcement learning for multiple constraints. *Advances in neural information processing systems*, 36: 19908–19939, 2023.

Puze Liu, Davide Tateo, Haitham Bou Ammar, and Jan Peters. Robot reinforcement learning on the constraint manifold. In *Conference on Robot Learning*, pages 1357–1366. PMLR, 2022.

Ying Liu, Wenlong Ding, Xin Liu, Lijun Wang, Li Zhao, Zhenyu Chen, and Li Zhang. Ipo: Interior-point policy optimization under constraints. In *AAAI Conference on Artificial Intelligence*, 2020.

Rémi Munos. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, pages 560–567, 2003.

Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7(1):2, 2019b.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning (ICML)*, 2015. Backbone for TRPO-Lagrangian variants.

Xiaoying Shi, Yinliang Xu, Guibin Chen, and Ye Guo. An augmented lagrangian-based safe reinforcement learning algorithm for carbon-oriented optimal scheduling of ev aggregators. *IEEE Transactions on Smart Grid*, 15(1):795–809, 2023.

Adam Stooke and Pieter Abbeel. Responsive safety in reinforcement learning via pid lagrangian methods. *arXiv preprint arXiv:2007.03964*, 2020.

Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by pid lagrangian methods. In *International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3 (1):9–44, 1988.

Chen Tessler, Daniel J Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *International Conference on Learning Representations (ICLR)*, 2019.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996a.

John N Tsitsiklis and Benjamin Van Roy. Feature-based methods for large scale dynamic programming. *Machine Learning*, 22(1):59–94, 1996b.

Saiwei Wang, Xin Jin, Shuai Mao, Athanasios V Vasilakos, and Yang Tang. Model-free event-triggered optimal consensus control of multiple euler-lagrange systems via reinforcement learning. *IEEE Transactions on Network Science and Engineering*, 8(1):246–258, 2020.

Tong Wu, Anna Scaglione, and Daniel Arnold. Constrained reinforcement learning for predictive control in real-time stochastic dynamic optimal power flow. *IEEE Transactions on Power Systems*, 39(3):5077–5090, 2023.

Pan Xu, Yingbin Liang, and Guanghui Lan. Crpo: A new approach for safe reinforcement learning with convergence guarantees. In *International Conference on Machine Learning (ICML)*, 2021.

Ziming Yan and Yan Xu. Real-time optimal power flow: A lagrangian based deep reinforcement learning approach. *IEEE Transactions on Power Systems*, 35(4):3270–3273, 2020.

Qisong Yang, Thiago D Simão, Simon H Tindemans, and Matthijs TJ Spaan. Wcsac: Worst-case soft actor critic for safety-constrained reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10639–10646, 2021a.

Tsung-Yen Yang, Justinian Rosca, Karthik Narasimhan, and Peter J Ramadge. Projection-based constrained policy optimization. *arXiv preprint arXiv:2010.03152*, 2020.

Yuhang Yang, Yitao Liu, Huan Xu, and Peng Cheng. Projection-free lagrangian methods for constrained reinforcement learning. In *ICML*, 2021b.

Hailong Zhang, Jiankun Peng, Huachun Tan, Hanxuan Dong, and Fan Ding. A deep reinforcement learning-based energy management framework with lagrangian relaxation for plug-in hybrid electric vehicle. *IEEE Transactions on Transportation Electrification*, 7(3):1146–1160, 2020.

Lixian Zhang, Ruixian Zhang, Tong Wu, Rui Weng, Minghao Han, and Ye Zhao. Safe reinforcement learning with stability guarantee for motion planning of autonomous vehicles. *IEEE transactions on neural networks and learning systems*, 32(12):5435–5444, 2021.

Yuxiang Zhang, Xiaoling Liang, Dongyu Li, Shuzhi Sam Ge, Bingzhao Gao, Hong Chen, and Tong Heng Lee. Adaptive safe reinforcement learning with full-state constraints and constrained adaptation for autonomous vehicles. *IEEE Transactions on Cybernetics*, 54(3):1907–1920, 2023.

A LIMITATIONS

A fundamental tension exists in safe RL between maximizing cumulative reward and satisfying multiple constraints—particularly in realistic, high-stakes domains. While our **Dedicated Critic** architecture significantly enhances constraint adherence and stabilizes training dynamics, it does not resolve the inherent trade-off: enforcing stricter safety often reduces achievable reward or slows convergence. The approach also introduces greater computational and memory requirements, as it maintains a separate critic for each constraint. This design may be less scalable in environments with many constraints or where critic updates are expensive. Although shared-backbone models with multiple heads offer a partial remedy, they require careful balancing and tuning to be effective.

Another limitation is the static one-to-one mapping between constraints and critics. In settings where constraints vary in relevance or activate sparsely, some critics may be under-trained, reducing sample efficiency (see Appendix F). Future directions could involve adaptive critic selection or shared-parameter architectures that dynamically reallocate capacity based on constraint salience.

The theoretical results provide asymptotic error bounds and bias characterisations, rather than full finite-time guarantees for PPO-style deep RL under realistic training regimes. However, this limitation is not specific to our approach: most practical deep RL algorithms rely on similar heuristic stabilisation mechanisms and likewise lack end-to-end finite-time guarantees. In our implementation, we mitigate these issues using standard techniques—advantage normalisation, conservative learning rates, PPO ratio clipping, gradient clipping, and bounding the dual variables—and empirically observe stable learning across seeds. Nevertheless, a more refined finite-time analysis that explicitly captures these practical design choices remains an open direction, so our current guarantees should be interpreted as qualitative guidance on critic design rather than a complete convergence certificate for deep safe RL.

Finally, although our theoretical results are first presented under linear function approximation for analytical clarity, we also extend them to nonlinear cases such as neural networks. Still, understanding the implications of gradient bias in deep architectures—particularly when critics share representations—remains an open question warranting further theoretical and empirical investigation.

B PBE

We justify that, for each fixed policy π_θ and each signal $x \in \{r, c_1, \dots, c_m\}$, the population target of the per-signal critic is the unique solution of a linear system $A(\theta)\,\omega^{x,*}(\theta) = b^x(\theta)$ with $A(\theta) \succeq \mu I$, and that $A(\cdot)$, $b^x(\cdot)$ are locally Lipschitz in θ under standard conditions.

Let $\mathcal{E}^x_{\pi_\theta}(s)$ denote the discounted state value for signal x under policy π_θ and let $\phi : \mathcal{S} \to \mathbb{R}^d$ be a fixed feature map. We approximate $\mathcal{E}^x_{\pi_\theta}(s) \approx \phi(s)^\top \omega^x$. Write $\Phi \in \mathbb{R}^{n\times d}$ for the matrix stacking feature rows $\phi(s)^\top$, $D = \operatorname{diag}(d_{\pi_\theta})$ for the diagonal matrix of the on-policy stationary distribution over states, and P_{π_θ} for the state transition kernel.

The projected fixed-point equation (PFE) in the D-weighted norm is

$$\Phi\omega^x \;=\; \Pi\left(\mathcal{T}^x_{\pi_\theta}(\Phi\omega^x)\right), \qquad \mathcal{T}^x_{\pi_\theta} v \;=\; r^x + \gamma P_{\pi_\theta} v,$$

where Π is the D-orthogonal projection onto $\operatorname{span}(\Phi)$ and $r^x \in \mathbb{R}^n$ is the immediate signal vector (r for $x = r$, c_i for $x = c_i$). The normal equations are

$$\Phi^\top D\Big(\Phi\omega^x - (r^x + \gamma P_{\pi_\theta}\Phi\omega^x)\Big) = 0 \iff \underbrace{\Phi^\top D(I - \gamma P_{\pi_\theta})\Phi}_{A(\theta)}\,\omega^x \;=\; \underbrace{\Phi^\top D\, r^x}_{b^x(\theta)}.$$

Hence the population target satisfies $A(\theta)\,\omega^{x,*}(\theta) = b^x(\theta)$.

Assume (i) *ergodicity*: the Markov chain under π_θ admits a stationary distribution d_{π_θ} with full support on the on-policy visited set; (ii) *feature non-degeneracy*: the columns of $D^{1/2}\Phi$ are linearly independent. Then $A(\theta) = \Phi^\top D(I - \gamma P_{\pi_\theta})\Phi$ is symmetric positive definite; in particular there exists $\mu > 0$ with $A(\theta) \succeq \mu I$, so the solution $\omega^{x,*}(\theta)$ is unique.

If π_θ is C^1 in θ and ergodicity holds on a neighbourhood, then P_{π_θ} and d_{π_θ} vary locally Lipschitzly in θ. Since Φ is fixed, $A(\theta) = \Phi^\top D(I - \gamma P_{\pi_\theta})\Phi$ and $b^x(\theta) = \Phi^\top D r^x$ inherit local Lipschitzness.

If the estimator depends on (s, a), take features $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, stack Φ over (s, a), let $D = \mathrm{diag}(d_{\pi_\theta}(s, a))$ be the on-policy state–action occupancy matrix, and use the state–action transition kernel P^{π_θ}. The same PFE derivation yields

$$A(\theta) \;=\; \Phi^\top D\,(I - \gamma P^{\pi_\theta})\,\Phi, \qquad b^x(\theta) \;=\; \Phi^\top D\, r^x,$$

so $A(\theta)\omega^{x,*}(\theta) = b^x(\theta)$ with $A(\theta) \succeq \mu I$ under the analogues of ergodicity and feature non-degeneracy for (s, a). Local Lipschitzness follows as above.

When training with advantages (e.g., GAE), two standard constructions lead to a linear system:

(1) *Difference-of-values:* Learn V^x (or Q^x) with the state/state–action equations above, and form $A^x = Q^x - V^x$; the critic parameters still solve $A(\theta)\omega^{x,*} = b^x(\theta)$.

(2) *Least-squares to generalized returns:* Regress $\phi(z)^\top \omega^x$ onto generalized returns $\hat{G}^x$ (e.g., GAE targets) in the D-weighted norm, i.e. $\min_{\omega^x} \mathbb{E}_{z\sim d_{\pi_\theta}}\big[\,(\phi(z)^\top\omega^x - \hat{G}^x(z))^2\,\big]$. The normal equations are

$$\underbrace{\Phi^\top D\,\Phi}_{A(\theta)}\,\omega^{x,*}(\theta) \;=\; \underbrace{\Phi^\top D\,\hat{G}^x}_{b^x(\theta)}.$$

Thus the linear model still holds (with a different A), and $A(\theta) \succeq \mu I$ under $D^{1/2}\Phi$ full column rank. If $\hat{G}^x$ depends smoothly on θ through π_θ, $b^x(\cdot)$ is locally Lipschitz.

Lemma B.1 (Positive definiteness of $A(\theta)$)**.** *Fix a policy π_θ and let $D = \mathrm{diag}(d_{\pi_\theta})$ be the diagonal matrix of the on-policy stationary distribution over states. Let P_{π_θ} be the corresponding state transition kernel (row-stochastic) satisfying $d_{\pi_\theta}^\top P_{\pi_\theta} = d_{\pi_\theta}^\top$. Let $\Phi \in \mathbb{R}^{n\times d}$ stack feature rows and assume $D^{1/2}\Phi$ has full column rank. For $\gamma \in [0, 1)$ define*

$$A(\theta) \;=\; \Phi^\top D\,(I - \gamma P_{\pi_\theta})\,\Phi.$$

Then $A(\theta)$ is symmetric positive definite and

$$v^\top A(\theta)\,v \;\geq\; (1-\gamma)\,\lambda_{\min}\big(\Phi^\top D\,\Phi\big)\,\|v\|_2^2 \qquad \textit{for all } v \in \mathbb{R}^d.$$

In particular, $A(\theta) \succeq \mu I$ with $\mu = (1-\gamma)\,\lambda_{\min}(\Phi^\top D\Phi) > 0$.

Proof. Let $y = \Phi v$. Using the D-weighted inner product $\langle u, w\rangle_D := u^\top D w$ and norm $\|u\|_D^2 := \langle u, u\rangle_D$,

$$v^\top A(\theta)\,v = y^\top D(I - \gamma P_{\pi_\theta})y = \|y\|_D^2 - \gamma\,\langle y, P_{\pi_\theta} y\rangle_D.$$

Because P_{π_θ} is a Markov operator with invariant measure d_{π_θ}, it is a non-expansion in $L_2(D)$, i.e., $\|P_{\pi_\theta} y\|_D \leq \|y\|_D$ and therefore $\langle y, P_{\pi_\theta} y\rangle_D \leq \|y\|_D\,\|P_{\pi_\theta} y\|_D \leq \|y\|_D^2$. Hence

$$v^\top A(\theta)\,v \;\geq\; \|y\|_D^2 - \gamma\,\|y\|_D^2 = (1-\gamma)\,\|y\|_D^2.$$

Finally, $\|y\|_D^2 = v^\top \Phi^\top D\Phi\,v \geq \lambda_{\min}(\Phi^\top D\Phi)\,\|v\|_2^2$ because $D^{1/2}\Phi$ has full column rank. Combining the inequalities yields the claim. □

Corollary B.2 (State–action variant)**.** *Let $D = \mathrm{diag}(d_{\pi_\theta}(s, a))$ be the on-policy state–action occupancy matrix, P^{π_θ} the state–action transition kernel (row-stochastic) with $d_{\pi_\theta}^\top P^{\pi_\theta} = d_{\pi_\theta}^\top$, and Φ stack features over (s, a) with $D^{1/2}\Phi$ full column rank. Define*

$$A(\theta) \;=\; \Phi^\top D\,(I - \gamma P^{\pi_\theta})\,\Phi.$$

Then $A(\theta) \succeq (1-\gamma)\,\lambda_{\min}(\Phi^\top D\Phi)\,I$ and is symmetric positive definite.

C DETAILED PROOF FOR MIXED-CRITIC ERROR BOUND

Lemma C.1 (Mixed-critic error bound)**.** *If $A(\theta) \succeq \mu I$ uniformly, and the stepsizes satisfy Robbins–Monro conditions with $\alpha_t/\eta_t \to 0$ and $\beta_t/\eta_t \to 0$, then there exist constants $C_\lambda, C_\theta < \infty$ such that*

$$\limsup_{t\to\infty} \mathbb{E}\|e_t\| \;\le\; \frac{C_\lambda}{\mu}\,\limsup_{t\to\infty}\frac{\beta_t}{\eta_t} \;+\; \frac{C_\theta}{\mu}\,\limsup_{t\to\infty}\frac{\alpha_t}{\eta_t} \;+\; O(1). \tag{16}$$

Proof. Recall the error recursion equation 4:

$$e_{t+1} = (I - \eta_t A(\theta_t))e_t + \underbrace{(\omega_t^\star - \omega_{t+1}^\star)}_{\text{target drift}} + \eta_t \zeta_{t+1} + \Delta_t^\theta.$$

Since $A(\theta_t) \succeq \mu I$, we have for all v, $\|(I - \eta_t A(\theta_t))v\| \le (1-\mu\eta_t)\|v\|$. Hence

$$\|(I - \eta_t A(\theta_t))e_t\| \;\le\; (1-\mu\eta_t)\,\|e_t\|. \tag{17}$$

Using the drift expansion equation 7,

$$\omega_t^\star - \omega_{t+1}^\star = -\sum_{i=1}^{m}(\lambda_{i,t+1} - \lambda_{i,t})\,\omega^{c_i}(\theta_t) + O(\|\theta_{t+1} - \theta_t\|).$$

Assuming the PBE solutions $\omega^{c_i}(\theta)$ and $\omega^r(\theta)$ are Lipschitz in θ (true under our linear/PBE setup with $A(\theta)$, $b^x(\theta)$ smoothly varying), there exist constants C_λ, C_θ s.t.

$$\|\omega_t^\star - \omega_{t+1}^\star\| \;\le\; C_\lambda\,\|\lambda_{t+1} - \lambda_t\| + C_\theta\,\|\theta_{t+1} - \theta_t\|. \tag{18}$$

By the definitions of the dual and actor steps, $\|\lambda_{t+1} - \lambda_t\| = O(\beta_t)$ and $\|\theta_{t+1} - \theta_t\| = O(\alpha_t)$, hence

$$\|\omega_t^\star - \omega_{t+1}^\star\| \;\le\; C_\lambda\,\beta_t + C_\theta\,\alpha_t. \tag{19}$$

Write $\delta_t := (\omega_t^\star - \omega_{t+1}^\star) + \Delta_t^\theta$. By equation 19 and Lipschitz variation of $A(\theta), b(\theta)$ (collected in Δ_t^θ), there exists $\widetilde{C}_\theta$ s.t.

$$\mathbb{E}\|\delta_t\| \;\le\; C_\lambda\,\beta_t + \widetilde{C}_\theta\,\alpha_t. \tag{20}$$

Using equation 17 and the triangle inequality,

$$\|e_{t+1}\| \;\le\; (1-\mu\eta_t)\|e_t\| + \|\delta_t\| + \eta_t\|\zeta_{t+1}\|.$$

Take conditional expectation and then total expectation. With $\mathbb{E}[\zeta_{t+1}|\mathcal{F}_t] = 0$ and $\mathbb{E}\|\zeta_{t+1}\|^2 \le \sigma^2$, standard SA arguments (via a mean-square detour or BDG inequality) yield

$$\mathbb{E}\big[\eta_t\|\zeta_{t+1}\|\big] \;\le\; C_{\text{noise}}\,\eta_t^2, \tag{21}$$

for some constant C_{noise} (intuitively, the "linear in η_t" noise can be handled through a square-norm contraction; in the first-moment recursion it appears as $O(\eta_t^2)$). Hence, taking total expectation and applying equation 20 gives

$$\mathbb{E}\|e_{t+1}\| \;\le\; (1-\mu\eta_t)\,\mathbb{E}\|e_t\| \;+\; C_\lambda\,\beta_t \;+\; \widetilde{C}_\theta\,\alpha_t \;+\; C_{\text{noise}}\,\eta_t^2. \tag{22}$$

We now use a standard comparison lemma: if a nonnegative sequence (x_t) satisfies

$$x_{t+1} \le (1-a_t)x_t + b_t, \quad a_t \in (0,1),\; \sum_t a_t = \infty,\; a_t \to 0,$$

then

$$\limsup_{t\to\infty} x_t \;\le\; \limsup_{t\to\infty}\frac{b_t}{a_t}.$$

Applying this to equation 22 with $x_t = \mathbb{E}\|e_t\|$, $a_t = \mu\eta_t$ and

$$b_t \;=\; C_\lambda\,\beta_t + \widetilde{C}_\theta\,\alpha_t + C_{\text{noise}}\,\eta_t^2,$$

gives

$$\limsup_{t\to\infty} \mathbb{E}\|e_t\| \;\le\; \frac{1}{\mu}\,\limsup_{t\to\infty}\left(\frac{C_\lambda\,\beta_t}{\eta_t} + \frac{\widetilde{C}_\theta\,\alpha_t}{\eta_t} + C_{\text{noise}}\,\eta_t\right).$$

Since $\eta_t \to 0$ and $\sum_t \eta_t^2 < \infty$, the last term contributes $O(1)$. Renaming $\widetilde{C}_\theta$ as C_θ yields equation 9. □

Lemma C.2 (dedicated-critic tracking error)**.** *Suppose $A(\theta) \succeq \mu I$ uniformly, Assumptions 4.1 and 4.3 hold, then there exists $\widetilde{C}_\theta < \infty$ such that for every $x \in \{r, c_1, \ldots, c_m\}$,*

$$\limsup_{t\to\infty} \mathbb{E}\|e_t^x\| \;\leq\; \frac{\widetilde{C}_\theta}{\mu}\; \limsup_{t\to\infty} \frac{\alpha_t}{\eta_t}. \tag{23}$$

Proof. Use $A(\theta_t) \succeq \mu I$:

$$\|e_{t+1}^x\| \;\leq\; \|(I - \eta_t A(\theta_t))e_t^x\| + \eta_t\|\zeta_{t+1}^x\| + \|\Delta_t^{\theta,x}\| \;\leq\; (1-\mu\eta_t)\|e_t^x\| + \eta_t\|\zeta_{t+1}^x\| + C_\theta\|\theta_{t+1} - \theta_t\|.$$

Take conditional expectation given $\mathcal{F}_t$ and then expectation; by Assumption 4.3, $\mathbb{E}[\eta_t\|\zeta_{t+1}^x\|] \leq c\,\eta_t$ and yields $O(\eta_t^2)$ at the level of first-moment recursion. By Assumption 4.1, $\|\theta_{t+1} - \theta_t\| = O(\alpha_t)$. Thus,

$$\mathbb{E}\|e_{t+1}^x\| \;\leq\; (1-\mu\eta_t)\,\mathbb{E}\|e_t^x\| \;+\; C_\theta\,\alpha_t \;+\; O(\eta_t^2).$$

Apply the standard SA comparison lemma for sequences of the form $x_{t+1} \leq (1-a_t)x_t + b_t$ with $a_t = \mu\eta_t$ and $b_t = C_\theta\alpha_t + O(\eta_t^2)$, using $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$, and $\alpha_t/\eta_t \to 0$. This yields

$$\limsup_{t\to\infty} \mathbb{E}\|e_t^x\| \;\leq\; \frac{C_\theta}{\mu}\; \limsup_{t\to\infty} \frac{\alpha_t}{\eta_t},$$

and absorbing constants into $\widetilde{C}_\theta$ gives equation 14. □

D Detailed Proof for Actor-gradient bias bound

Lemma D.1 (Actor-gradient bias bound)**.** *Let $\widehat{g}_t$ and $g_t^\star$ be the actor's estimated and ideal gradients,*

$$\widehat{g}_t = \mathbb{E}_t\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top \omega_t\big], \qquad g_t^\star = \mathbb{E}_t\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top \omega_t^\star\big], \tag{24}$$

with critic error $e_t = \omega_t - \omega_t^\star$. Assume the score and features are bounded as

$$\|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\| \le G, \qquad \|\phi(s_t,a_t)\| \le L_\phi \quad a.s.$$

Then the actor-gradient bias

$$B_t \;:=\; \widehat{g}_t - g_t^\star$$

satisfies

$$\|B_t\| \;\le\; G L_\phi\,\|e_t\|. \tag{25}$$

Proof. Recall the definitions

$$\widehat{g}_t = \mathbb{E}\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top \omega_t \,\big|\, \mathcal{F}_t\big], \quad g_t^\star = \mathbb{E}\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top \omega_t^\star \,\big|\, \mathcal{F}_t\big],$$

and the critic error $e_t = \omega_t - \omega_t^\star$. Here $\mathcal{F}_t$ is the sigma-field generated by everything up to time t; in particular, $\theta_t, \lambda_t, \omega_t, \omega_t^\star, e_t$ are $\mathcal{F}_t$-measurable, while (s_t, a_t) are drawn from π_{θ_t} at time t and are not $\mathcal{F}_t$-measurable.

Using linearity of conditional expectation and $e_t = \omega_t - \omega_t^\star$,

$$\begin{aligned} B_t &\coloneqq \widehat{g}_t - g_t^\star \\ &= \mathbb{E}\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top (\omega_t - \omega_t^\star) \,\big|\, \mathcal{F}_t\big] \\ &= \mathbb{E}\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top e_t \,\big|\, \mathcal{F}_t\big]. \end{aligned}$$

Since e_t is $\mathcal{F}_t$-measurable, we can factor it outside the conditional expectation: for any random matrix/vector X and $\mathcal{F}_t$-measurable (deterministic under $\mathbb{E}[\cdot|\mathcal{F}_t]$) vector Y,

$$\mathbb{E}[X\,Y \mid \mathcal{F}_t] \;=\; \mathbb{E}[X \mid \mathcal{F}_t]\,Y.$$

Applying this with $X := \nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top$ and $Y := e_t$,

$$B_t = \mathbb{E}\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top \,\big|\, \mathcal{F}_t\big]\; e_t.$$

Equivalently, without explicitly pulling out the matrix, we can directly bound the norm inside the conditional expectation as follows.

From the triangle inequality for norms, we have

$$\|B_t\| = \Big\|\mathbb{E}\big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top e_t \,\big|\, \mathcal{F}_t\big]\Big\| \;\le\; \mathbb{E}\big[\big\|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top e_t\big\| \,\big|\, \mathcal{F}_t\big],$$

where we used Jensen's inequality for the convex function $x \mapsto \|x\|$ and conditional expectation.

Now use submultiplicativity of operator/vector norms:

$$\big\|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top e_t\big\| \;\le\; \|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\|\;\|\phi(s_t,a_t)\|\;\|e_t\|.$$

Here we regard the product $\psi\,\phi^\top e_t$ (with $\psi := \nabla_\theta \log \pi_{\theta_t}(a_t|s_t)$) as $(\psi\,\phi^\top)e_t$; the operator norm of the rank-1 matrix $\psi\,\phi^\top$ is $\|\psi\|\,\|\phi\|$.

Therefore,

$$\|B_t\| \;\le\; \mathbb{E}\big[\|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\|\,\|\phi(s_t,a_t)\|\,\|e_t\| \,\big|\, \mathcal{F}_t\big].$$

Assume the standard boundedness conditions hold almost surely:

$$\|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\| \le G, \qquad \|\phi(s_t,a_t)\| \le L_\phi.$$

Since $\|e_t\|$ is $\mathcal{F}_t$-measurable, we can treat it as a constant inside the conditional expectation. Hence,

$$\|B_t\| \;\le\; \mathbb{E}\big[G\,L_\phi\,\|e_t\| \,\big|\, \mathcal{F}_t\big] \;=\; G\,L_\phi\,\|e_t\|.$$

This establishes the claimed Lipschitz bound

$$\|B_t\| \le G L_\phi \|e_t\|,$$

which is precisely equation 11. $\square$

E Detailed Proof for Actor-gradient bias bound

Theorem E.1 (Bias from a Mixed Critic)**.** *Assume the conditions of Lemma 4.5 hold, and the score/features are uniformly bounded $\|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\| \le G$, $\|\phi(s_t,a_t)\| \le L_\phi$ a.s. Then the actor-gradient bias $B_t = \widehat{g}_t - g_t^\star$ satisfies*

$$\limsup_{t\to\infty} \mathbb{E}\|B_t\| \;\le\; GL_\phi \left(\frac{C_\lambda}{\mu} \limsup_{t\to\infty} \frac{\beta_t}{\eta_t} \;+\; \frac{C_\theta}{\mu} \limsup_{t\to\infty} \frac{\alpha_t}{\eta_t} \right). \tag{26}$$

Proof. By Lemma 4.6,

$$\|B_t\| \;\le\; GL_\phi \,\|e_t\|.$$

Taking expectations preserves the inequality (monotonicity of $\mathbb{E}$):

$$\mathbb{E}\|B_t\| \;\le\; GL_\phi \,\mathbb{E}\|e_t\|. \tag{27}$$

Lemma 4.5 states that, for some finite C_λ, C_θ,

$$\limsup_{t\to\infty} \mathbb{E}\|e_t\| \;\le\; \frac{C_\lambda}{\mu} \limsup_{t\to\infty} \frac{\beta_t}{\eta_t} \;+\; \frac{C_\theta}{\mu} \limsup_{t\to\infty} \frac{\alpha_t}{\eta_t} \;+\; O(1). \tag{28}$$

Here the $O(1)$ term collects vanishing contributions such as $O(\eta_t)$ from the noise control (cf. the proof of Lemma 4.5).

Taking $\limsup_{t\to\infty}$ on both sides of equation 27 and using equation 28 yields

$$\limsup_{t\to\infty} \mathbb{E}\|B_t\| \;\le\; GL_\phi \;\limsup_{t\to\infty} \mathbb{E}\|e_t\| \;\le\; GL_\phi \left(\frac{C_\lambda}{\mu} \limsup_{t\to\infty} \frac{\beta_t}{\eta_t} \;+\; \frac{C_\theta}{\mu} \limsup_{t\to\infty} \frac{\alpha_t}{\eta_t} \;+\; O(1) \right).$$

Since $G, L_\phi, C_\lambda, C_\theta, \mu$ are constants (independent of t), and $\limsup(O(1)) = 0$, we can drop the vanishing term to obtain exactly equation 12.

In Lemma 4.5, the $\frac{\beta_t}{\eta_t}$ contribution arises from the *dual-driven target drift* in the mixed-critic error dynamics (see the decomposition of $\omega_t^\star - \omega_{t+1}^\star$). Thus the bound equation 12 explicitly exposes the additional bias component inherited from the mixed critic's dependence on λ. □

Theorem E.2 (dedicated-critic bias)**.** *Suppose Assumptions 4.1–4.3 hold, $A(\theta) \succeq \mu I$ uniformly, and $\|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\| \le G$, $\|\phi(s_t,a_t)\| \le L_\phi$ a.s. Let $\widehat{g}_t$ and $g_t^\star$ be the actor's estimated and ideal gradients, the dedicated-critic actor bias be $B_t^{multi} := \widehat{g}_t^{multi} - g_t^\star$. Then*

$$\limsup_{t\to\infty} \mathbb{E}\|B_t^{multi}\| \;\le\; GL_\phi \,\frac{\widetilde{C}_\theta}{\mu} \;\limsup_{t\to\infty} \frac{\alpha_t}{\eta_t}, \tag{29}$$

Proof. Define the ideal (mixed) gradient at time t and its estimator as

$$g_t^\star = \mathbb{E}_t\Big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\big(\phi(s_t,a_t)^\top \omega^r(\theta_t) - \sum_{i=1}^m \lambda_{i,t}\,\phi(s_t,a_t)^\top \omega^{c_i}(\theta_t)\big)\Big],$$

$$\widehat{g}_t^{\text{multi}} = \mathbb{E}_t\Big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\big(\phi(s_t,a_t)^\top \omega_t^r - \sum_{i=1}^m \lambda_{i,t}\,\phi(s_t,a_t)^\top \omega_t^{c_i}\big)\Big]. \tag{30}$$

From equation 30 and linearity,

$$B_t^{\text{multi}} = \mathbb{E}_t\Big[\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\,\phi(s_t,a_t)^\top \Big(e_t^r - \sum_{i=1}^m \lambda_{i,t}\,e_t^{c_i}\Big)\Big],$$

where $e_t^x = \omega_t^x - \omega^x(\theta_t)$. Using Jensen, submultiplicativity, and boundedness of score and features,

$$\|B_t^{\text{multi}}\| \;\le\; \mathbb{E}_t\Big[\|\nabla_\theta \log \pi_{\theta_t}(a_t|s_t)\|\,\|\phi(s_t,a_t)\|\Big(\|e_t^r\| + \sum_{i=1}^m \|\lambda_{i,t}\|\,\|e_t^{c_i}\|\Big)\Big] \;\le\; GL_\phi\Big(\|e_t^r\| + \Lambda \max_i \|e_t^{c_i}\|\Big),$$

where $\Lambda = \sup_t \|\lambda_t\| < \infty$ due to projection onto a compact set. Taking expectations and $\limsup$,

$$\limsup_{t\to\infty} \mathbb{E}\|B_t^{\text{multi}}\| \;\le\; GL_\phi\Big(1+\Lambda\Big) \;\limsup_{t\to\infty} \max_{x\in\{r,c_i\}} \mathbb{E}\|e_t^x\|.$$

Apply Lemma 4.8 to bound each $\mathbb{E}\|e_t^x\|$ by $\frac{\widetilde{C}_\theta}{\mu} \limsup \frac{\alpha_t}{\eta_t}$, and absorb $(1+\Lambda)$ into $\widetilde{C}_\theta$ (renaming the constant) to get equation 15. No β_t term appears and the target drift involves only θ (rate α_t), not λ. □

F Dual-induced drift and linearity

Let the mixed critic be trained by minimizing any smooth population loss

$$\mathcal{L}_{\text{mix}}(\omega;\, \theta, \lambda) \quad \text{(e.g., TD loss, Monte-Carlo/GAE regression, etc.).}$$

Because the scalarized signal is $r_\lambda := r - \sum_{i=1}^{m} \lambda_i c_i$, this loss depends *explicitly* on λ. Denote the population minimizer by $\omega^\star(\theta, \lambda) \in \arg\min_\omega \mathcal{L}_{\text{mix}}(\omega; \theta, \lambda)$. At any (strict) local minimum, the first-order condition holds:

$$\nabla_\omega \mathcal{L}_{\text{mix}}(\omega^\star(\theta, \lambda); \theta, \lambda) = 0.$$

Assume the Hessian $H(\theta, \lambda) := \nabla^2_{\omega\omega} \mathcal{L}_{\text{mix}}(\omega^\star(\theta, \lambda); \theta, \lambda)$ is nonsingular (standard local strong convexity around the solution). Then by the implicit function theorem, $\omega^\star$ is differentiable in (θ, λ) and

$$\frac{\partial \omega^\star}{\partial \lambda} \;=\; -\,H(\theta, \lambda)^{-1} \underbrace{\nabla^2_{\omega\lambda}\, \mathcal{L}_{\text{mix}}(\omega^\star(\theta, \lambda); \theta, \lambda)}_{\neq 0 \text{ generically}}.$$

Hence for small updates $(\Delta\theta, \Delta\lambda)$,

$$\omega^\star(\theta + \Delta\theta, \lambda + \Delta\lambda) - \omega^\star(\theta, \lambda) \;=\; \underbrace{\frac{\partial \omega^\star}{\partial \theta}\, \Delta\theta}_{\text{policy-induced drift}} \;+\; \underbrace{\frac{\partial \omega^\star}{\partial \lambda}\, \Delta\lambda}_{\text{dual-induced drift}} \;+\, o(\|\Delta\theta\| + \|\Delta\lambda\|).$$

The key point is that $\nabla^2_{\omega\lambda} \mathcal{L}_{\text{mix}} \neq 0$ whenever the training targets or TD errors inside $\mathcal{L}_{\text{mix}}$ depend on r_λ (which they do for any mixed critic). Therefore, $\partial \omega^\star / \partial \lambda \neq 0$ generically, and the *dual-induced drift* term proportional to $\Delta\lambda$ appears *regardless of linearity*. The linear case analysed in the main text is just the special instance where $\mathcal{L}_{\text{mix}}$ yields normal equations $A(\theta)\omega = b^r(\theta) - \sum_i \lambda_i b^{c_i}(\theta)$, so that $\partial \omega^\star / \partial \lambda = -A(\theta)^{-1} [b^{c_1}(\theta), \ldots, b^{c_m}(\theta)]$ explicitly.

Why dedicated critics avoid it. For per-signal critics, each loss $\mathcal{L}_x(\omega^x; \theta)$ *does not* involve λ:

$$\nabla_\omega \mathcal{L}_x(\omega^{x,\star}(\theta); \theta) = 0 \qquad \Rightarrow \qquad \frac{\partial \omega^{x,\star}}{\partial \lambda} = 0.$$

Thus their targets drift only through θ (policy-induced), with *no* dual-induced component. When the actor later combines the already-computed per-signal estimates as $\omega^{\text{mix}} = \omega^r - \sum_i \lambda_i \omega^{c_i}$, the λ's appear *outside* the critics and do not change the critics' own population optima.

G Computational Resources

We implement all experiments using PyTorch-1.12 on an Ubuntu 18.04 server with two Intel Xeon Gold 6142M CPUs with 16 cores, 24G memory, and one NVIDIA 3090 GPU.

To further clarify constraint satisfaction during testing, Fig. 5 reports the measured computation time per training epoch under different numbers of Lagrangian critics (i.e., constraints). The results exhibit a clear linear trend, quantified by the fitted regression:

$$y = 0.00169x + 0.00776, \tag{31}$$

indicating stable and predictable scaling as the number of constraints increases. Importantly, the variance bars are small across all cases, showing that the training remains stable even when more constraints are introduced.

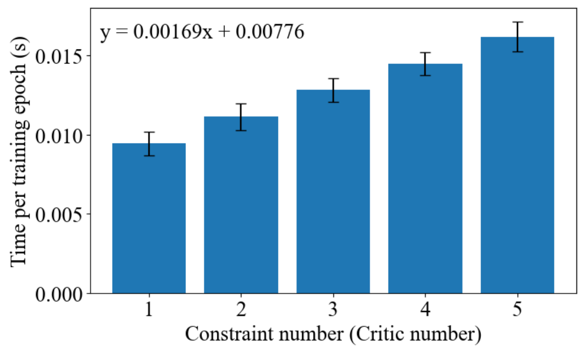


Figure 5: Time consumption of the proposed method with different number of critics (constraints).

H Experiment Detail - CMDP Bandit

Single mixed critic: A *single* scalar critic per action is trained on the mixed signal

$$r_\lambda(a) \;:=\; r(a) - \lambda_1 c_1(a) - \lambda_2 c_2(a).$$

With $\gamma = 0$, a TD(0) bandit update reduces to exponential averaging:

$$Q^{\text{mixed}}_{t+1}(a) \;=\; Q^{\text{mixed}}_t(a) + \eta_{\text{mixed}}\Big(r_\lambda(a_t) \;-\; Q^{\text{mixed}}_t(a_t)\Big)\,\mathbf{1}\{a_t = a\}. \tag{32}$$

The actor readout in equation 5.1 uses $\widehat{Q}_t(a) = Q^{\text{mixed}}_t(a)$.

Dedicated-critic: We train *separate* per-action critics for reward and each cost:

$$Q^{r}_{t+1}(a) = Q^{r}_t(a) + \eta_{\text{multi}}\big(r(a_t) - Q^{r}_t(a_t)\big)\,\mathbf{1}\{a_t = a\}, \tag{33}$$

$$Q^{c_1}_{t+1}(a) = Q^{c_1}_t(a) + \eta_{\text{multi}}\big(c_1(a_t) - Q^{c_1}_t(a_t)\big)\,\mathbf{1}\{a_t = a\}, \tag{34}$$

$$Q^{c_2}_{t+1}(a) = Q^{c_2}_t(a) + \eta_{\text{multi}}\big(c_2(a_t) - Q^{c_2}_t(a_t)\big)\,\mathbf{1}\{a_t = a\}. \tag{35}$$

The actor combines them *at readout time* with the *current* multipliers:

$$\widehat{Q}_t(a) \;=\; Q^{r}_t(a) \;-\; \lambda_{1,t}\,Q^{c_1}_t(a) \;-\; \lambda_{2,t}\,Q^{c_2}_t(a). \tag{36}$$

We run $T = 5000$ steps per seed and average over $S = 15$ random seeds for the main curves. For the *timescale ablation* (Sec. H.2), we sweep critic and dual learning rates and average over 8 seeds per grid point. For a mixed scalar summary of conditional alignment (reported once), we optionally use $S = 20$ seeds to reduce variance. $\alpha = 0.02, \beta = 0.02, \eta_{\text{mixed}} = 0.03, \eta_{\text{multi}} = 0.03, \theta_0 = 0, \lambda_{1,0} = \lambda_{2,0} = 0.1, Q(\cdot) = 0$ for all heads at $t = 0$.

H.1 Evaluation Metrics

Expected reward. We report the *on-policy* expected reward $J_r = \sigma(\theta)\,r(a_1) + (1 - \sigma(\theta))\,r(a_2)$ as a function of steps.

Constraint violation. Instantaneous expected violation is

$$\mathrm{Viol}_t \;:=\; \max\big(0,\, J_{c_1}(\pi_{\theta_t}) - d_1\big) + \max\big(0,\, J_{c_2}(\pi_{\theta_t}) - d_2\big) \;=\; \max(0,\, 1 - \sigma(\theta_t) - 0.5) + \max(0,\, \sigma(\theta_t) - 0.5). \tag{37}$$

Unconditional gradient alignment. We compute a *moving* Pearson correlation between the estimated actor gradient $\widehat{g}_t$ (from equation 5.1 with the appropriate critic readout) and the true gradient g_t, using a centered window of width $w = 201$ with boundary normalization:

$$\mathrm{corr}_t(\widehat{g}, g) \;=\; \frac{\mathrm{Cov}_t(\widehat{g}, g)}{\sqrt{\mathrm{Var}_t(\widehat{g})\,\mathrm{Var}_t(g)}}, \tag{38}$$

with $\mathrm{Cov}_t(\cdot,\cdot),\ \mathrm{Var}_t(\cdot)$ computed over the window and normalized by its effective length.

Conditional gradient alignment. Same as equation 38, but *restricted* to timesteps in the window where the ground-truth magnitude exceeds a threshold $\varepsilon = 10^{-3}$:

$$\mathrm{corr}_t^{\mathrm{cond}}(\widehat{g}, g) \;=\; \mathrm{corr}\big(\{\widehat{g}_\tau : |g_\tau| > \varepsilon\},\ \{g_\tau : |g_\tau| > \varepsilon\}\big) \tag{39}$$

This metric emphasizes periods with a meaningful learning signal, computed with boundary-normalized counts; windows with < 5 effective samples are masked.

Dual oscillation magnitude. We quantify multiplier oscillations via the *moving standard deviation* of the signed gap $|\lambda_2 - \lambda_1|$, again using a boundary-normalized window of width $w = 201$:

$$\mathrm{Osc}_t \;=\; \sqrt{\max\Big(0,\ \mathbb{E}_t[\Delta^2] - \big(\mathbb{E}_t[\Delta]\big)^2\Big)}, \qquad \Delta_\tau := |\lambda_{2,\tau} - \lambda_{1,\tau}|. \tag{40}$$

Smoothing (for curves) and uncertainty bands. For reward and violation we plot *boundary-normalized* running means:

$$\widetilde{x}_t \;=\; \frac{\sum_{\tau=t-\lfloor w/2\rfloor}^{t+\lfloor w/2\rfloor} x_\tau}{\#\{\tau \text{ inside range}\}}, \tag{41}$$

then average $\widetilde{x}_t$ across seeds and show ± 1 standard deviation bands across seeds.

H.2 Timescale Ablation (Critic vs. Dual)

To mirror the theory's timescale conclusions, we sweep critic and dual learning rates on a grid:

$$\eta \in \{0.01,\, 0.03,\, 0.10\}, \qquad \beta \in \{0.005,\, 0.02,\, 0.08\},$$

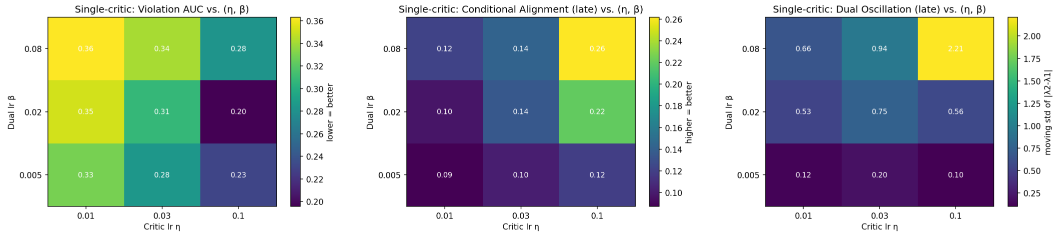
holding the actor step $\alpha = 0.02$ fixed. For each (η, β), we run the *mixed-critic* variant for $T = 5000$ steps with 8 seeds and report:

1. **Violation AUC:** $\sum_{t=1}^{T} \mathrm{Viol}_t / T$,
2. **Late conditional alignment:** mean of equation 39 over the last 500 steps,
3. **Late dual oscillation:** mean of equation 40 over the last 500 steps.

Results are visualized as heatmaps over (η, β).

H.3 Compute, Randomization, and Reproducibility

All runs are CPU-only and complete within seconds. Random seeds $s \in \{1000, \ldots, 1000 + S - 1\}$ control action sampling only (initial parameters are deterministic). Each figure reports the mean across seeds with ± 1 standard deviation. We save raw arrays (per-seed trajectories for reward, violation, gradients, and multipliers) to a serialized file for exact reproduction of all plots.



(a) Violation AUC (b) Late conditional alignment (c) Late dual oscillation

Figure 6: Performance for CMDP Bandit.

I Algorithm: Dedicated-Critic PPO-Lag

Algorithm 1 Dedicated-Critic Lagrangian PPO (multi-constraint, single-constraint is special case)

Initialize policy parameters θ; Initialize reward critic parameters ϕ_r; Initialize cost critic parameters ϕ_{c_i} for $i = 1, \dots, m$; Initialize dual variables $\lambda_i \leftarrow \lambda_{\text{init}} \geq 0$ for $i = 1, \dots, m$
for iteration $k = 0, 1, 2, \dots$ **do**
  Reset buffers $\{s_t, a_t, r_t, c_t^{(i)}, \text{done}_t, \log \pi_t^{\text{old}}, V_t^r, V_t^{c_i}\}_{t=0}^{T-1}$; $s_0 \leftarrow \mathcal{E}.\text{reset}()$
  for $t = 0, \dots, T-1$ **do**
    Sample $a_t \sim \pi_\theta(\cdot \mid s_t)$; $\log \pi_t^{\text{old}} \leftarrow \log \pi_\theta(a_t \mid s_t)$; $V_t^r \leftarrow V_r(s_t; \phi_r)$
    for $i = 1, \dots, m$ **do**
      $V_t^{c_i} \leftarrow V_{c_i}(s_t; \phi_{c_i})$
    end for
    $s_{t+1}, r_t, \{c_t^{(i)}\}_{i=1}^m, \text{done}_t \leftarrow \mathcal{E}.\text{step}(a_t)$
    Store $(s_t, a_t, r_t, \{c_t^{(i)}\}, \text{done}_t, \log \pi_t^{\text{old}}, V_t^r, \{V_t^{c_i}\})$ in buffer
  end for
  if s_T is terminal **then**
    $V_T^r \leftarrow 0, \quad V_T^{c_i} \leftarrow 0 \ \ \forall i$
  else
    $V_T^r \leftarrow V_r(s_T; \phi_r)$
    $V_T^{c_i} \leftarrow V_{c_i}(s_T; \phi_{c_i}) \ \ \forall i$
  end if
  Initialize $A_T^r \leftarrow 0$ and $A_T^{c_i} \leftarrow 0$ for all i
  for $t = T-1, \dots, 0$ **do**
    $\delta_t^r \leftarrow r_t + \gamma(1 - \text{done}_{t+1}) V_{t+1}^r - V_t^r$
    $A_t^r \leftarrow \delta_t^r + \gamma \lambda_{\text{GAE}} (1 - \text{done}_{t+1}) A_{t+1}^r$
    for $i = 1, \dots, m$ **do**
      $\delta_t^{c_i} \leftarrow c_t^{(i)} + \gamma(1 - \text{done}_{t+1}) V_{t+1}^{c_i} - V_t^{c_i}$
      $A_t^{c_i} \leftarrow \delta_t^{c_i} + \gamma \lambda_{\text{GAE}} (1 - \text{done}_{t+1}) A_{t+1}^{c_i}$
    end for
    $R_t^r \leftarrow A_t^r + V_t^r$; $R_t^{c_i} \leftarrow A_t^{c_i} + V_t^{c_i} \ \ \forall i$
  end for
  for $t = 0, \dots, T-1$ **do**
    $A_t^{\text{Lag}} \leftarrow A_t^r - \sum_{i=1}^m \lambda_i A_t^{c_i}$
  end for
  for PPO epoch $e = 1, \dots, K$ **do**
    for minibatch $\mathcal{M}$ **do**
      for $(s_t, a_t, \log \pi_t^{\text{old}}, A_t^{\text{Lag}}) \in \mathcal{M}$ **do**
        $\log \pi_t \leftarrow \log \pi_\theta(a_t \mid s_t)$
        $\rho_t \leftarrow \exp(\log \pi_t - \log \pi_t^{\text{old}})$
        $\hat{L}_t \leftarrow \min\left(\rho_t A_t^{\text{Lag}}, \ \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon) A_t^{\text{Lag}}\right)$
      end for
      $L_\pi \leftarrow -\frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \hat{L}_t$
      Update $\theta \leftarrow \theta - \alpha_\pi \nabla_\theta L_\pi$
      $L_V^r \leftarrow \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \left(V_r(s_t; \phi_r) - R_t^r\right)^2$
      for $i = 1, \dots, m$ **do**
        $L_V^{c_i} \leftarrow \frac{1}{|\mathcal{M}|} \sum_{t \in \mathcal{M}} \left(V_{c_i}(s_t; \phi_{c_i}) - R_t^{c_i}\right)^2$
      end for
      $L_V \leftarrow L_V^r + \sum_{i=1}^m L_V^{c_i}$
      Update $\phi_r, \{\phi_{c_i}\} \leftarrow \phi_r, \{\phi_{c_i}\} - \alpha_V \nabla L_V$
    end for
  end for
  $\theta_{\text{old}} \leftarrow \theta$
  for $i = 1, \dots, m$ **do**
    Estimate average cost $\widehat{J}_{c_i} \leftarrow \frac{1}{T} \sum_{t=0}^{T-1} c_t^{(i)}$
    $\lambda_i \leftarrow \max\left(0, \ \lambda_i + \alpha_\lambda (\widehat{J}_{c_i} - d_i)\right)$
  end for
end for

J EXPERIMENT DETAILS - CASE 1

J.1 SYSTEM DESCRIPTION

We model a radial distribution network with high rooftop PV penetration, where a set of community battery energy storage systems (CBESSs) are coordinated to ensure operational safety and efficiency. Each CBESS is constrained by efficiency, power, and state-of-charge (SOC) limits, and can exchange energy with the upstream grid within trading bounds, incurring both trading and degradation costs. When storage is saturated, PV curtailment at the bus level is introduced with fairness considerations to avoid disproportionate restrictions. The distribution network is described using the LinDistFlow approximation, including power balance, voltage regulation, and branch thermal limits. Voltage violations and line loading are penalized in the objective. The overall scheduling problem minimizes the aggregated penalties and costs associated with CBESS operations, grid trading, and PV curtailment fairness.

We consider a radial distribution network $(\mathcal{N}, \mathcal{L})$ operated by a DNSP over intra-day periods $t \in \mathcal{T} = \{1, \ldots, T\}$. The system model consists of CBESS operation, PV curtailment, and PDN-level constraints.

J.1.1 CBESS

Let $\mathcal{M}$ denote the set of CBESSs. Each CBESS $m \in \mathcal{M}$ is connected to bus $\xi(m)$, with charging/discharging efficiencies $(\eta_m^{\text{ch}}, \eta_m^{\text{dis}})$, charging and discharging limits $(\overline{P}_m^{\text{ch}}, \overline{P}_m^{\text{dis}})$, and SOC range $[\underline{SOC}_m, \overline{SOC}_m]$. The charging/discharging power are $p_{m,t}^{\text{ch}}$ and $p_{m,t}^{\text{dis}}$, the reactive support is $q_{m,t}^{\text{CB}}$, and stored energy is $E_{m,t}$ with capacity E_m^{Cap}. Their dynamics are:

$$E_{m,t+1} = E_{m,t} + \eta_m^{\text{ch}} p_{m,t}^{\text{ch}} \Delta t - \frac{1}{\eta_m^{\text{dis}}} p_{m,t}^{\text{dis}} \Delta t, \tag{42a}$$

$$SOC_{m,t} = \frac{E_{m,t}}{E_m^{\text{Cap}}}, \quad \underline{SOC}_m \leq SOC_{m,t} \leq \overline{SOC}_m, \tag{42b}$$

$$0 \leq p_{m,t}^{\text{ch}} \leq \overline{P}_m^{\text{ch}}, \quad 0 \leq p_{m,t}^{\text{dis}} \leq \overline{P}_m^{\text{dis}}, \tag{42c}$$

$$p_{m,t}^{\text{ch}} \cdot p_{m,t}^{\text{dis}} = 0, \tag{42d}$$

$$(p_{m,t}^{\text{dis}} - p_{m,t}^{\text{ch}})^2 + (q_{m,t}^{\text{CB}})^2 \leq (S_m^{\text{CB}})^2, \tag{42e}$$

$$E_{m,0} = E_m^{\text{init}}. \tag{42f}$$

CBESSs also trade with the main grid through a ratio $\rho_{m,t}^{\text{trade}} \in [0, 1]$. With buy/sell prices $(\phi_t^{\text{buy}}, \phi_t^{\text{sell}})$, the trading cost is:

$$f_t^{\text{ET}} = \sum_{m \in \mathcal{M}} f_{m,t}^{\text{trade}}, \tag{43a}$$

$$f_{m,t}^{\text{trade}} = \phi_t^{\text{buy}} p_{m,t}^{\text{ch}} \rho_{m,t}^{\text{trade}} - \phi_t^{\text{sell}} p_{m,t}^{\text{dis}} \rho_{m,t}^{\text{trade}}, \tag{43b}$$

$$0 \leq p_{m,t}^{\text{ch}} \rho_{m,t}^{\text{trade}} \leq \overline{P}_m^{\text{trade,ch}}, \quad 0 \leq p_{m,t}^{\text{dis}} \rho_{m,t}^{\text{trade}} \leq \overline{P}_m^{\text{trade,dis}}. \tag{43c}$$

Battery degradation is approximated linearly:

$$f_t^{\text{BD}} = \sum_{m \in \mathcal{M}} c_m^{\text{deg}} (p_{m,t}^{\text{ch}} + p_{m,t}^{\text{dis}}), \tag{44a}$$

where $c_m^{\text{deg}} > 0$ is the degradation cost coefficient.

J.1.2 PV CURTAILMENT

When all CBESSs are full, PV generation is curtailed via ratio $\gamma_{i,t} \in [0, 1]$:

$$\tilde{p}_{i,t}^{\text{PV}} = (1 - \gamma_{i,t}) p_{i,t}^{\text{PV}}, \tag{45a}$$

$$0 \leq \gamma_{i,t} \leq 1. \tag{45b}$$

Fairness is enforced by comparing each bus's curtailed ratio π_i^{curt} with its proportional target π_i^{tar}:

$$f^{\text{PVF}} = \sum_{i \in \mathcal{N}} (\pi_i^{\text{curt}} - \pi_i^{\text{tar}})^2. \tag{46}$$

J.1.3 PDN

The PDN is described by lossless LinDistFlow. For each branch $(i, j) \in \mathcal{L}$:

$$p_{ij,t} = \sum_{k:(j,k)\in\mathcal{L}} p_{jk,t} + p_{j,t}^{\text{load}} - \tilde{p}_{j,t}^{\text{PV}} - \sum_{m:\xi(m)=j} (p_{m,t}^{\text{dis}} - p_{m,t}^{\text{ch}}), \tag{47a}$$

$$q_{ij,t} = \sum_{k:(j,k)\in\mathcal{L}} q_{jk,t} + q_{j,t}^{\text{load}} - \sum_{m:\xi(m)=j} q_{m,t}^{\text{CB}}. \tag{47b}$$

Voltage drop is given by:

$$V_{j,t} = V_{i,t} - 2(r_{ij}p_{ij,t} + x_{ij}q_{ij,t}), \tag{48}$$

with bounds $\underline{V} \le V_{i,t} \le \overline{V}$. Penalties for voltage violations are:

$$f_t^{\text{VD}} = \sum_{i\in\mathcal{N}} \big([V_{i,t} - \overline{V}]^+ + [\underline{V} - V_{i,t}]^+\big), \tag{49a}$$

$$f_t^{\text{VN}} = \sum_{i\in\mathcal{N}} \mathbb{I}(V_{i,t} > \overline{V} \vee V_{i,t} < \underline{V}). \tag{49b}$$

Line loading penalty is:

$$f_t^{\text{LL}} = \sum_{(i,j)\in\mathcal{L}} r_{ij} \frac{p_{ij,t}^2 + q_{ij,t}^2}{V_0^2}, \tag{50a}$$

$$p_{ij,t}^2 + q_{ij,t}^2 \le \overline{S}_{ij}^2. \tag{50b}$$

J.1.4 Objective

The goal is to coordinate CBESS operation under PV-rich PDNs to ensure network safety and efficiency. At each time step, CBESSs decide charging/discharging and grid trading ratios. The optimization problem is:

$$\min_{p^{\text{ch}}, p^{\text{dis}}, q_{m,t}^{\text{CB}}, \rho^{\text{trade}}} \sum_{t\in\mathcal{T}} \big(f_t^{\text{VD}} + f_t^{\text{VN}} + f_t^{\text{LL}} + f_t^{\text{BD}} + f_t^{\text{ET}}\big) + f^{\text{PVF}}, \tag{51a}$$

$$\text{s.t.} \quad \text{equation 42, equation 43c, equation 45, equation 47, and equation 48.} \tag{51b}$$

J.2 CMDP Modeling with Dedicated-Critic Lagrangian RL

We cast the CBESS coordination as a constrained Markov decision process (CMDP) $\big(\mathcal{S}, \mathcal{A}, P, r, \{c_i\}_{i=1}^m, \gamma, \{d_i\}_{i=1}^m\big)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(\cdot|s,a)$ the transition kernel, $\gamma \in (0, 1)$ the discount factor, $r(s, a)$ the reward, and $c_i(s, a)$ the cost signal for constraint i with threshold d_i. Given a stochastic policy $\pi_\theta(a|s)$, define the discounted returns

$$J_r(\pi_\theta) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big], \quad J_{c_i}(\pi_\theta) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t)\Big]. \tag{52}$$

The CMDP objective is

$$\max_\theta \; J_r(\pi_\theta) \quad \text{s.t.} \quad J_{c_i}(\pi_\theta) \le d_i, \;\; i = 1, \dots, m. \tag{53}$$

Reward & costs from the PDN model. Let the instantaneous penalties/costs at time t be those defined in the system model: $f_t^{\text{VD}}, f_t^{\text{VN}}, f_t^{\text{LL}}, f_t^{\text{BD}}, f_t^{\text{ET}}$ and the daily PV-curtailment fairness term f^{PVF}. A practical partition is:

$$r(s_t, a_t) = -\big(\alpha_{\text{BD}} f_t^{\text{BD}} + \alpha_{\text{ET}} f_t^{\text{ET}}\big), \tag{54}$$

$$c_1(s_t, a_t) = f_t^{\text{VD}}, \quad c_2(s_t, a_t) = f_t^{\text{VN}}, \quad c_3(s_t, a_t) = f_t^{\text{LL}}, \tag{55}$$

and an episodic fairness constraint

$$C_4(\tau) \triangleq f^{\text{PVF}} \;\text{ with }\; \mathbb{E}_\pi[C_4(\tau)] \le d_4, \tag{56}$$

where τ denotes a full episode (day). If desired, f^{PVF} can be spread as a per-step density $c_4(s_t, a_t)$ so that $\sum_t \gamma^t c_4(s_t, a_t)$ recovers the same daily target. The weights $\alpha_{\text{BD}}, \alpha_{\text{ET}} > 0$ reflect economic preferences. Alternative partitions (e.g., moving f^{ET} into constraints) are also supported without changing the derivations below.

Lagrangian relaxation with per-constraint critics. Introduce multipliers $\lambda = (\lambda_1, \ldots, \lambda_m) \succeq 0$ and define

$$\mathcal{L}(\theta, \lambda) = J_r(\pi_\theta) - \sum_{i=1}^{m} \lambda_i \big(J_{c_i}(\pi_\theta) - d_i\big). \tag{57}$$

We perform the standard primal–dual updates:

$$\theta \text{ update: } \nabla_\theta \mathcal{L}(\theta, \lambda) = \nabla_\theta J_r(\pi_\theta) - \sum_{i=1}^{m} \lambda_i \nabla_\theta J_{c_i}(\pi_\theta), \tag{58}$$

$$\lambda \text{ update: } \lambda_i \leftarrow \Pi_{[0, \lambda_{\max}]}\Big(\lambda_i + \beta\,[J_{c_i}(\pi_\theta) - d_i]\Big), \tag{59}$$

where Π denotes projection to stabilize λ.

Signal-wise value functions and advantages. For each signal $x \in \{r, c_1, \ldots, c_m\}$ define

$$Q_\pi^x(s, a) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t x(s_t, a_t) \,\Big|\, s_0{=}s, a_0{=}a\Big], \tag{60}$$

$$V_\pi^x(s) = \mathbb{E}_{a\sim\pi}[Q_\pi^x(s, a)], \;\; A_\pi^x(s, a) = Q_\pi^x(s, a) - V_\pi^x(s). \tag{61}$$

Using the policy score function, the actor gradient becomes

$$\nabla_\theta \mathcal{L}(\theta, \lambda) = \mathbb{E}_\pi\big[\nabla_\theta \log \pi_\theta(a|s)\; \underbrace{\big(A_\pi^r(s, a) - \sum_{i=1}^{m} \lambda_i A_\pi^{c_i}(s, a)\big)}_{\widetilde{A}_\pi(s,a)}\,\big]. \tag{62}$$

Per-constraint critics. We learn one critic per signal $x \in \{r, c_1, \ldots, c_m\}$ with parameters ω_x:

$$Q_{\omega_x}(s, a) \approx Q_\pi^x(s, a), \qquad \delta_t^x = x_t + \gamma\, Q_{\omega_x}(s_{t+1}, a_{t+1}) - Q_{\omega_x}(s_t, a_t), \tag{63}$$

and minimize $\mathbb{E}[(\delta_t^x)^2]$ (or use GAE to reduce variance). Advantages are estimated by A_t^x (e.g., GAE(λ)) and plugged into equation 62.

PPO-style actor (with dedicated-critic advantage). Let $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ and $\widetilde{A}_t = A_t^r - \sum_i \lambda_i A_t^{c_i}$. The clipped surrogate is

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}\Big[\min\big(r_t(\theta)\,\widetilde{A}_t,\; \text{clip}(r_t(\theta), 1{-}\epsilon, 1{+}\epsilon)\,\widetilde{A}_t\big)\Big] + \eta\,\mathbb{E}[\mathcal{H}(\pi_\theta(\cdot|s_t))], \tag{64}$$

where $\mathcal{H}$ is policy entropy and $\eta \ge 0$.

Episodic fairness constraint. If keeping f^{PVF} as episodic, use the per-episode estimator $\widehat{J}_{c_4} = \frac{1}{N}\sum_{k=1}^{N} C_4(\tau^{(k)})$ in equation 82. A practical alternative is to define a per-step density $c_4(s_t, a_t)$ whose discounted sum equals the daily fairness value, enabling a standard critic update as in equation 87.

Table 1: Key hyperparameters, reward, and CMDP constraints for the energy management case study.

Category	Hyperparameter / Term	Value	Notes / Definition
Reward & constraints			
Reward r_t	$-f_t^{\text{ET}}$	–	$f_t^{\text{ET}} = \sum_m \left(\phi_t^{\text{buy}} p_{m,t}^{\text{ch}} \rho_{m,t}^{\text{trade}} - \phi_t^{\text{sell}} p_{m,t}^{\text{dis}} \rho_{m,t}^{\text{trade}}\right)$
Constraint 1 c_t^{VN}	$\frac{1}{\lvert\mathcal{N}\rvert}\sum_i \mathbb{I}\big(V_{i,t} \notin [\underline{V}, \overline{V}]\big)$	$[0,1]$	Count of voltage violations (normalized by bus count)
Constraint 2 c_t^{VD}	$\sum_i\big([V_{i,t}-\overline{V}]^+ + [\underline{V}-V_{i,t}]^+\big)$	–	Degree of voltage violation (no extra scaling)
Constraint 3 c_t^{LL}	$\sum_{(i,j)}[\,\ell_{ij,t} - \tau^{\text{line}}\,]^+$	–	Line thermal overload beyond threshold (p.u. or %)
Constraint 4 c_t^{BD}	$\dfrac{\sum_m (p_{m,t}^{\text{ch}} + p_{m,t}^{\text{dis}})\Delta t}{\lvert\mathcal{M}\rvert \overline{P}^{\text{CB}} \Delta t}$	$[0,1]$	Battery degradation (throughput, normalized)
Constraint 5 c_t^{PVF}	$\dfrac{\mathrm{var}(\{\gamma_{i,t}\}_{i\neq 0})}{0.25}$	$[0,1]$	PV curtailment unfairness (variance normalized by max 0.25)
Lag-PPO constraint	$c_t^{\text{VN}} + c_t^{\text{VD}} + c_t^{\text{LL}} + c_t^{\text{BD}} + c_t^{\text{PVF}}$	-	Summation of all constraints
General training parameters			
Learning rate α/η	–	3e-4	Shared by actor/critic
Clip coefficient	–	0.2	Ratio clipping $[1-\epsilon, 1+\epsilon]$
Target KL	–	0.015	Early stop when approx-KL exceeds threshold
Value loss coeff.	–	0.5	Weight on value loss
Entropy coeff.	–	0.0	Entropy regularization
Grad norm clip	–	0.5	Global gradient clipping
Hidden sizes	–	(256, 256)	MLP for actor/critic
Init log-std	–	-0.5	Gaussian policy init
Discount γ, GAE	–	0.99, 0.95	For returns and advantages
Dual learning rate λ	–	5e-3	Step size for dual updates in Lagrangian RL paradigm
λ init / max	–	0.0 / 10^4	Projected to $[0, \lambda_{\max}]$
Training schedule & environment			
PPO episodes	–	2000	Total training episodes
Steps / episode	–	288	$\Delta t = 5$ min $\Rightarrow$ one day per episode
Env time step	–	5 min	Day length = 288 steps

Concrete instantiation for this problem. With equation 74–equation 56, we have $m \in \{3, 4\}$ constraints:

$$\text{Critics: } Q_{\omega_r} \text{ for reward, } \; Q_{\omega_{c_1}}, Q_{\omega_{c_2}}, Q_{\omega_{c_3}} \text{ (and } Q_{\omega_{c_4}} \text{ if episodic fairness is densified);} \tag{65}$$

$$\text{Advantage: } \widetilde{A}_t = A_t^r - \lambda_1 A_t^{c_1} - \lambda_2 A_t^{c_2} - \lambda_3 A_t^{c_3} \;(-\lambda_4 A_t^{c_4} \text{ if used);} \tag{66}$$

$$\text{Dual: } \lambda_i \leftarrow \Pi_{[0,\lambda_{\max}]}\Big(\lambda_i + \beta\,[\widehat{J}_{c_i} - d_i]\Big), \quad i = 1, \ldots, m. \tag{67}$$

Notes on stability and practice. (i) Use separate target networks or Polyak averaging for each critic to stabilize TD. (ii) Normalize every A_t^x before forming $\widetilde{A}_t$ to balance scales across constraints. (iii) Choose d_i from engineering limits (e.g., allowable daily voltage violation budget, line loading budget); start with conservative d_i then relax. (iv) Bound λ via projection or log-parameterization to avoid runaway dual ascent; optionally add a small L2 penalty on λ. (v) For mixed episodic/step constraints, update episodic multipliers once per episode and stepwise ones per minibatch.

J.3 Experimental Parameters

Symbols. The key parameters of the power system management case study are provided in Table 3. $\phi_t^{\text{buy}}, \phi_t^{\text{sell}}$: upstream buy/sell prices; $\rho_{m,t}^{\text{trade}} \in [0,1]$: trading ratio for CBESS m; $[x]^+ = \max\{x, 0\}$; $\mathbb{I}(\cdot)$: indicator; $\underline{V}, \overline{V}$: voltage bounds (e.g., $[1-\nu, 1+\nu]$ p.u., $\nu > 0$); $V_{i,t}$: bus-i voltage; $\ell_{ij,t}$:

loading of line (i, j) (p.u. or %); τ^{line}: overload threshold (default 0); $\gamma_{i,t} \in [0, 1]$: PV curtailment ratio at bus i; Δt: step duration (5 min); T: daily horizon (288 steps); $|\mathcal{N}|$: number of buses; $|\mathcal{M}|$: number of CBESS; $\overline{P}^{\text{CB}}$: nameplate active-power rating for normalization.

J.4 Two-Tiered Statistics

The two-tiered statistical results in Table 2 highlight a clear trade-off between economic performance and system safety. Specifically, the DC-Lag-PPO variant achieves substantial improvements across all five constraint metrics. The violation ratio (c1) and violation degree (c2) of bus voltages are reduced by approximately 33% and 52%, respectively, while the line loading rate (c3) decreases by 11.6%. Similarly, the battery degradation cost (c4) drops by 46%, and the PV curtailment unfairness (c5) improves by 52.7%. These reductions indicate that DC-Lag-PPO enforces network security and operational fairness more effectively than the baseline Lag-PPO.

In contrast, the reward, which reflects the economic cost, declines significantly (-87.2%). Since higher reward is preferred, this suggests that DC-Lag-PPO sacrifices economic efficiency to achieve stronger compliance with safety and fairness constraints. The mechanism is likely due to more conservative charging, discharging, and trading behaviors encouraged by the tightened constraint handling.

In terms of stability, the across-run standard deviations of c1, c2, and c3 decrease considerably, demonstrating more consistent performance in voltage and line-loading metrics. However, the standard deviation of PV curtailment fairness (c5) increases, implying reduced consistency across different runs in this aspect. This suggests that while DC-Lag-PPO reliably improves most safety indicators, its fairness outcomes may vary depending on specific training trajectories.

Overall, DC-Lag-PPO demonstrates its effectiveness as a safer policy with stronger constraint satisfaction, albeit at the cost of economic performance. Future work may seek to balance this trade-off by tuning constraint thresholds, adjusting the dual update step size, or normalizing advantage signals across constraints to prevent overly conservative policies.

Table 2: Two-tiered test statistics, where across-run mean $\pm$ across-run std; The higher reward is better, while the lower constraints are better. Δ = (DC-Lag-PPO $-$ Lag-PPO). Positive improvement % is computed as Lag-PPO $-$ DC-Lag-PPO$/$Lag-PPO $\times$ 100%.

Metric	Lag-PPO (n=9)	DC-Lag-PPO (n=9)	Δ	Improvement %
Economic cost (reward)	25.95 $\pm$ 63.25	3.32 $\pm$ 34.68	**-22.64**	**-87.23%**
Volt. vio. ratio (c1)	62.96 $\pm$ 29.41	42.17 $\pm$ 11.61	**-20.80**	**+33.03%**
Volt. vio. degree (c2)	54.77 $\pm$ 36.31	26.47 $\pm$ 10.69	**-28.30**	**+51.67%**
Line load rate (c3)	0.405 $\pm$ 0.062	0.358 $\pm$ 0.024	**-0.047**	**+11.59%**
Battery degradation (c4)	32.23 $\pm$ 13.47	17.35 $\pm$ 11.21	**-14.88**	**+46.17%**
PV curt. unfairness (c5)	210.95 $\pm$ 23.09	99.82 $\pm$ 62.79	**-111.13**	**+52.68%**

J.5 Training Curves

As shown in Fig. 7 - 12, the training curves across multiple runs consistently highlight the strengths of DC-Lag-PPO in terms of constraint satisfaction. While the reward trajectories show that DC-Lag-PPO tends to converge to lower economic returns compared to the baseline Lag-PPO, the improvement in constraint metrics is substantial.

First, the voltage violation metrics (both ratio and degree) are markedly reduced under DC-Lag-PPO. The curves demonstrate faster convergence to lower levels of violations and maintain stability across episodes, especially in Fig. 11 and 12. This indicates that the dual-critic structure effectively penalizes unsafe voltage states, leading to more secure system operation.

Second, the line loading rates remain consistently lower for DC-Lag-PPO. Although the difference is modest compared to voltage metrics, the reduced variance in the curves reflects more stable utilization of line capacity, especially in Fig. 8 and 10.

Third, battery degradation under DC-Lag-PPO is substantially lower. The curves show that the algorithm learns to avoid excessive charging and discharging cycles, which not only improves system longevity but also reduces long-term operational costs.

Finally, PV curtailment unfairness also benefits significantly from DC-Lag-PPO. Although variance is occasionally higher across runs, the overall trajectory converges to much lower unfairness compared to Lag-PPO. This suggests that DC-Lag-PPO is able to balance curtailment more evenly across the network, enhancing fairness.

Overall, the training results confirm that DC-Lag-PPO enforces operational safety and fairness more effectively than the baseline. The cost of this improvement is a reduction in reward, implying that the method prioritizes constraint satisfaction over immediate economic gains. From a practical perspective, this trade-off can be acceptable or even desirable in safety-critical power systems, where violations may carry severe penalties or risks.

Future extensions could explore adaptive balancing mechanisms, such as dynamic adjustment of dual learning rates or reward re-weighting—to recover part of the economic performance while maintaining the strong safety guarantees observed here.

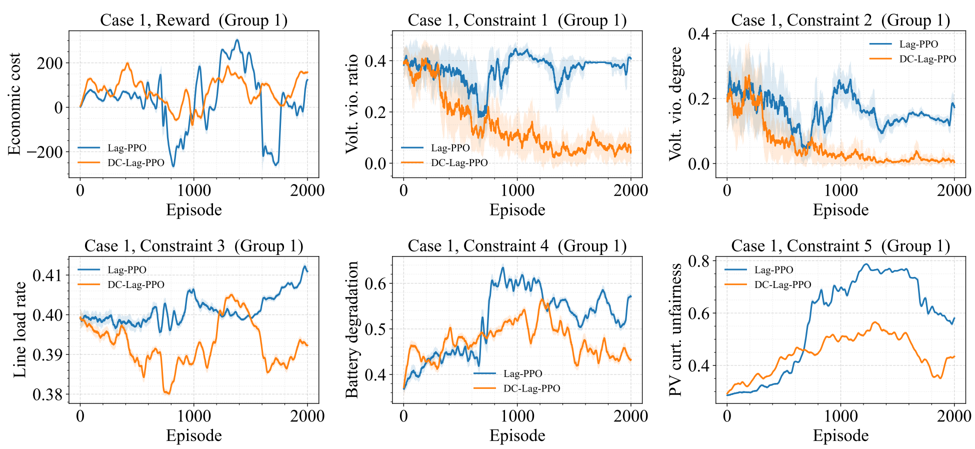


Figure 7: Training curves on Lagrangian cost threshold set: [9,9,0.1,30,30].

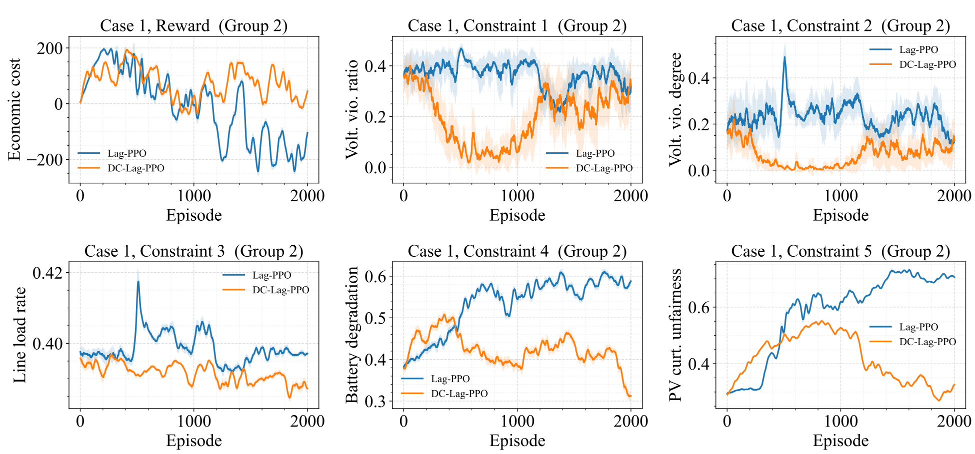


Figure 8: Training curves on Lagrangian cost threshold set: [12,12,0.1,30,30].

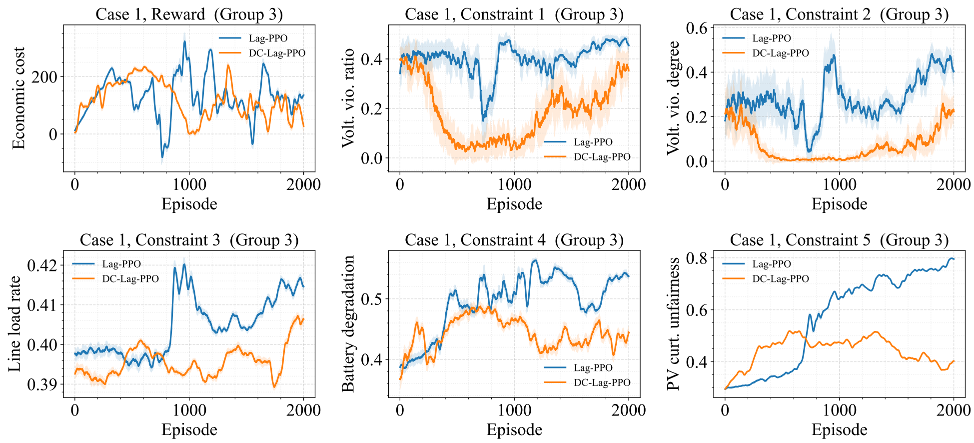


Figure 9: Training curves on Lagrangian cost threshold set: [15,15,0.1,30,30].

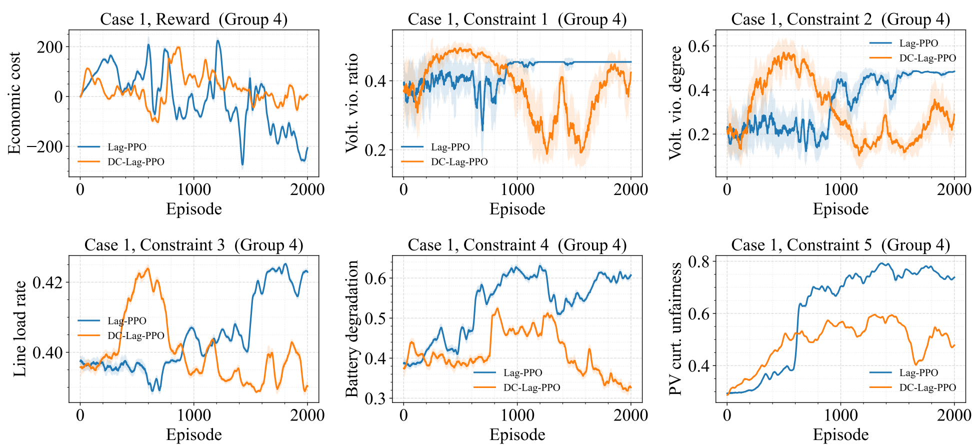


Figure 10: Training curves on Lagrangian cost threshold set: [18,18,0.1,20,30].

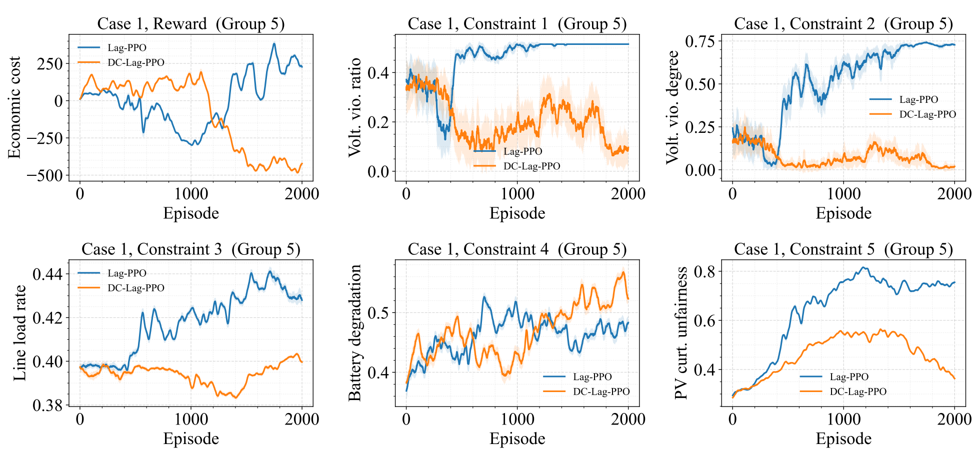


Figure 11: Training curves on Lagrangian cost threshold set: [9,9,0.1,30,20].

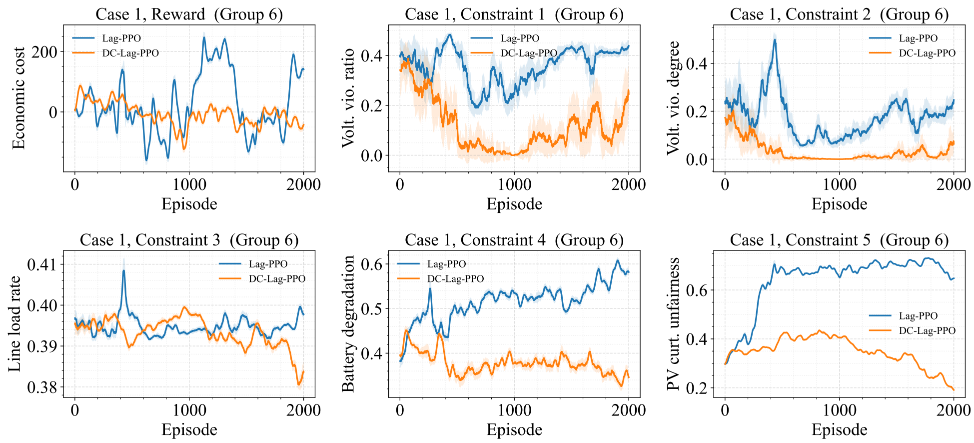


Figure 12: Training curves on Lagrangian cost threshold set: [7,7,0.1,15,15].

Additionally, J_c curves are demonstrated in Fig. 13-18. Across all parameter settings, the plots consistently show a clear difference between Lag-PPO and our DC-Lag-PPO. The J_c curve of Lag-PPO represents the summed constraint cost, and it frequently drifts far above the allowed threshold, exhibiting large fluctuations during training (e.g., group 1 and group 5). This indicates that a single shared Lagrange multiplier cannot effectively regulate multiple heterogeneous constraints.

In contrast, DC-Lag-PPO decomposes the constraint cost into five independent components, each with its own critic. The corresponding J_c curves tightly track their respective thresholds across all settings, for large thresholds (e.g., 18, 20, 30) and even for very small ones (d = 0.1). This demonstrates precise constraint satisfaction and significantly improved stability. Therefore, the decomposed multi-critic structure is fundamentally more effective for enforcing multi-constraint safety compared to the single-critic Lag-PPO.

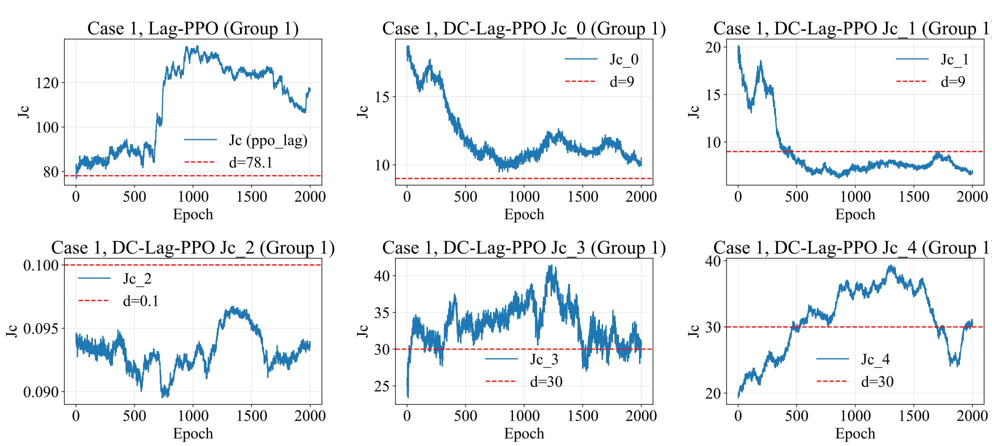


Figure 13: Training curves of J_c on Lagrangian cost threshold set: [9,9,0.1,30,30].

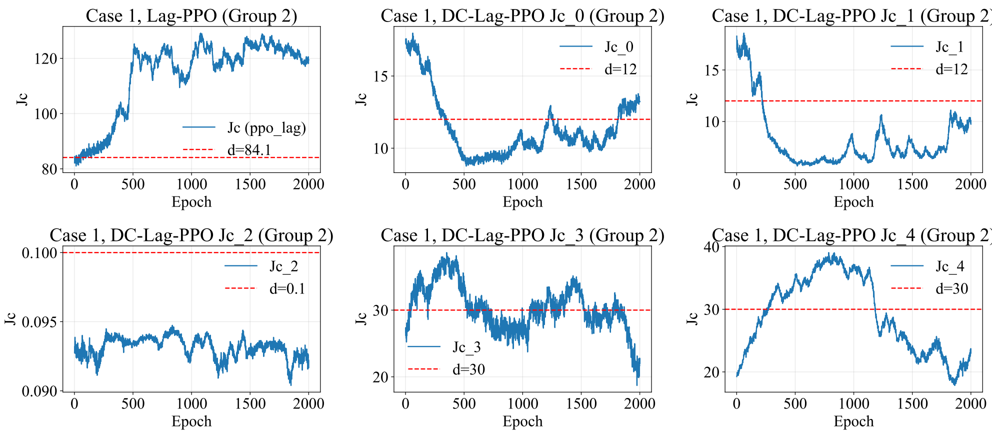


Figure 14: Training curves of J_c on Lagrangian cost threshold set: [12,12,0.1,30,30].

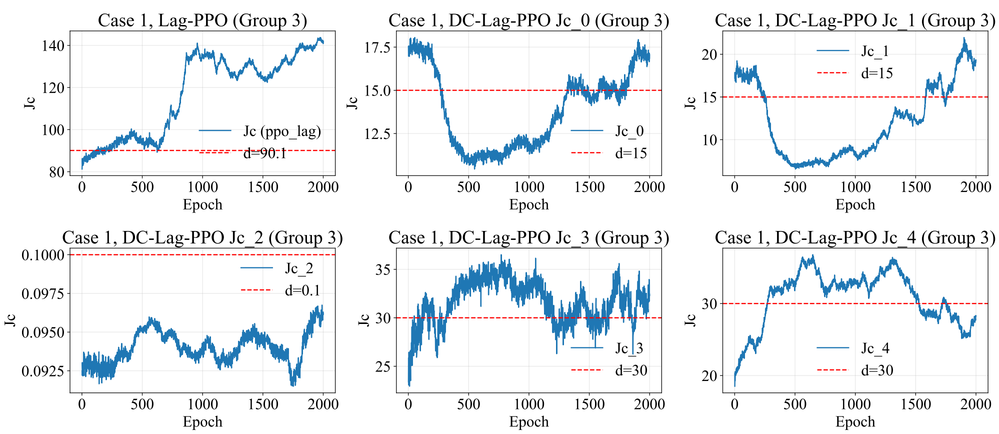


Figure 15: Training curves of J_c on Lagrangian cost threshold set: [15,15,0.1,30,30].

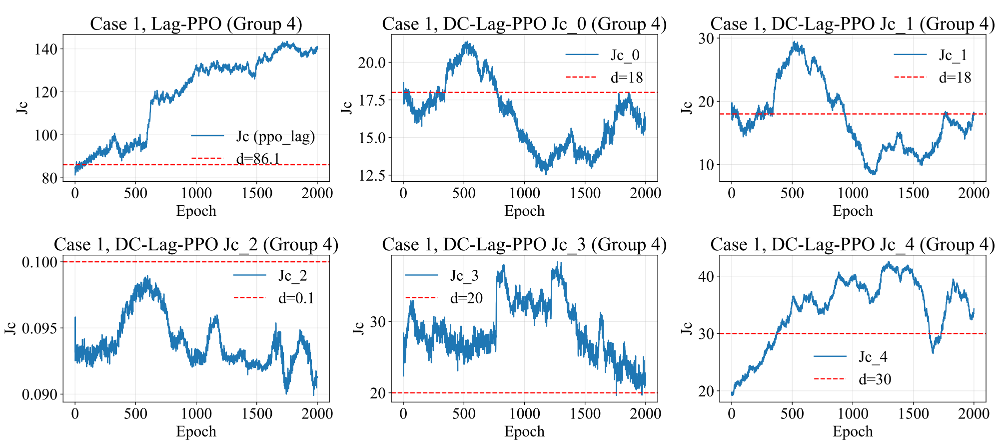


Figure 16: Training curves of J_c on Lagrangian cost threshold set: [18,18,0.1,20,30].

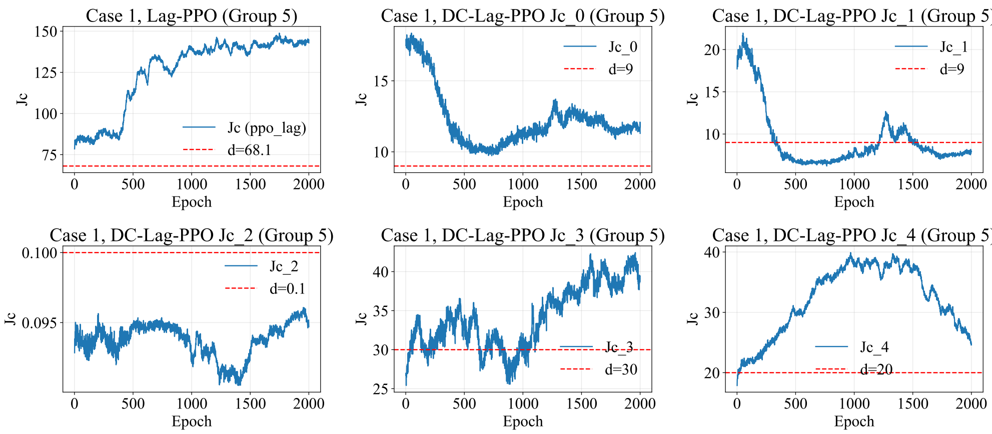


Figure 17: Training curves of J_c on Lagrangian cost threshold set: [9,9,0.1,30,20].

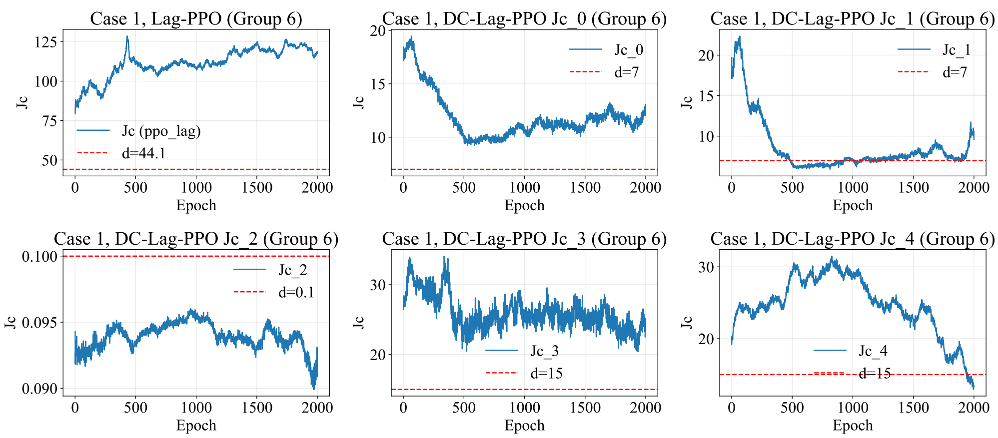


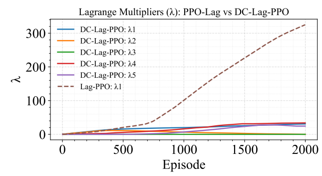
Figure 18: Training curves of J_c on Lagrangian cost threshold set: [7,7,0.1,15,15].

J.6 Lagrangian Multiplier Learning Curves

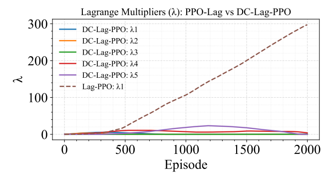
As shown in Fig. 19, the trajectories of the Lagrange multipliers provide insight into how Lag-PPO and DC-Lag-PPO enforce constraints during training. In the baseline Lag-PPO, the single multiplier tends to grow rapidly and exhibit instability, reflecting difficulty in balancing multiple heterogeneous constraints with a single aggregated signal. In contrast, DC-Lag-PPO assigns a dedicated multiplier to each constraint, and the resulting curves show more moderate growth and better separation among the multipliers. This indicates that the algorithm is able to distinguish between constraints of varying tightness and adjust enforcement accordingly.

Although some multipliers in DC-Lag-PPO still reach relatively high values, the spread across constraints suggests that the framework avoids over-penalizing all dimensions uniformly. Instead, it allocates stricter penalties only where violations are more prevalent. This aligns with the earlier observation that DC-Lag-PPO substantially reduces voltage violations, line overloads, and battery degradation, even though economic rewards are diminished.

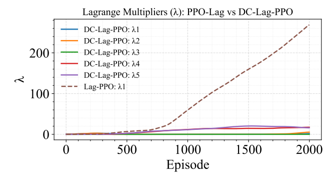
Overall, the Lagrange multiplier dynamics confirm that DC-Lag-PPO enforces constraints in a more structured and interpretable way than Lag-PPO. By disentangling constraint signals, it achieves stronger and more balanced compliance with operational limits, providing a safer and more reliable control policy for power system management.



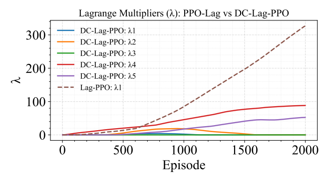
(a) Lagrangian cost threshold set [9,9,0.1,30,30]



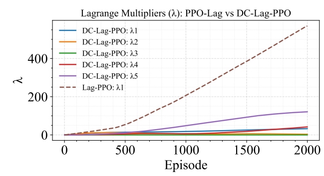
(b) Lagrangian cost threshold set [12,12,0.1,30,30].



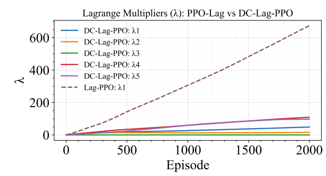
(c) Lagrangian cost threshold set [15,15,0.1,30,30]



(d) Lagrangian cost threshold set [18,18,0.1,20,30]



(e) Lagrangian cost threshold set [9,9,0.1,30,20]



(f) Lagrangian cost threshold set [7,7,0.1,15,15]

Figure 19: Lagrangian multiplier λ learning curves.

J.7 Pareto Fronts

Across all Pareto fronts shown in test cases in Fig. 20-27, which are obtained from different training runs, the DC-Lag-PPO fronts are typically shifted toward lower constraint values for the same (or nearby) reward levels, especially on the voltage metrics (violation ratio c1 and degree c2), indicating stronger constraint satisfaction without requiring large additional sacrifices in reward at the efficient frontier. This shift is visible in the reward-c1/c2 plots and also in the c1-c2, c1-c3, and c1–c4 pairings, where the dedicated-critic front envelopes or nearly envelopes the single-critic front.

A consistent pattern also emerges when examining the $J_c - d$ Pareto views across all parameter settings, where the points are the J_c values of each constraint cost, and d, the black dotted line, is the target threshold of each constraint (see Figs. 28-35). For every constraint dimension, the DC-Lag-PPO solutions cluster tightly around, or slightly below, the threshold d, forming a compact Pareto front near the lower-left region. In contrast, Lag-PPO's J_c values frequently lie well above the thresholds, and its Pareto front stretches diagonally upward, revealing strong trade-off tensions that arise from using a single shared multiplier. These $J_c - d$ relations provide direct evidence that DC-Lag-PPO not only finds better reward-constraint trade-offs but also fundamentally attains closer adherence to the prescribed limits on every constraint dimension.

Knee regions and policy selection Several plots exhibit knee points on the DC-Lag-PPO front (most clearly on voltage and line-loading axes), where a small relaxation in reward yields a disproportionate drop in violations. These knees are natural operating points for deployment, offering strong safety gains at modest economic cost.

In the $J_c - d$ space, knee behavior manifests as sharp transitions where J_c collapses rapidly once the policy enters the feasible region. These knees appear consistently in DC-Lag-PPO but rarely in Lag-PPO, further indicating that decomposed critics create a more controllable and interpretable constraint landscape.

Voltage safety trade-offs (C1, C2) For voltage ratio and degree, DC-Lag-PPO consistently attains lower violations at comparable reward, producing a "left/downward" movement of the frontier relative to Lag-PPO. The paired-constraint views (C1 vs. C2) show a visibly tighter cloud and a frontier closer to the origin, suggesting better joint compliance.

The corresponding $J_c - d$ results reinforce this pattern: across all runs, DC-Lag-PPO keeps J_c (C1) and J_c (C2) very near the voltage thresholds, whereas Lag-PPO exhibits persistent overshoot. This aligns with the training curves and confirms that decomposed voltage critics effectively isolate and regulate the two voltage-related risks during testing.

Line loading and degradation (C3, C4) On line loading (C3) and battery degradation (C4), DC-Lag-PPO fronts again tend to sit below the Lag-PPO fronts for similar reward ranges, implying reduced thermal stress and milder throughput for batteries. The cross-constraint plots (e.g., C2-C3, C3-C4) also show that dedicated-critic solutions better balance these two operational risks simultaneously.

The $J_c - d$ plots show the same effect: DC-Lag-PPO pushes J_c(C3) and J_c(C4) tightly toward their respective thresholds, often forming extremely compact clusters around d, while Lag-PPO's distributions remain dispersed and systematically above the limits. This confirms that multi-critic Lagrangian updates mitigate cross-constraint interference that otherwise destabilizes single-multiplier methods.

PV curtailment unfairness (C5) In the reward C5 panels and the mixed-constraint views involving C5, the dedicated-critic frontier usually dominates or matches the single-critic frontier for a broad range, indicating more equitable PV curtailment at similar reward. That said, dispersion varies across runs, hinting that fairness may remain sensitive to training seed or tariff profiles.

The $J_c - d$ comparisons show that DC-Lag-PPO frequently holds J_c(C5) near the fairness threshold, whereas Lag-PPO often overshoots or displays large variance. This confirms that separating the fairness critic prevents it from being overshadowed by voltage/thermal constraints during optimization.

Discussion DC-Lag-PPO delivers stronger and more balanced constraint satisfaction than Lag-PPO, most prominently on voltage safety and with consistent advantages on line loading, degradation, and fairness. The additional $J_c - d$ evidence strengthens this conclusion: the dedicated-critic design yields systematically lower J_c values tightly aligned with target thresholds, while the single-critic baseline exhibits structural difficulty simultaneously controlling heterogeneous constraints. The Pareto frontier shifts indicate that many safe operating points do not require drastic reward compromises once the policy is tuned to the knee region. Fairness (C5) gains are evident, though variability suggests room for additional stabilization (e.g., densifying episodic fairness or smoothing dual updates) in future runs.

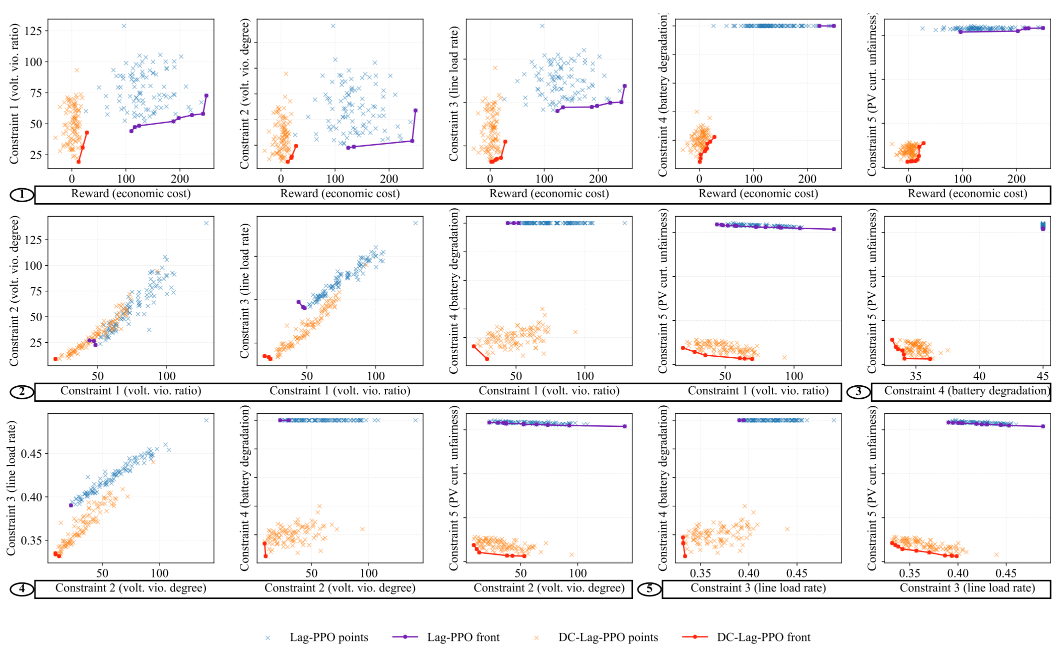


Figure 20: Pareto fronts from the test results on Lagrangian cost threshold set [9,9,0.1,30,30].

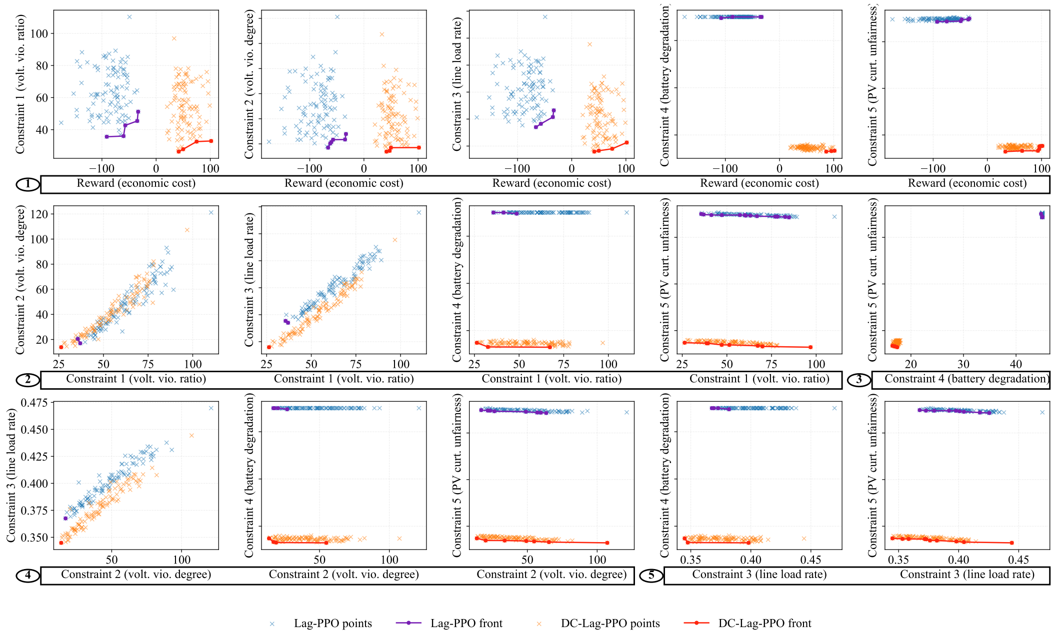


Figure 21: Pareto fronts from the test results on Lagrangian cost threshold set [12,12,0.1,30,30].

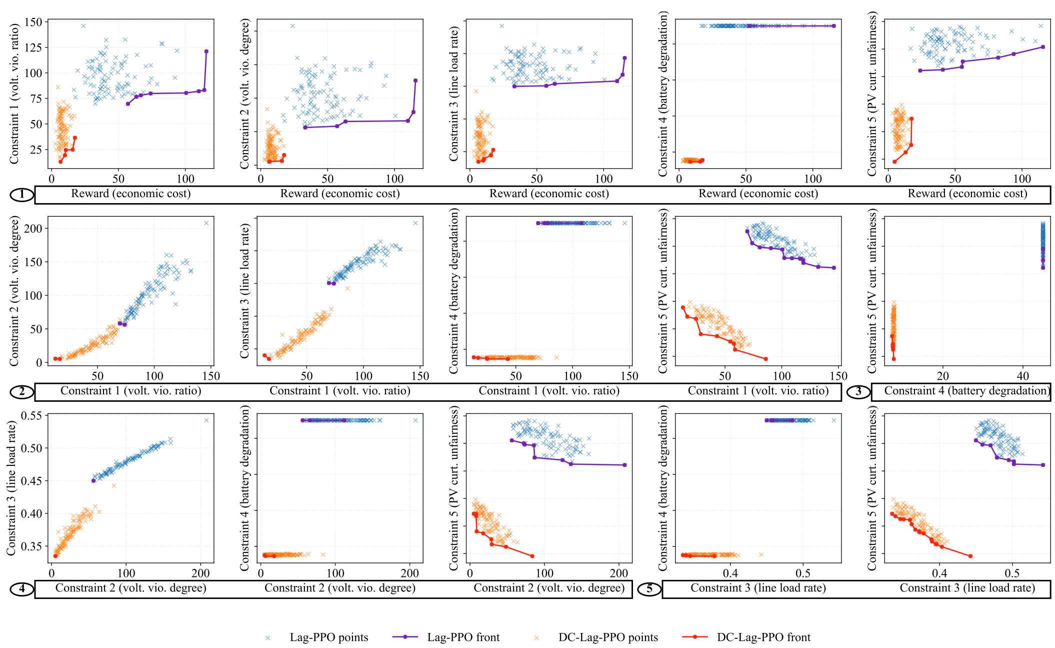


Figure 22: Pareto fronts from the test results on Lagrangian cost threshold set [15,15,0.1,20,30].

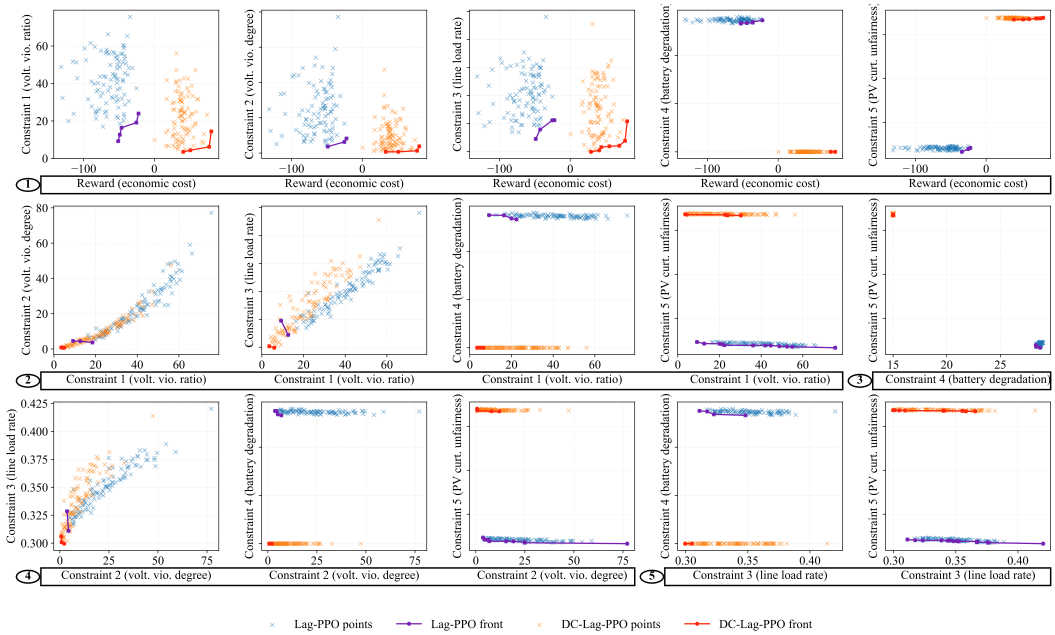


Figure 23: Pareto fronts from the test results on Lagrangian cost threshold set [18,18,0.1,20,30].

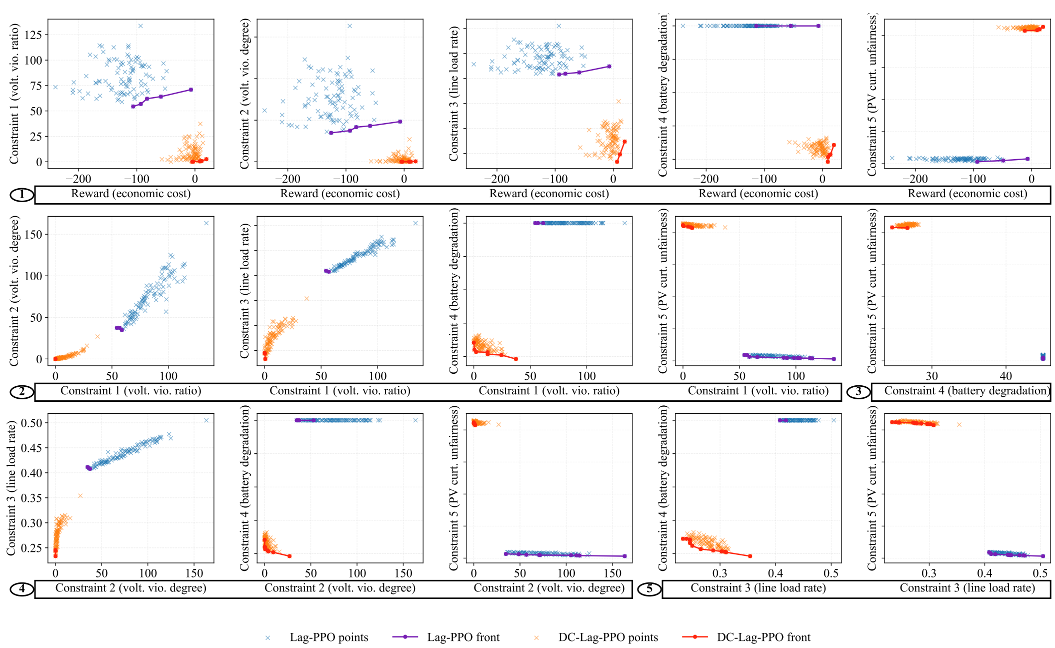


Figure 24: Pareto fronts from the test results on Lagrangian cost threshold set [9,9,0.1,30,20].

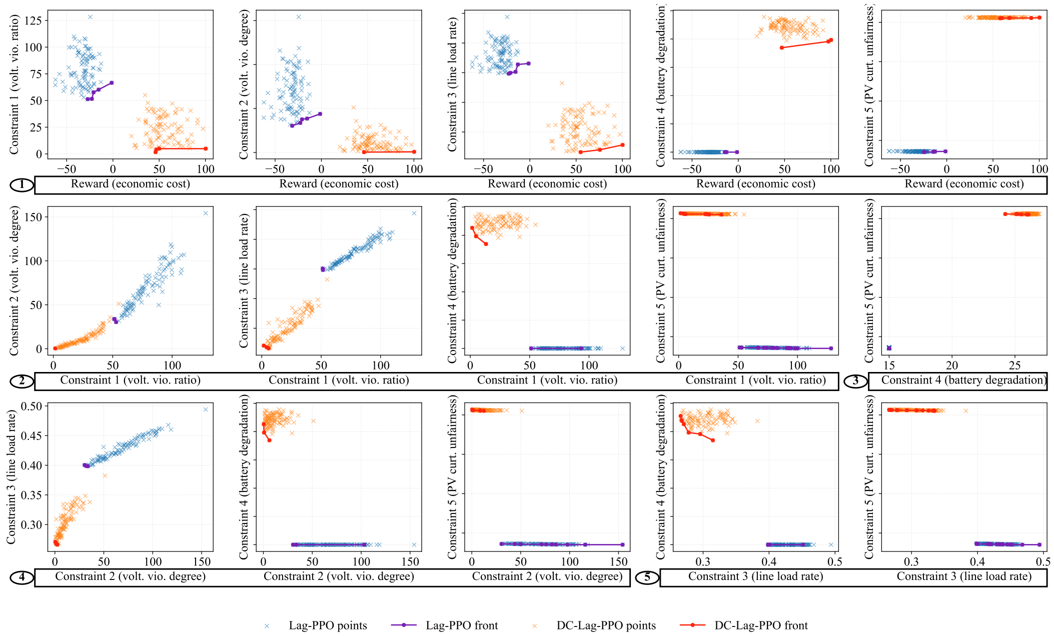


Figure 25: Pareto fronts from the test results on Lagrangian cost threshold set [12,12,0.1,30,20].

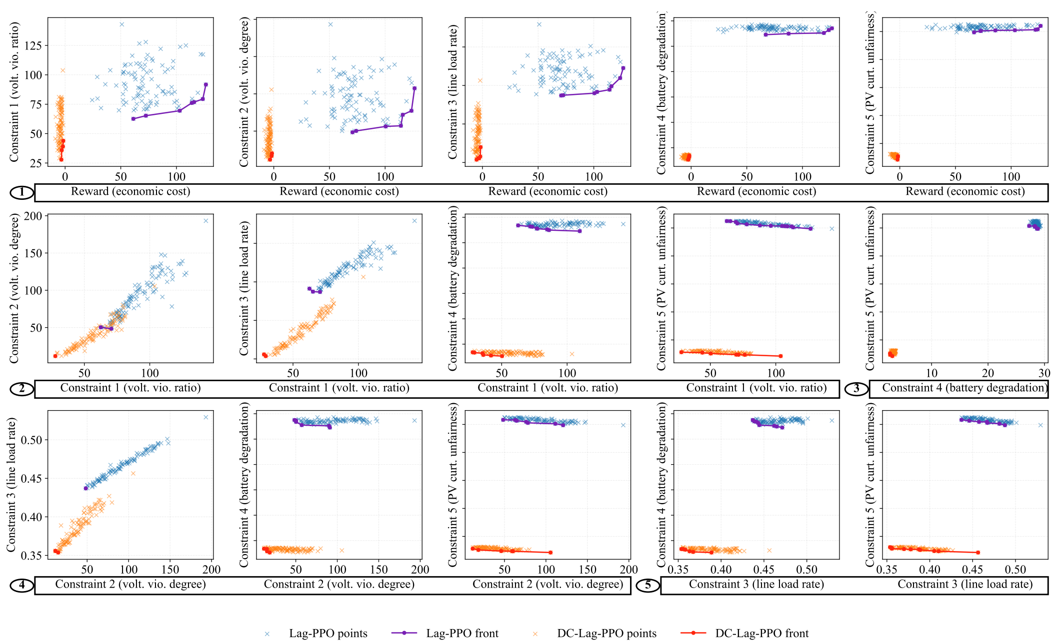


Figure 26: Pareto fronts from the test results on Lagrangian cost threshold set [9,9,0.1,20,20].

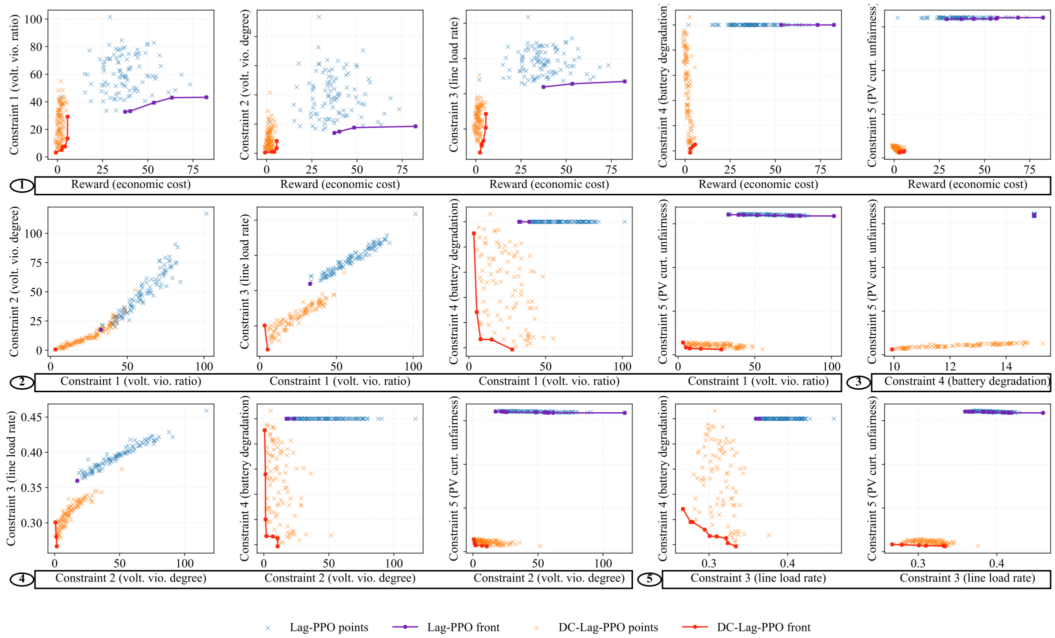


Figure 27: Pareto fronts from the test results on Lagrangian cost threshold set [7,7,0.1,15,15].

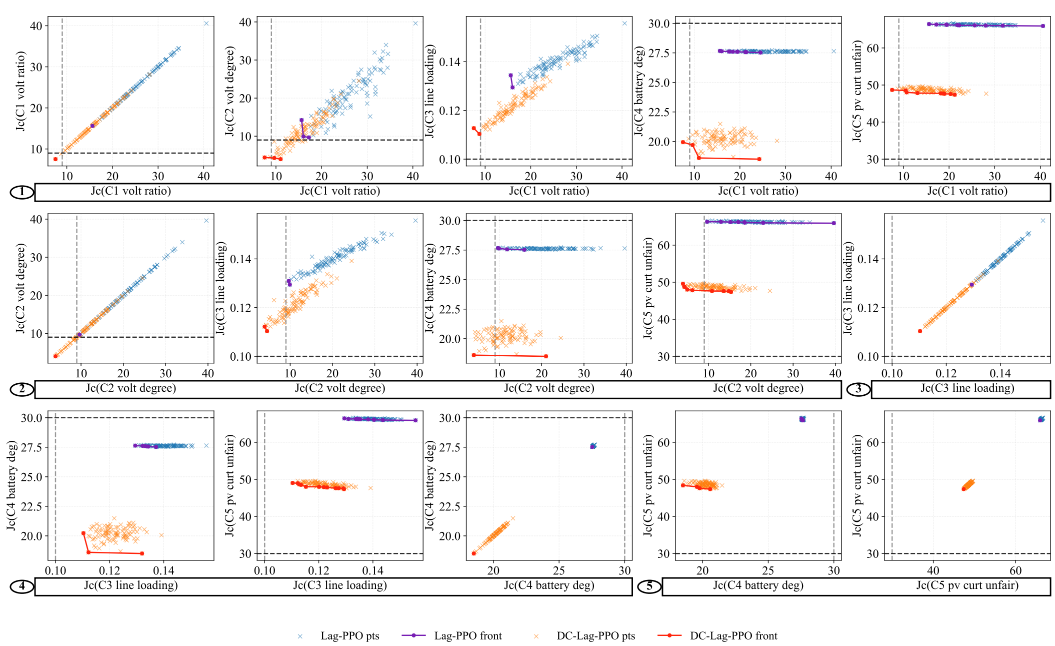


Figure 28: Pareto fronts from the test results of J_c on Lagrangian cost threshold set [9,9,0.1,30,30], where the black dotted lines are the thresholds d.

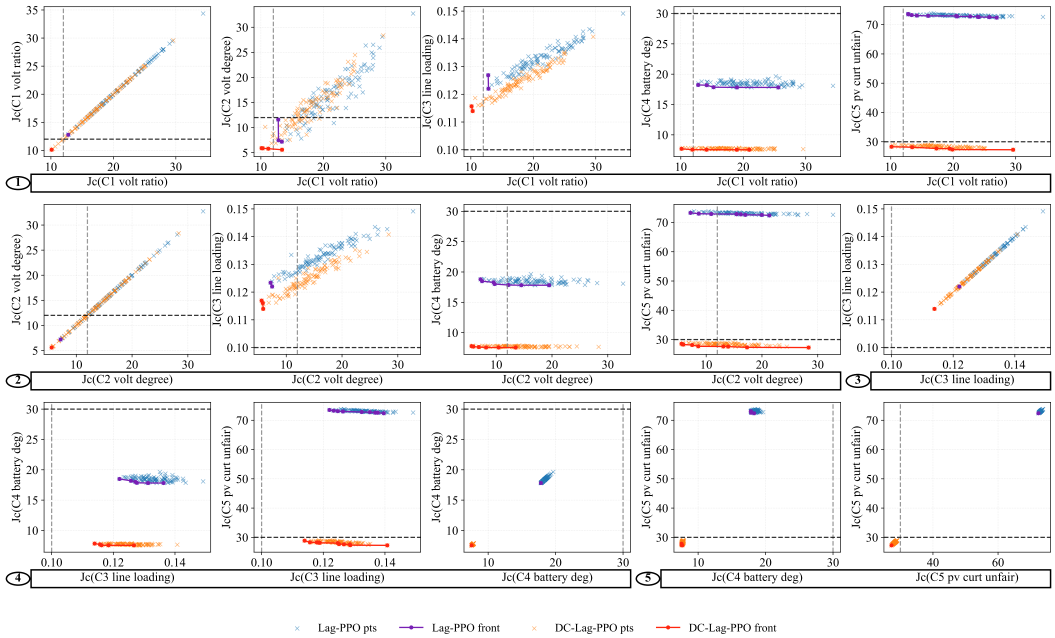


Figure 29: Pareto fronts from the test results of J_c on Lagrangian cost threshold set [12,12,0.1,30,30], where the black dotted lines are the thresholds d.

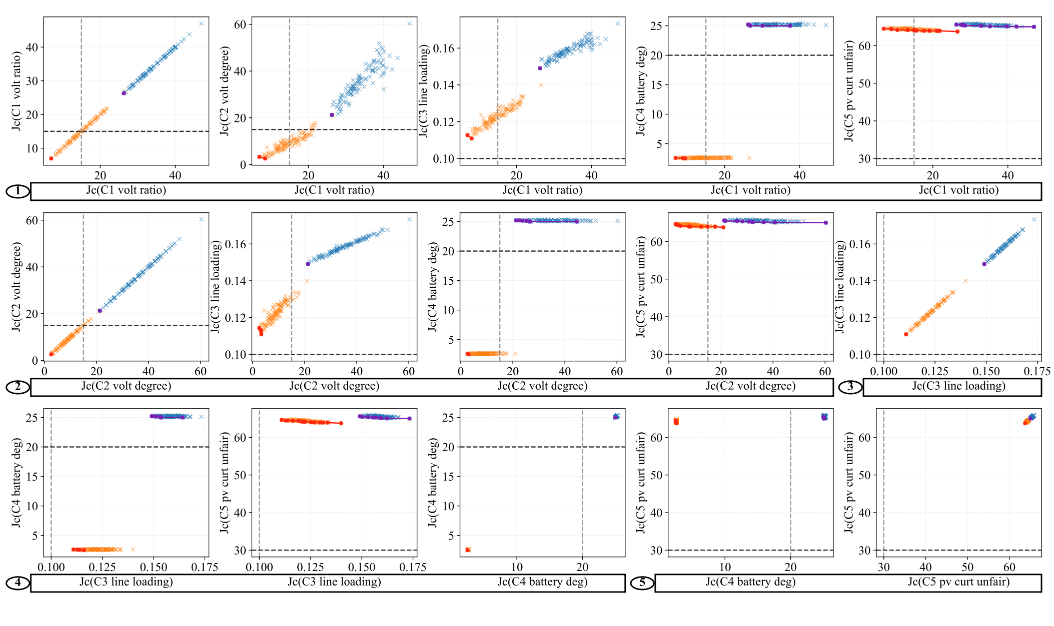


Figure 30: Pareto fronts from the test results of J_c on Lagrangian cost threshold set [15,15,0.1,20,30], where the black dotted lines are the thresholds d.

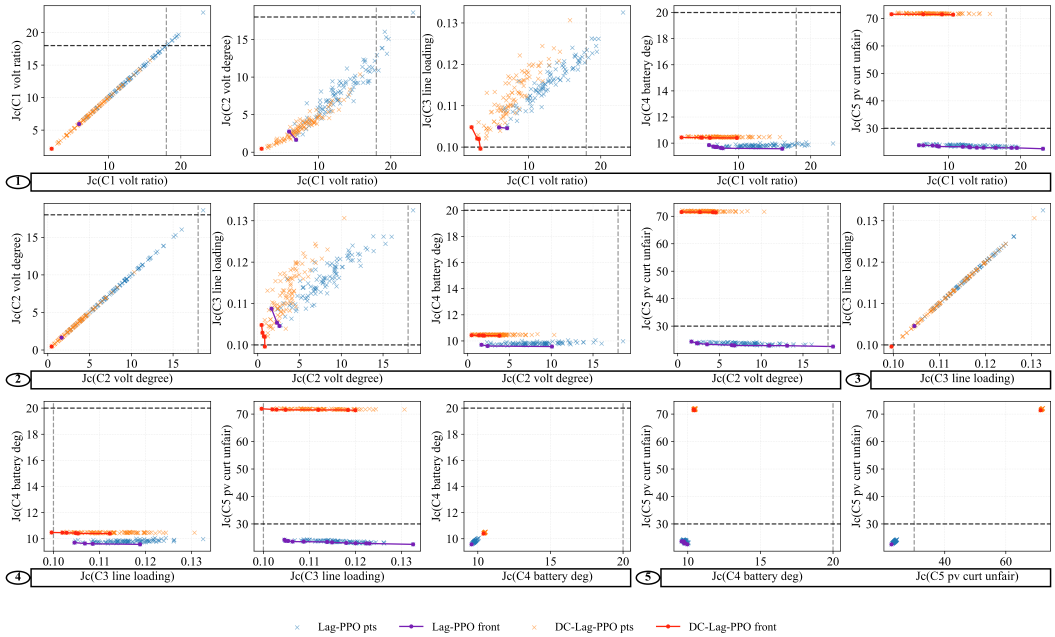


Figure 31: Pareto fronts from the test results of J_c on Lagrangian cost threshold set [18,18,0.1,20,30], where the black dotted lines are the thresholds d.

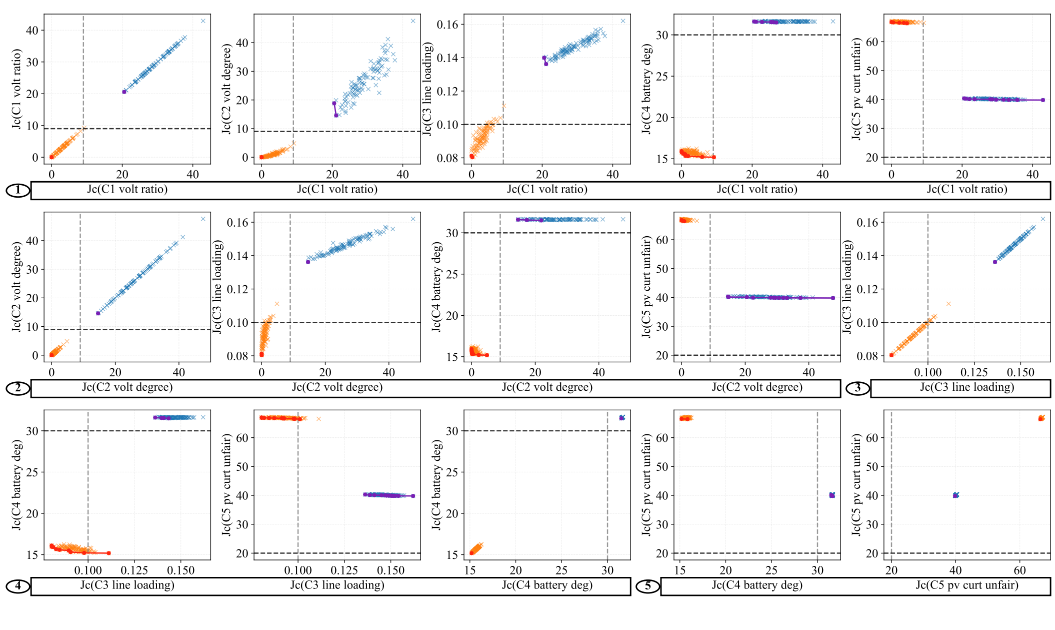


Figure 32: Pareto fronts from the test results of J_c on Lagrangian cost threshold set [9,9,0.1,30,20], where the black dotted lines are the thresholds d.

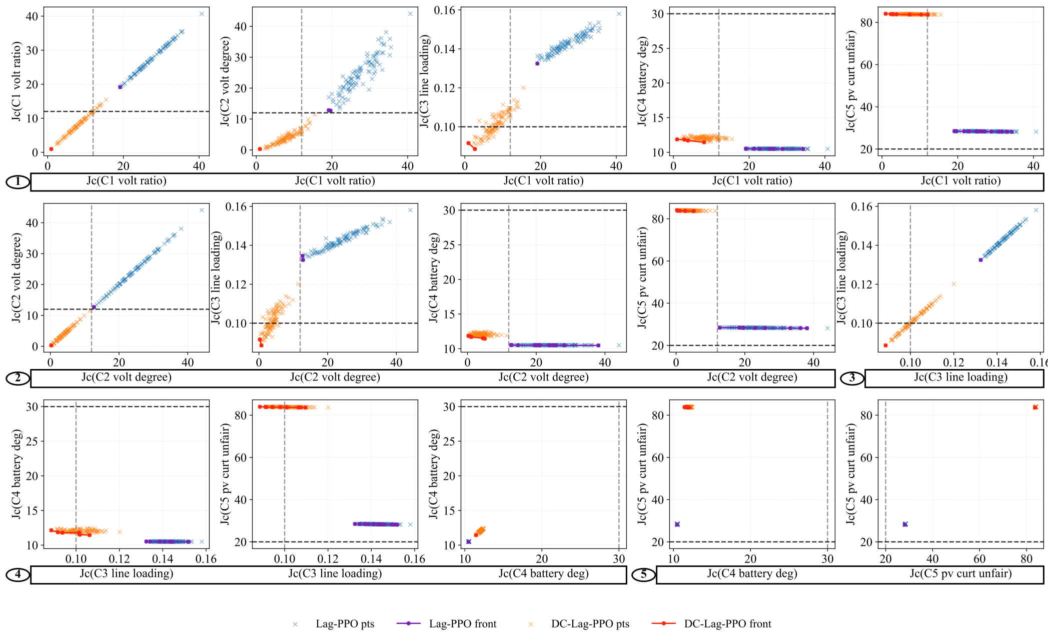


Figure 33: Pareto fronts from the test results of J_c on Lagrangian cost threshold set [12,12,0.1,30,20], where the black dotted lines are the thresholds d.

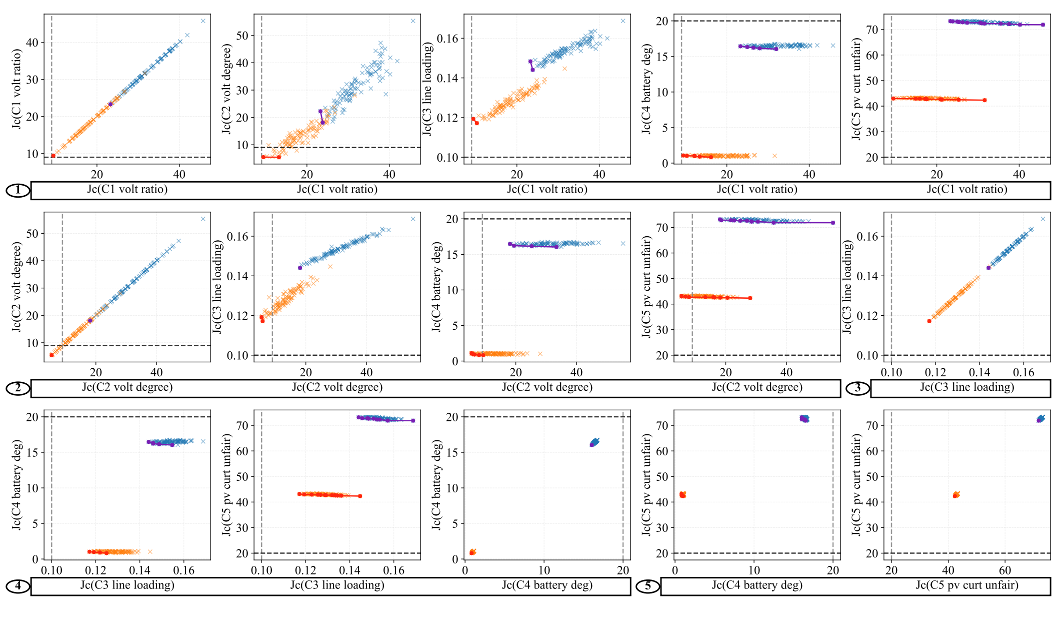


Figure 34: Pareto fronts from the test results of J_c on Lagrangian cost threshold set [9,9,0.1,20,20], where the black dotted lines are the thresholds d.

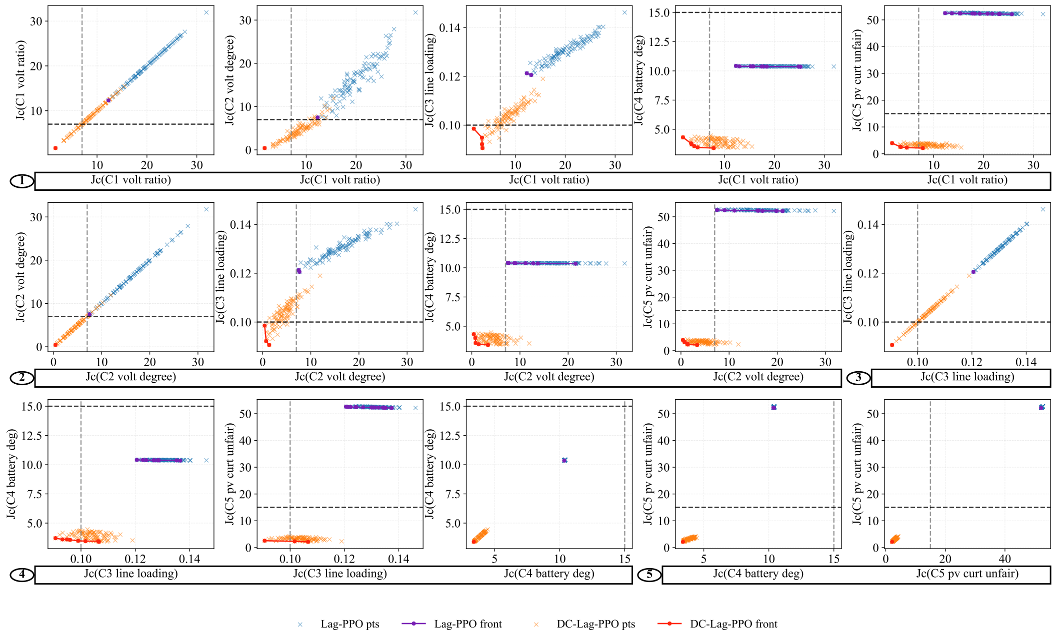


Figure 35: Pareto fronts from the test results of J_c on Lagrangian cost threshold set [7,7,0.1,15,15], where the black dotted lines are the thresholds d.

K Experiment Details - Case 2

We extend our experiments to a more complex electric vehicle charging station (EVCS) problem to further evaluate how well the dedicated-critic approach scales and whether its benefits persist in realistic multi-constraint settings (Power system environment $\rightarrow$ observation dimension: 105, action dimension: 24; Electric Vehicle Charging $\rightarrow$ Observation dimension: 219; action dimension: 40). This environment models coordination problems, where the goal is to minimize charging costs while enforcing multiple charging related constraints, including voltage limits, EV battery degradation, and charging demand satisfaction. In contrast to the power system problem, this setting involves coordinating multiple EVCSs, each operating dozens of chargers and responding to highly stochastic and heterogeneous EV behaviours (arrival and departure times, charging demands, battery capacities, etc.). As a result, both the state and action spaces are substantially higher-dimensional, and the additional uncertainty introduced by EV dynamics makes the EVCS coordination task considerably more challenging than community battery scheduling.

K.1 System Description

K.1.1 PDN

The power distribution network (PDN) consists of a set of buses $\mathcal{N} = \{1, \dots, N\}$ interconnected via distribution lines $\mathcal{L} \subseteq \mathcal{N} \times \mathcal{N}$. The system evolves over a discrete time horizon $\mathcal{T} = \{1, \dots, T\}$. At each bus $i \in \mathcal{N}$ and time $t \in \mathcal{T}$, let $p_{i,t}$ and $v_{i,t}$ denote the net active power injection and voltage magnitude, respectively.

All nodes must satisfy the standard voltage bounds:

$$V^{\min} \le v_{i,t} \le V^{\max}, \qquad \forall i \in \mathcal{N},\ t \in \mathcal{T}. \tag{68}$$

K.1.2 EVCS Deployment and Neighborhood Structure

The DSO manages a set of EV charging stations (EVCSs) $\mathcal{K} = \{1, \dots, K\}$, each equipped with rooftop PV generation and a set of chargers $\mathcal{C}_k = \{1, \dots, C_k\}$. EVCS k is placed at exactly one PDN bus, represented by the binary deployment matrix $\mathbf{K} \in \{0,1\}^{N \times K}$:

$$K_{ik} = 1 \text{ if EVCS } k \text{ is located at bus } i, \quad K_{ik} = 0 \text{ otherwise,}$$

with the physical constraint that no two EVCSs colocate:

$$\sum_{k \in \mathcal{K}} K_{ik} \le 1, \qquad \forall i \in \mathcal{N}.$$

For each bus i, its one-hop neighborhood is defined as

$$\mathcal{N}_i^{(1)} = \{j \in \mathcal{N} \setminus \{i\} \mid (i,j) \in \mathcal{L}\}.$$

If EVCS k is located at bus i ($K_{ik} = 1$), its accessible neighborhood is

$$\mathcal{N}_k^{(1)} = \mathcal{N}_i^{(1)},$$

representing all physically adjacent buses whose aggregate voltage and load information is available.

K.1.3 EV Charging Model

Each charger serves EVs that arrive, park for a duration, and leave with a required energy level. Let $T_{c_k}^{\text{arr}}$ and $T_{c_k}^{\text{dep}}$ denote the arrival and departure times of EV c_k, and let $\text{SoC}_{c_k}^{\text{arr}}$ and $\text{SoC}_{c_k}^{\text{dep}}$ be the corresponding SoC levels. Their target SoC trajectory is modeled via linear interpolation:

$$\text{SoC}_{c_k}^{\text{target}}(t) = \text{SoC}_{c_k}^{\text{arr}} + \frac{t - T_{c_k}^{\text{arr}}}{T_{c_k}^{\text{dep}} - T_{c_k}^{\text{arr}}} \left(\text{SoC}_{c_k}^{\text{dep}} - \text{SoC}_{c_k}^{\text{arr}} \right).$$

Each charger must satisfy:

$$0 \le p^{\mathrm{ch}}_{c_k}(t) \le P^{\mathrm{ch,max}}_{c_k}, \tag{69a}$$
$$0 \le p^{\mathrm{dis}}_{c_k}(t) \le P^{\mathrm{dis,max}}_{c_k}, \tag{69b}$$
$$p^{\mathrm{ch}}_{c_k}(t)\, p^{\mathrm{dis}}_{c_k}(t) = 0, \tag{69c}$$
$$\mathrm{SoC}^{\min}_{c_k} \le \mathrm{SoC}_{c_k}(t) \le \mathrm{SoC}^{\max}_{c_k}, \tag{69d}$$
$$\Delta\mathrm{SoC}_{c_k}(t) = \eta^{\mathrm{ch}} p^{\mathrm{ch}}_{c_k}(t) - \frac{1}{\eta^{\mathrm{dis}}} p^{\mathrm{dis}}_{c_k}(t). \tag{69e}$$

K.1.4 Voltage Violation Metrics

In addition to total voltage violation, we also consider the number of voltage-violating buses:

$$f^{\mathrm{NV}}(t) = \sum_{i\in\mathcal{N}} \mathbf{1}\{v_{i,t} \notin [V^{\min}, V^{\max}]\}.$$

This discrete stability metric counts the extent of widespread voltage deviations across the PDN.

K.1.5 Demand Satisfaction Violation Rate

Let N^{EV} denote the total number of EVs served during the horizon. Define a violation indicator for each EV:

$$\delta^{\mathrm{DS}}_c = \mathbf{1}\Big\{\mathrm{SoC}_c(T^{\mathrm{dep}}_c) < 0.95\,\mathrm{SoC}^{\mathrm{dep}}_c\Big\},$$

i.e., the EV fails to achieve at least 95% of its desired departure SoC. The demand satisfaction violation rate is then

$$f^{\mathrm{VR}}_{\mathrm{DS}} = \frac{1}{N^{\mathrm{EV}}} \sum_{c=1}^{N^{\mathrm{EV}}} \delta^{\mathrm{DS}}_c.$$

K.1.6 Operational Cost Functions

At each EVCS k, the DSO controls charging and discharging powers $\{p^{\mathrm{ch}}_{c_k}(t),\, p^{\mathrm{dis}}_{c_k}(t)\}_{c_k\in\mathcal{C}_k}$. The cost components are:

$$f^{\mathrm{TD}}_k(t) = \begin{cases} \lambda^{\mathrm{buy}}_t\, p^{\mathrm{TD}}_k(t), & p^{\mathrm{TD}}_k(t) > 0, \\ \lambda^{\mathrm{sell}}_t\, p^{\mathrm{TD}}_k(t), & \text{otherwise}, \end{cases}$$

$$f^{\mathrm{DG}}_k(t) = \alpha_e \sum_{c_k\in\mathcal{C}_k} \left([p^{\mathrm{ch}}_{c_k}(t)]^2 + [p^{\mathrm{dis}}_{c_k}(t)]^2\right),$$

$$f^{\mathrm{VT}}(t) = \sum_{i\in\mathcal{N}} \Big([v_{i,t} - V^{\max}]^+ + [V^{\min} - v_{i,t}]^+\Big),$$

$$f^{\mathrm{DS}}_k(t) = \sum_{c_k\in\mathcal{C}_k} \left[\mathrm{SoC}^{\mathrm{target}}_{c_k}(t) - \mathrm{SoC}_{c_k}(t)\right]^+,$$

with traded power

$$p^{\mathrm{TD}}_k(t) = \sum_{c_k\in\mathcal{C}_k} \left(p^{\mathrm{ch}}_{c_k}(t) - p^{\mathrm{dis}}_{c_k}(t)\right) - p^{\mathrm{PV}}_{k,t}.$$

K.1.7 Objective

The DSO seeks to minimize aggregated operational costs and violation penalties:

$$\min_{\mathbf{p}^{\text{ch}},\,\mathbf{p}^{\text{dis}}} \sum_{t\in\mathcal{T}} \left[\sum_{k\in\mathcal{K}} \left(\beta_1 f_k^{\text{TD}}(t) + \beta_2 f_k^{\text{DG}}(t) + \beta_3 f_k^{\text{DS}}(t)\right) + \beta_4 f^{\text{VT}}(t) + \beta_5 f^{\text{NV}}(t) \right] + \beta_6 f_{\text{DS}}^{\text{VR}}. \quad (70)$$

This objective captures energy costs, degradation, voltage safety, spatial extent of voltage violations, and global demand satisfaction reliability.

K.2 CMDP Formulation with Dedicated-Critic Lagrangian RL

The EVCS coordination problem is modeled as a constrained Markov decision process (CMDP)

$$\left(\mathcal{S}, \mathcal{A}, P, r, \{c_i\}_{i=1}^{m}, \gamma, \{d_i\}_{i=1}^{m}\right),$$

where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, $P(\cdot|s,a)$ the transition kernel, $r(s,a)$ the reward signal, $c_i(s,a)$ the cost signal for constraint i with threshold d_i, and $\gamma \in (0,1)$ the discount factor.

For a policy $\pi_\theta(a|s)$, define the discounted returns:

$$J_r(\pi_\theta) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t\, r(s_t, a_t) \right], \quad (71)$$

$$J_{c_i}(\pi_\theta) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t\, c_i(s_t, a_t) \right], \quad i = 1, \ldots, m. \quad (72)$$

The CMDP objective is

$$\max_{\theta}\; J_r(\pi_\theta) \qquad \text{s.t.} \qquad J_{c_i}(\pi_\theta) \le d_i, \;\; i = 1, \ldots, m. \quad (73)$$

Reward and cost signals derived from the system model. Let the instantaneous cost components from the system description be:

- f_t^{VT}: total voltage violation magnitude,
- f_t^{NV}: number of voltage-violating buses,
- f_t^{LL}: line-loading stress,
- f_t^{DG}: battery degradation,
- f_t^{TD}: energy trading cost,
- f_t^{DS}: per-step EV dissatisfaction,
- $f_{\text{DS}}^{\text{VR}}$: demand satisfaction violation rate (episodic).

A practical reward–cost decomposition aligning with operational goals is:

$$r(s_t, a_t) = -\left(\alpha_{\text{TD}}\, f_t^{\text{TD}} + \alpha_{\text{DG}}\, f_t^{\text{DG}} + \alpha_{\text{DS}}\, f_t^{\text{DS}}\right), \quad (74)$$

$$c_1(s_t, a_t) = f_t^{\text{VT}}, \qquad \text{(voltage violation magnitude)} \quad (75)$$

$$c_2(s_t, a_t) = f_t^{\text{NV}}, \qquad \text{(number of violating buses)} \quad (76)$$

$$c_3(s_t, a_t) = f_t^{\text{LL}}, \qquad \text{(line loading)} \quad (77)$$

$$c_4(s_t, a_t) = f_t^{\text{DG}}, \qquad \text{(battery degradation)} \quad (78)$$

Additionally, the demand-satisfaction violation rate $f_{\text{DS}}^{\text{VR}}$ is an episodic cost:

$$C_5(\tau) \triangleq f_{\text{DS}}^{\text{VR}}, \qquad \mathbb{E}_\pi[C_5(\tau)] \le d_5, \quad (79)$$

where τ denotes a full episode. If preferred, $f_{\text{DS}}^{\text{VR}}$ can be distributed as a per-step cost $c_5(s_t, a_t)$ such that its discounted sum recovers the same episodic value.

Lagrangian relaxation with per-constraint critics. Introduce dual multipliers $\lambda = (\lambda_1, \dots, \lambda_m) \succeq 0$ and form the Lagrangian:

$$\mathcal{L}(\theta, \lambda) = J_r(\pi_\theta) - \sum_{i=1}^{m} \lambda_i \left(J_{c_i}(\pi_\theta) - d_i\right). \tag{80}$$

Primal–dual updates follow:

$$\nabla_\theta \mathcal{L}(\theta, \lambda) = \nabla_\theta J_r(\pi_\theta) - \sum_{i=1}^{m} \lambda_i \nabla_\theta J_{c_i}(\pi_\theta), \tag{81}$$

$$\lambda_i \leftarrow \Pi_{[0,\lambda_{\max}]}\Big(\lambda_i + \beta\,(\widehat{J}_{c_i} - d_i)\Big), \tag{82}$$

with Π denoting projection for stability.

Value functions and signal-specific advantages. For each signal $x \in \{r, c_1, \dots, c_m\}$, define:

$$Q^x_\pi(s, a) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t x(s_t, a_t) \mid s_0 = s, a_0 = a\Big], \tag{83}$$

$$V^x_\pi(s) = \mathbb{E}_{a\sim\pi}[Q^x_\pi(s, a)], \tag{84}$$

$$A^x_\pi(s, a) = Q^x_\pi(s, a) - V^x_\pi(s). \tag{85}$$

The actor gradient becomes:

$$\nabla_\theta \mathcal{L}(\theta, \lambda) = \mathbb{E}_\pi\big[\nabla_\theta \log \pi_\theta(a|s) \left(A^r_\pi(s, a) - \sum_{i=1}^{m} \lambda_i A^{c_i}_\pi(s, a)\right)\big]. \tag{86}$$

Dedicated critics for each signal. Each signal $x \in \{r, c_1, \dots, c_m\}$ is assigned a separate critic Q_{ω_x}:

$$\delta^x_t = x_t + \gamma Q_{\omega_x}(s_{t+1}, a_{t+1}) - Q_{\omega_x}(s_t, a_t), \tag{87}$$

and the critic minimizes $\mathbb{E}[(\delta^x_t)^2]$. Advantage estimates (e.g., GAE) are computed per signal and combined through the Lagrangian structure.

PPO-style actor update. Let $r_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\theta_{\text{old}}}(a_t|s_t)$ and

$$\widetilde{A}_t = A^r_t - \sum_{i=1}^{m} \lambda_i A^{c_i}_t.$$

The clipped surrogate is

$$\mathcal{J}_{\text{PPO}}(\theta) = \mathbb{E}\Big[\min\big(r_t(\theta)\,\widetilde{A}_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\widetilde{A}_t\big)\Big] + \eta\,\mathbb{E}[\mathcal{H}(\pi_\theta(\cdot|s_t))], \tag{88}$$

where $\mathcal{H}$ denotes policy entropy.

Instantiated constraints for this problem. With the reward and cost mapping above, we typically have

$$m = 5,$$

corresponding to:

- c_1: voltage violation magnitude f^{VT}_t,
- c_2: violating node count f^{NV}_t,
- c_3: line loading f^{LL}_t,
- c_4: battery degradation f^{DG}_t,
- c_5: demand-satisfaction violation rate $f^{\text{VR}}_{\text{DS}}$ (episodic or densified).

Table 3: Key hyperparameters, reward structure, and CMDP constraints for the EVCS coordination case study.

Category	Term / Parameter	Value	Definition / Description
Reward & CMDP constraints			
Reward r_t	$-(\alpha_{\text{TD}} f_t^{\text{TD}} + \alpha_{\text{DG}} f_t^{\text{DG}} + \alpha_{\text{DS}} f_t^{\text{DS}})$	–	Trading, degradation, and dissatisfaction penalties
Constraint c_1 (Voltage magnitude violation)	$f_t^{\text{VT}} = \sum_i ([v_{i,t} - V^{\max}]^+ + [V^{\min} - v_{i,t}]^+)$	–	Voltage violation magnitude across buses
Constraint c_2 (Count of violating buses)	$f_t^{\text{NV}} = \sum_i \mathbf{1}\{v_{i,t} \notin [V^{\min}, V^{\max}]\}$	$[0, N]$	Number of buses violating voltage limits
Constraint c_3 (Line loading)	$f_t^{\text{LL}} = \sum_{(i,j)} [\ell_{ij,t} - \tau^{\text{line}}]^+$	–	Thermal overload above permissible threshold
Constraint c_4 (Battery degradation)	$f_t^{\text{DG}} = \alpha_e \sum_{c_k} ([p_{c_k,t}^{\text{ch}}]^2 + [p_{c_k,t}^{\text{dis}}]^2)$	–	Throughput-based quadratic degradation
Constraint c_5 (Demand satisfaction violation rate)	$f_{\text{DS}}^{\text{VR}}$	$[0, 1]$	Fraction of EVs leaving with SoC $< 0.95\,\text{SoC}^{\text{dep}}$
Lag-PPO baseline constraint	$\sum_{i=1}^{5} c_i$	–	Single aggregated constraint in standard Lag-PPO
General training hyperparameters			
Learning rate	–	3×10^{-4}	For both actor and critics
PPO clip ϵ	–	0.2	Ratio clipping: $[1-\epsilon, 1+\epsilon]$
Target KL	–	0.015	Early stopping threshold
Value loss coefficient	–	0.5	Weight for critic loss
Entropy coefficient	–	0.0	Entropy regularization
Gradient norm clip	–	0.5	Global clipping limit
Hidden sizes	–	(256, 256)	MLP layers for all networks
Init log-std	–	-0.5	Gaussian policy initialization
Discount γ / GAE λ	–	0.99 / 0.95	Returns and advantage estimation
Dual learning rate	–	5×10^{-3}	Step size for multiplier update
λ init / max	–	0.0 / 10^4	Multiplier projection range
Training schedule & environment			
Episodes	–	2000	Total PPO training episodes
Steps per episode	–	288	One full day (5 min resolution)
Environment step	–	5 min	Sampling interval

The corresponding critics are:

$$Q_{\omega_r},\ Q_{\omega_{c_1}}, Q_{\omega_{c_2}}, Q_{\omega_{c_3}}, Q_{\omega_{c_4}}, Q_{\omega_{c_5}}.$$

The combined advantage is:

$$\widetilde{A}_t = A_t^r - \lambda_1 A_t^{c_1} - \lambda_2 A_t^{c_2} - \lambda_3 A_t^{c_3} - \lambda_4 A_t^{c_4} - \lambda_5 A_t^{c_5}.$$

Dual multipliers update via equation 82.

Practical considerations. To stabilize learning: (i) use target networks or Polyak averaging for each critic; (ii) normalize each advantage $A_t^{c_i}$ before aggregation; (iii) constrain multipliers via projection or softplus parameterization; (iv) for episodic costs, update multipliers once per episode; stepwise costs update per batch.

K.3 Experimental Parameters

Symbols. $\phi_t^{\text{buy}}, \phi_t^{\text{sell}}$: buy/sell electricity prices; $[x]^+ = \max(x, 0)$; $\mathbf{1}\{\cdot\}$: indicator; $v_{i,t}$: voltage at bus i; $V^{\min}, V^{\max}$: voltage bounds; $\ell_{ij,t}$: loading of line (i,j); τ^{line}: overload threshold; $p_{c_k,t}^{\text{ch}}, p_{c_k,t}^{\text{dis}}$: charging/discharging powers; $\gamma_{i,t}$: PV curtailment ratio; Δt: step duration (5 min); N: number of buses; $|\mathcal{C}_k|$: chargers at EVCS k.

K.4 Two-Tiered Statistics

See Table 4 for details. This table summarizes the two-tiered evaluation statistics, reporting the mean and standard deviation across three independent training runs. Overall, DC-Lag-PPO consistently outperforms Lag-PPO across all constraint metrics while also achieving better economic performance, confirming that decomposing the critics alleviates the interference between heterogeneous constraints and stabilizes the dual updates.

DC-Lag-PPO reduces the economic cost substantially, outperforming Lag-PPO by 19.3 units on average, despite the inherently high variance of cost signals. The improvement of -109% (negative because higher reward is better) indicates that DC-Lag-PPO not only avoids the reward degradation often observed in constrained RL, but actually discovers more cost-efficient charging strategies while still satisfying the operational constraints.

For both voltage violation ratio (c1) and degree (c2), DC-Lag-PPO achieves consistent and significant reductions: -23.14 in violation ratio (+16.9% improvement), -44.67 in violation degree (+22.8% improvement). These gains validate the core motivation of the dedicated-critic design: each voltage-related critic captures its own risk landscape, preventing the single-critic baseline from being dominated by a few severe constraints. The larger improvement on violation degree (c2) suggests that DC-Lag-PPO not only reduces the frequency of violations but also suppresses their severity, producing safer voltage profiles across the entire PDN.

Battery throughput and degradation drop from 41.43 to 25.62, yielding the largest improvement among all instantaneous constraints (+38.2%). This indicates that DC-Lag-PPO is better at distributing the charging/discharging workload across EV chargers, avoiding the overuse of individual chargers or time windows. The result also aligns with DC-Lag-PPO's smoother dual updates, which prevent oscillatory behaviors commonly seen in single-multiplier methods.

The dissatisfaction volume is reduced from 41.07 to 36.72 (+12.2%), demonstrating that DC-Lag-PPO better supports EV users' charging requirements. The relatively low variance of DC-Lag-PPO also implies improved training stability and more consistent performance across runs.

Among all constraints, the dedicated-critic method yields one of the most significant improvements on dissatisfaction number from 53.13 to 35.44 (+33.9%). This confirms that DC-Lag-PPO not only reduces instantaneous dissatisfaction but also lowers the probability of EVs failing to meet their departure SoC requirement, complementing the improvement in dissatisfaction volume.

Table 4: Two-tiered test statistics, where across-run mean $\pm$ across-run std; The higher reward is better, while the lower constraints are better. Δ = (DC-Lag-PPO $-$ Lag-PPO). Positive improvement % is computed as (Lag-PPO $-$ DC-Lag-PPO)/Lag-PPO $\times$ 100%, except reward where the value is negated because higher is better.

Metric	Lag-PPO (n=3)	DC-Lag-PPO (n=3)	Δ	Improvement %
Economic cost (reward)	17.66 $\pm$ 40.88	-1.61 $\pm$ 10.02	**-19.27**	**-109.18%**
Volt. vio. ratio (c1)	136.67 $\pm$ 6.61	113.53 $\pm$ 10.72	**-23.14**	**+16.93%**
Volt. vio. degree (c2)	196.41 $\pm$ 14.06	151.74 $\pm$ 47.41	**-44.67**	**+22.75%**
Battery deg. (c3)	41.43 $\pm$ 4.86	25.62 $\pm$ 7.81	**-15.82**	**+38.18%**
Dissat. vol. (c3)	41.07 $\pm$ 9.01	36.72 $\pm$ 3.16	**-9.89**	**+12.17%**
Dissat. num. (c5)	53.13 $\pm$ 8.25	35.44 $\pm$ 5.22	**-11.68**	**+33.87%**

K.5 Training Curves

See Figure 36 and Figure 37 for detail. It can be seen that DC-Lag-PPO learns faster, stabilizes earlier, satisfies constraints better, and avoids the late-stage divergence exhibited by Lag-PPO.

First, DC-Lag-PPO achieves lower and more stable economic cost, whereas Lag-PPO exhibits large oscillations and late-stage degradation. Second, for all constraint metrics, including voltage ratio/degree, line loading, battery degradation, and demand dissatisfaction, DC-Lag-PPO maintains lower violation levels throughout training, with clearly reduced variance. In contrast, Lag-PPO's curves drift upward or fluctuate heavily, indicating unstable constraint handling under the aggregated-critic formulation.

The J_c curves further confirm this advantage: every DC-Lag-PPO constraint return steadily moves toward its target threshold d crosses it, and eventually stabilizes near or below the limit. Lag-PPO, however, shows a single aggregated J_c that quickly rises above the feasibility threshold and fails to recover, demonstrating an inability to control multiple constraints simultaneously.

In summary, DC-Lag-PPO learns faster, stabilizes earlier, satisfies constraints more reliably, and avoids the divergence observed in Lag-PPO, showing clear benefits of using dedicated critics for multi-constraint RL.

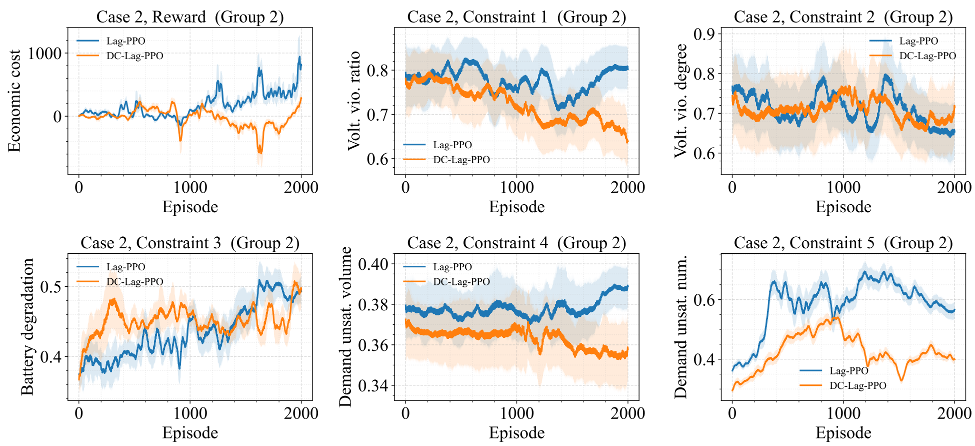


Figure 36: Training curves on Lagrangian cost threshold set: [30,39,20,20,30].

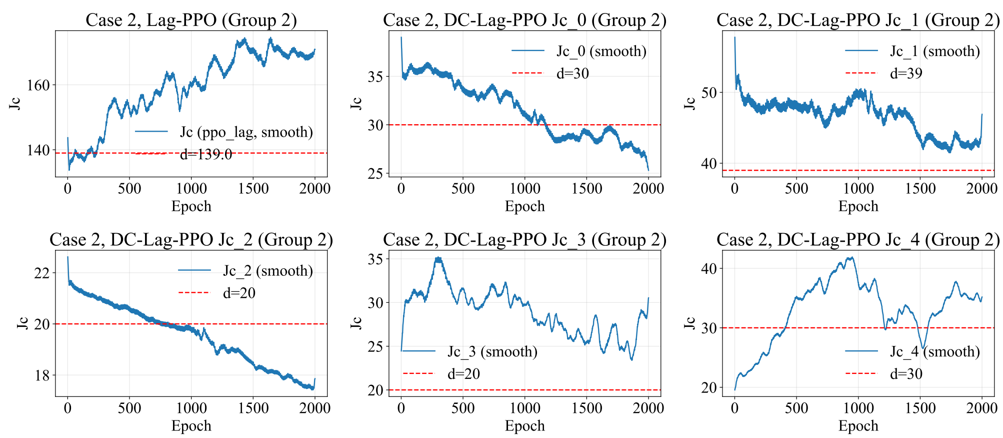


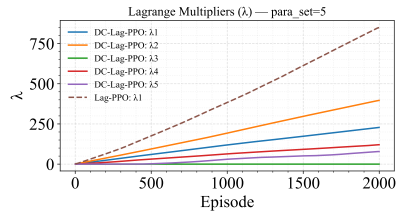
Figure 37: Training curves of J_c on Lagrangian cost threshold set: [30,39,20,20,30].

K.6 Lagrangian Multiplier Learning Curves

See Figure 38a.

K.7 Pareto Fronts

See Fig. 39. Across all reward-constraint and constraint-constraint pairs in this figure, DC-Lag-PPO exhibits consistently superior Pareto fronts. Its frontier lies uniformly closer to the lower-left region, indicating lower violations for comparable (or better) economic cost. Reward vs. all constraint: DC-Lag-PPO achieves strict dominance, producing solutions with both lower cost and lower violations, whereas Lag-PPO spreads widely and lacks a coherent frontier. Voltage-related pairs (c1-c2): DC-Lag-PPO forms a tighter and clearly improved front, showing better joint voltage safety. Battery



(a) Lagrangian cost threshold set [30,39,20,20,30].

Figure 38: Lagrangian multiplier λ learning curves.

degradation & dissatisfaction (c3-c4-c5): DC-Lag-PPO consistently pushes the front downward, reducing both degradation and unmet demand simultaneously. Lag-PPO fronts are often fragmented or upward-sloping, reflecting unstable trade-offs caused by aggregated-critic interference. Overall, the DC-Lag-PPO front either envelops or strictly improves upon Lag-PPO across all dimensions, confirming its ability to maintain safer and more efficient trade-offs under multi-constraint settings.

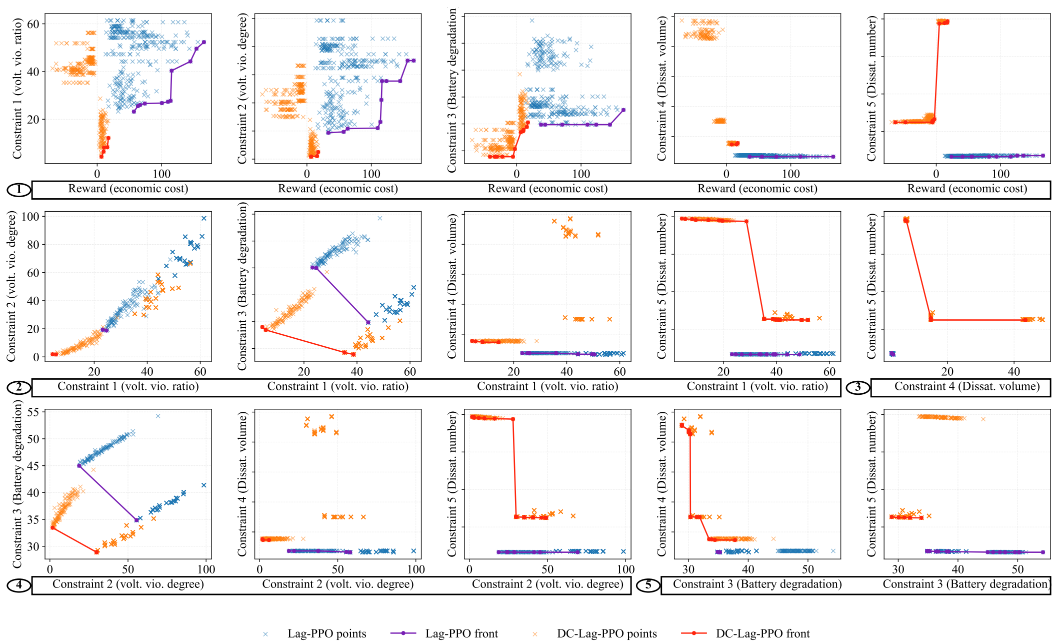


Figure 39: Pareto fronts from the test results on Lagrangian cost threshold set [30,39,20,20,30], where the black dotted lines are the thresholds d.

L NOTATION SUMMARY

We summarize the key notations used throughout the paper.

Symbol	Meaning
$\mathcal{S}$	State space of the CMDP.
$\mathcal{A}$	Action space of the CMDP.
$P(\cdot\|s,a)$	Transition kernel, probability of next state given (s, a).
$\gamma \in (0,1)$	Discount factor.
$r(s,a)$	Reward function.
$c_i(s,a)$	Cost function for constraint $i \in \{1,\dots,m\}$.
d_i	Threshold for constraint i.
π_θ	Stochastic policy parameterized by θ.
$\nabla_\theta \log \pi_\theta(a\|s)$	Policy score function.
$J_x(\pi_\theta)$	Expected discounted return of signal x.
$\mathcal{L}(\theta,\lambda)$	Lagrangian objective.
$\lambda = (\lambda_1,\dots,\lambda_m)$	Vector of Lagrange multipliers.
$Q^x_\pi(s,a)$	State–action value function for signal x under π.
T^x_π	Bellman operator for signal x.
$\phi(s,a) \in \mathbb{R}^d$	Feature vector for linear function approximation.
Φ	Feature matrix stacking $\phi(s,a)$ for all (s,a).
$Q_\omega(s,a) = \phi(s,a)^\top \omega$	Linear critic parameterized by ω.
D	Diagonal weighting matrix with stationary distribution $d_\pi(s,a)$.
$A(\theta)$	System matrix $\Phi^\top D(I - \gamma P_\pi)\Phi$.
$b^x(\theta)$	Right-hand side vector $\Phi^\top D\, x$.
$\omega^x(\theta)$	PBE solution for signal x: $A(\theta)\omega^x(\theta) = b^x(\theta)$.
$\omega^\lambda(\theta,\lambda)$	Mixed critic solution.
η_t	Critic stepsize.
α_t	Actor stepsize.
β_t	Dual stepsize.
θ_t	Actor parameters at iteration t.
λ_t	Dual variables at iteration t.
ω_t	Critic parameters at iteration t.
$\omega^\star_t$	Instantaneous mixed-critic target at iteration t,
e_t	Mixed critic error: $e_t = \omega_t - \omega^\star_t$.
e^x_t	Dedicated critic error: $e^x_t = \omega^x_t - \omega^x(\theta_t)$.
ζ_t, ζ^x_t	Martingale-difference noise terms in critic updates.
Δ^θ_t	Variation from changes in θ, eq. equation 4.
$\Delta^{\theta,x}_t$	Drift term for dedicated critic x.
$g^\star_t$	True actor gradient at iteration t.
$\widehat{g}_t$	Actor gradient estimate using mixed critic.
$\widehat{g}^{\text{multi}}_t$	Actor gradient estimate using dedicated-critic, eq. equation 62.
B_t	Actor-gradient bias (mixed critic): $\widehat{g}_t - g^\star_t$.
B^{multi}_t	Actor-gradient bias (dedicated-critic): $\widehat{g}^{\text{multi}}_t - g^\star_t$.
G	Uniform bound on $\\|\nabla_\theta \log \pi_\theta(a\|s)\\|$.
L_ϕ	Uniform bound on $\\|\phi(s,a)\\|$.
μ	Uniform lower bound on eigenvalues of $A(\theta)$.
$C_\lambda, C_\theta, \widetilde{C}_\theta$	Lipschitz / drift constants from error bounds.

Table 6: Performance comparison between dedicated-critic PPO-Lag and mixed-critic PPO-Lag. Values are mean ± standard deviation over evaluation episodes.

Env	Dedicated critics		Mixed critic	
	Reward	Cost	Reward	Cost
SafetyCarGoal1-v0	14.56 ± 8.97	21.72 ± 32.06	1.12 ± 9.23	55.34 ± 102.32
SafetyCarButton1-v0	0.36 ± 1.81	51.40 ± 82.14	1.51 ± 3.64	107.14 ± 132.22
SafetyAntVelocity-v1	3324.67 ± 83.21	13.01 ± 6.32	2821.72 ± 201.91	28.52 ± 8.37
SafetyHalfCheetahV-v1	3035.76 ± 287.42	4.14 ± 2.37	2234.245 ± 345.73	45.82 ± 7.15
SafetyHopperVelocity-v1	1002.73 ± 723.64	14.87 ± 20.74	1238.83 ± 465.35	17.21 ± 12.23
SafetyPointGoal1-v0	13.03 ± 7.15	23.97 ± 33.16	15.78 ± 3.18	52.81 ± 17.10

M Additional Experiments for the Environment with Single Constraints

To assess the practical impact of critic design on safe RL performance, we compare our dedicated-critic PPO-Lagrangian (separate value functions for reward and constraint cost) against a standard mixed-critic PPO-Lagrangian baseline across a diverse set of Safety-Gymnasium and velocity-control benchmarks. Specifically, we evaluate on the navigation tasks `SafetyCarGoal1-v0`, `SafetyCarButton1-v0`, and `SafetyPointGoal1-v0`, as well as the continuous-control environments `SafetyAntVelocity-v1`, `SafetyHalfCheetahVelocity-v1` (SafetyHalfCheetahV-v1), and `SafetyHopperVelocity-v1`, which together span both sparse goal-reaching rewards with collision costs and dense velocity-tracking settings with safety penalties. For each environment and method, we report the mean ± standard deviation of episodic reward and episodic cost over multiple evaluation rollouts after training, so that higher reward and lower cost indicate a better reward–safety trade-off. As summarised in Table 6, these experiments allow us to directly test whether separating reward and cost critics improves constraint satisfaction and stabilises learning compared to the widely used mixed-critic formulation.

Across all six benchmarks, the dedicated-critic PPO–Lagrangian consistently improves safety and often improves reward relative to the mixed-critic baseline. On the navigation tasks, `SafetyCarGoal1-v0` shows the clearest win: separating reward and cost critics raises the mean return from 1.12 to 14.56 while *also* reducing mean cost from 55.34 to 21.72, and it substantially shrinks the very large cost variance of the mixed critic. A similar pattern appears on `SafetyCarButton1-v0`: both methods obtain very low rewards (reflecting task difficulty), but the dedicated critic roughly halves the average cost (51.4 vs. 107.14) and reduces variability, indicating more reliable constraint satisfaction even when the policy is far from optimal. On `SafetyPointGoal1-v0`, the mixed critic achieves slightly higher reward (15.78 vs. 13.03) but at the price of more than double the mean cost (52.81 vs. 23.97), so the dedicated critic offers a strictly safer solution with only a modest reward gap.

The MuJoCo velocity environments highlight the benefit of dedicated critics even more strongly. On `SafetyAntVelocity-v1`, the dedicated-critic agent improves reward from 2821.72 to 3324.67 *and* cuts mean cost by more than half (28.52 to 13.01). On `SafetyHalfCheetahVelocity-v1`, the effect is even more pronounced: reward increases from 2234.25 to 3035.76, while cost drops from 45.82 to 4.14, giving a dramatically better reward–safety trade-off. `SafetyHopperVelocity-v1` is the only case where the mixed critic slightly outperforms in reward (1238.83 vs. 1002.73), but the dedicated critic still attains lower cost (14.87 vs. 17.21) and comparable variance. Overall, these results align with our theoretical claim: by removing the λ-induced target drift, dedicated critics provide more stable value estimates, which in practice translates into systematically lower constraint violations and, in most tasks, equal or higher task performance than the mixed-critic formulation.

N Asymptotic Vanishing of Mixed–Critic Bias

Remark N.1 (Asymptotic vanishing of mixed–critic bias). Under Assumption 4.1, the step–size ratios satisfy $\alpha_t/\eta_t \to 0$ and $\beta_t/\eta_t \to 0$ as $t \to \infty$. Hence both limsup terms on the right-hand

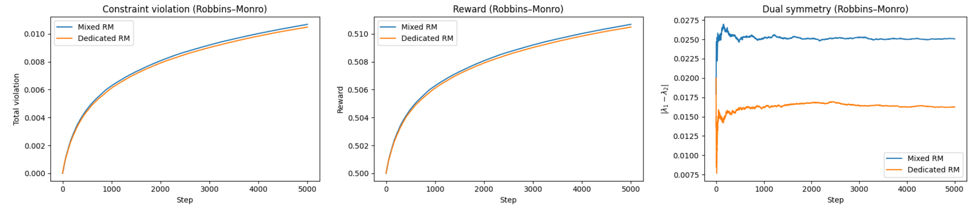


Figure 40: Robbins–Monro Lag-PPO in a two-constraint CMDP: comparison between mixed and dedicated critics

side of Theorem 4.7 are zero, and we obtain $\limsup_{t\to\infty} \mathbb{E}\big[\|B_t\|\big] = 0$. That is, in the idealised linear SA regime of Assumption 4.1, the mixed–critic actor-gradient bias vanishes asymptotically, so mixed–critic Lag-PPO and its dedicated–critic variant coincide in the limit. By contrast, this regime requires Robbins–Monro step sizes and strong timescale separation between critic, actor, and dual, which is *not* how deep PPO–Lagrangian is used in practice with (effectively) constant learning rates and Adam. Our experiments follow this practical setting, where a non-negligible mixed–critic bias can persist, which is why Theorem 4.7 is stated in terms of the ratios β_t/η_t and α_t/η_t, keeping the bound informative for realistic, non-asymptotic schedules.

To more directly connect our empirical results to the asymptotic setting in Assumption 4.1 and Theorem 4.7, we also study a Robbins–Monro (RM) variant in a simplified, idealised environment. The goal of these experiments is to approximate the stochastic-approximation regime assumed in the theory and to compare mixed and dedicated critics under those conditions.

Experimental design. We consider a small constrained MDP with two constraints and a low-dimensional state and action space, for which we can reliably measure reward, total constraint violation, and dual-variable behaviour over training(same as Appendix H). In this setting we implement two variants:

- **Mixed RM**: a single mixed critic trained on the scalarised signal $r_\lambda = r - \sum_i \lambda_i c_i$.
- **Dedicated RM**: separate critics V_r and V_{c_i} trained on reward and each constraint cost, respectively.

Both variants use the *same* data (trajectories), and differ only in how the value function is parameterised. To align with Assumption 4.1, we use Robbins–Monro learning-rate schedules for the actor, critics, and dual variables:

$$\alpha_t = \frac{\alpha_0}{1 + k_\pi t}, \quad \eta_t = \frac{\eta_0}{1 + k_V t}, \quad \beta_t = \frac{\beta_0}{1 + k_\lambda t},$$

chosen such that $\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$ (and similarly for η_t, β_t), and with a clear time-scale separation $\eta_t \gg \alpha_t \gg \beta_t$. All other hyperparameters are held fixed across the two variants. We track three metrics over training:

1. expected reward (per step),
2. total constraint violation (per step),
3. the absolute difference between dual variables (to monitor symmetry of the Lagrange multipliers in the two-constraint case).

Figure 40 shows these three quantities for both Mixed RM and Dedicated RM.

In this idealised, near-linear Robbins–Monro setting, the mixed-critic and dedicated-critic variants behave very similarly. Their reward curves almost overlap, both methods achieve comparable levels of constraint satisfaction, and the dual variables converge to a very similar symmetric configuration. This is consistent with Theorem 4.7: when the learning rates satisfy the stochastic-approximation conditions and the critics can track their targets on the fastest time scale, the additional dual-induced drift term in the mixed critic does not translate into a noticeable difference in the limiting behaviour. In other words, this toy Robbins–Monro experiment can be seen as a direct empirical realisation of the asymptotic linear-theory predictions, and it serves as a sanity check that our finite-sample implementation matches the behaviour analysed in the theoretical section.

O ROBBINS–MONRO LAG-PPO

In RM Lag-PPO, we replace Adam with plain SGD and use diminishing step-size schedules for the actor, critic, and dual updates. The data-collection and PPO objective remain unchanged; only the optimiser and step-size schedules differ. We instantiate a *mixed* RM Lag-PPO, Lag-PPO (one critic) and a *Dedicated* RM variant (separate critics), and train them on the complex power system environments. However, as we cannot run for effectively unlimited time, and strict Robbins–Monro schedules cause step sizes to become very small within a realistic training budget, the RM Lag-PPO variant performs quite similarly with the standard constant–stepsize Lag-PPO(Figure 41). This similarity is further reinforced by PPO's ratio clipping (and gradient clipping), which effectively bounds the size of each policy update even under a nominally constant learning rate, making standard Lag-PPO behave in practice like a conservatively damped method whose effective step sizes are not far from those induced by a Robbins–Monro schedule over a finite training horizon.

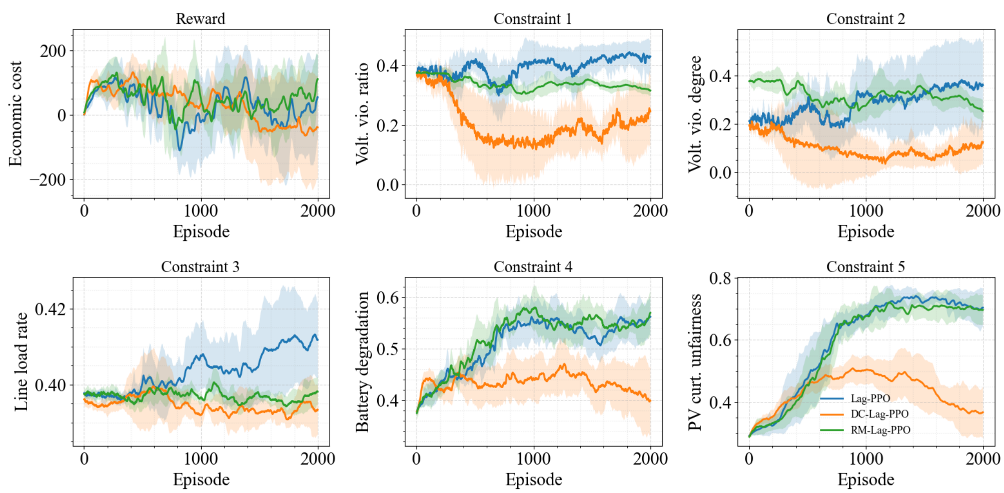


Figure 41: Training curves of Lag-PPO.

P STATEMENT ON LLM USAGE

Large language models (LLMs), such as ChatGPT, were used solely for editorial assistance in this work. Their role was limited to improving grammar, rephrasing sentences, and enhancing clarity and readability of the authors' original text. No LLM was used to generate original scientific content, analysis, or results. The authors take full responsibility for the integrity and validity of the work presented.