Learning Deterministic Policies with Policy Gradients in Constrained Markov Decision Processes

Alessandro Montenegro[a,*], Leonardo Cesani[a], Marco Mussi[a], Matteo Papini[a] and Alberto Maria Metelli[a]

[a]Politecnico di Milano, Piazza Leonardo Da Vinci 32, Milan, 20133, Italy

ARTICLE INFO

Keywords:
Reinforcement Learning
Constrained RL
Last-Iterate Global Convergence
Deterministic Policies
Policy Gradients
Primal-Dual

Abstract

Constrained Reinforcement Learning (CRL) addresses sequential decision-making problems where agents are required to achieve goals by maximizing the expected return while meeting domain-specific constraints. In this setting, *policy-based* methods are widely used thanks to their advantages when dealing with continuous-control problems. These methods search in the policy space with an *action-based* or a *parameter-based* exploration strategy, depending on whether they learn the parameters of a stochastic policy or those of a stochastic hyperpolicy. We introduce an exploration-agnostic algorithm, called `C-PG`, which enjoys *global last-iterate convergence* guarantees under gradient domination assumptions. Furthermore, under specific noise models where the (hyper)policy is expressed as a *stochastic perturbation* of the actions or of the parameters of an underlying deterministic policy, we additionally establish global last-iterate convergence guarantees of `C-PG` to the *optimal deterministic policy*. This holds when learning a stochastic (hyper)policy and subsequently *switching off the stochasticity* at the *end* of training, thereby deploying a deterministic policy. Finally, we empirically validate both the action-based (`C-PGAE`) and parameter-based (`C-PGPE`) variants of `C-PG` on constrained control tasks, and compare them against state-of-the-art baselines, demonstrating their effectiveness, in particular when deploying deterministic policies after training.[1]

1. Introduction

When applying Reinforcement Learning (RL, Sutton and Barto, 2018) to real-world scenarios, we aim at solving large-scale continuous control problems where, in addition to reaching a goal, it is necessary to meet structural or utility-based constraints. For instance, an autonomous-driving car has its main objective of getting to the desired destination (i.e., goal) while avoiding collisions, ensuring the safety of people on the streets, adhering to traffic rules, and respecting the physical requirements of the engine to avoid damaging it (i.e., constraints) (Likmeta, Metelli, Tirinzoni, Giol, Restelli and Romano, 2020). To pursue such an objective, it is necessary to extend the RL problem formulation with the possibility to account for constraints. Constrained Reinforcement Learning (CRL, Uchibe and Doya, 2007) aims at solving this family of problems by employing RL techniques to tackle Constrained Markov Decision Processes (CMDPs, Altman, 1999), which provide an established and widely-used framework for modeling constrained control tasks. The conventional CRL framework primarily focuses on constraints related directly to *expected costs* (Stooke, Achiam and Abbeel, 2020; Ding, Zhang, Basar and Jovanovic, 2020; Ying, Ding and Lavaei, 2022; Ding, Wei, Zhang and Ribeiro, 2024).

Among the RL methods applicable to CMDPs, Policy Gradients (PGs, Deisenroth, Neumann and Peters, 2013) are particularly appealing. Indeed, PGs have demonstrably achieved impressive results in continuous-control problems due to several advantages that make them well-suited for real-world applications. These advantages include the ability to handle continuous state and action spaces (Peters and Schaal, 2006), resilience to sensor and actuator noise (Gravell, Esfahani and Summers, 2020), robustness in partially-observable environments (Azizzadenesheli, Yue and Anandkumar, 2018), and the possibility of incorporating expert knowledge during the policy design

[1]This work extends the preliminary version presented in (Montenegro, Mussi, Papini and Metelli, 2024b) by providing a theoretical analysis and empirical evaluation of deterministic policy deployment.

*Corresponding author

alessandro.montenegro@polimi.it (A. Montenegro); leonardo.cesani@polimi.it (L. Cesani); marco.mussi@polimi.it (M. Mussi); matteo.papini@polimi.it (M. Papini); albertomaria.metelli@polimi.it (A.M. Metelli)

ORCID(s): 0009-0000-2034-7106 (A. Montenegro); 0009-0009-9329-5349 (L. Cesani); 0000-0001-8356-6744 (M. Mussi); 0000-0002-3807-3171 (M. Papini); 0000-0002-3424-5212 (A.M. Metelli)

phase (Ghavamzadeh and Engel, 2006), thus improving the efficacy, safety, and interpretability of the learned policy (Likmeta et al., 2020). PGs can be categorized into two key families depending on the way exploration is carried out in the policy space (Montenegro, Mussi, Metelli and Papini, 2024a). Following their taxonomy, we distinguish between the *action-based* (AB) and the *parameter-based* (PB) exploration paradigms. The former, employed by `REINFORCE` (Williams, 1992) and `GPOMDP` (Baxter and Bartlett, 2001), focuses on directly learning the parameters of a parametric stochastic *policy*. The latter, employed by `PGPE` (Sehnke, Osendorfer, Rückstieß, Graves, Peters and Schmidhuber, 2010), is tasked with learning the parameters of a parametric stochastic *hyperpolicy* from which the parameters of the actual policy (often deterministic) are sampled.

PGs have gained significant popularity in solving constrained control problems (Achiam, Held, Tamar and Abbeel, 2017). Within this field, algorithms are primarily developed using *primal-dual* methods (Chow, Ghavamzadeh, Janson and Pavone, 2017; Tessler, Mankowitz and Mannor, 2019; Ding et al., 2020; Ding, Wei, Yang, Wang and Jovanovic, 2021; Bai, Bedi, Agarwal, Koppel and Aggarwal, 2022), which can be formulated through *Lagrangian optimization* of the primal (i.e., policy or hyperpolicy parameters) and dual variable (i.e., Lagrange multipliers). Even though the distinction between the exploration paradigms is well known in the PG methods literature, the current state of the art in policy-based CRL focuses only on the *action-based* exploration approach (Achiam et al., 2017; Stooke et al., 2020; Bai, Bedi and Aggarwal, 2023), while the *parameter-based* one remains unexplored (Montenegro et al., 2024b). A critical challenge for policy-based Lagrangian optimization algorithms is ensuring convergence guarantees. Existing works have spent a notable effort in this direction (Ying et al., 2022; Gladin, Lavrik-Karmazin, Zainullina, Rudenko, Gasnikov and Takác, 2023; Ding et al., 2024). Recently, Ying et al. (2022), Gladin et al. (2023), and Ding et al. (2024) manage to ensure *global last-iterate* convergence guarantees. However, these approaches are affected by some notable limitations: (i) the provided convergence rates depend on the cardinality of the state and action spaces, limiting their applicability to tabular CMPDs and preventing scaling to realistic continuous control problems; (ii) they focus on *softmax* policies only, disregarding other more realistic policy models (e.g., Gaussian ones); (iii) ensure convergence when a single constraint only is present (Ding et al., 2024; Rozada, Ding, Marques and Ribeiro, 2025).

Real-world problems not only require RL algorithms to produce policies satisfying constraints, but they also often demand the resulting policy to be deterministic to meet reliability, safety, and traceability requirements. To this end, PGs remain a recommended choice. Considering unconstrained scenarios, the challenge of learning deterministic policies was first addressed by (Silver, Lever, Heess, Degris, Wierstra and Riedmiller, 2014), that introduced the *deterministic policy gradient* (DPG) method, later inspiring successful deep RL algorithms such as DDPG (Lillicrap, Hunt, Pritzel, Heess, Erez, Tassa, Silver and Wierstra, 2015; Fujimoto, van Hoof and Meger, 2018). However, DPG-based approaches present notable drawbacks due to their off-policy nature, which makes the theoretical analysis complex and limits local convergence guarantees to restrictive assumptions (Xiong, Xu, Zhao, Liang and Zhang, 2022). More recently, Montenegro et al. (2024a) proposed a unified framework for deterministic policy deployment that bridges action-based and parameter-based exploration paradigms. Their approach is grounded on specific noise models that represent stochastic policies and hyperpolicies as *perturbations* of the actions or the parameters of an underlying deterministic policy. The core idea is to *train stochastic* (hyper)policies via policy gradient algorithms and *deploy* their *deterministic* counterparts by *switching off* the stochasticity, thereby offering a principled methodology for incorporating deterministic policies within the policy gradient framework. The latter contribution is a recent advancement in the unconstrained setting, while in the constrained one, despite the advancements in policy-based CRL, the integration of deterministic policies within this framework remains mostly unexplored. A recent contribution in this direction is presented by Rozada et al. (2025), who introduce the *Deterministic Policy Gradient Primal-Dual* (D-PGPD) algorithm, a novel approach designed to directly learn deterministic policies in CMDPs with continuous state and action spaces. D-PGPD is a primal-dual algorithm that incorporates entropy regularization w.r.t. the policy and ridge regularization w.r.t. the dual variable. Unlike traditional stochastic policy-based CRL methods, which introduce exploration through policy randomness, D-PGPD learns a fully deterministic policy and relies solely on the inherent stochasticity of the environment. While this design ensures stable and consistent policy execution, it may face limitations in environments where intrinsic stochasticity is insufficient to explore the state-action space effectively. Moreover, the algorithm is developed for single-constraint settings, which may reduce its applicability to real-world problems involving multiple constraints. In the considered setup, and under additional assumptions, such as boundedness of the action space, the sample-based version of D-PGPD requires a sample complexity of order $\mathcal{O}(\epsilon^{-18})$ to the optimal feasible deterministic policy in the last iterate. This unsatisfactory rate highlights the need for further research on more efficient methods.

Original Contribution. The goal of this work is to introduce a framework for solving constrained continuous control problems using policy-based primal-dual algorithms that operate in both the *action-based* and *parameter-based* policy gradient exploration scenarios, while providing global last-iterate convergence guarantees with general (hyper)policy parameterization to both optimal stochastic (hyper)policies and deterministic policies. Specifically, the main contributions of this work can be summarized as follows:

- In Section 2, we introduce a general constrained optimization problem, which is agnostic w.r.t. both the *action-based* and *parameter-based* exploration paradigm.
- In Section 3, we introduce `C-PG`, a general policy-based primal-dual algorithm optimizing the regularized Lagrangian function associated with the general constrained optimization problem shown in Section 2. We show that, under (weak) gradient domination assumptions, it simultaneously achieves the following: (i) *last-iterate* convergence guarantees to a globally optimal feasible policy (i.e., satisfying all the constraints); (ii) compatibility with CMDPs having *continuous state and action spaces*; (iii) the ability to handle *multiple constraints*.
- In Section 4, we restrict action-based and parameter-based exploration paradigms as white-noise perturbations applied to the actions or parameters of an underlying parametric deterministic policy. Based on this characterization, we define *deterministic deployment* as the process of *switching off* the noise (Montenegro et al., 2024a) in the learned stochastic (hyper)policies. Within this framework, we derive all the conditions required to ensure the last-iterate global convergence of `C-PG`, as presented in Section 3. Finally, we show that this approach guarantees last-iterate convergence to the optimal deterministic policy in the constrained setting.

In Section 6, we numerically validate the parameter-based and the action-based variants of `C-PG` against state-of-the-art baselines in constrained control problems. Related work is discussed in Section 5. Omitted proofs and additional technical results are reported in Appendix A.

2. Preliminaries

In this section, we present the notation we will use throughout this manuscript and the preliminaries needed to understand its content. Moreover, after having introduced the the AB and PB exploration paradigms, we introduce the exploration-agnostic constrained optimization problem we aim at solving with the introduced method.

Notation. For a measurable set $\mathcal{X}$, we denote as $\Delta(\mathcal{X})$ the set of probability measures over $\mathcal{X}$. For $P \in \Delta(\mathcal{X})$, we denote with p its density function w.r.t. a reference measure that we assume to exist whenever needed. With a little abuse of notation, we will interchangeably use $x \sim P$ or $x \sim p$ to express that random variable x is distributed according to P. For $n, m \in \mathbb{N}$ with $n \leq m$, we denote $[\![n]\!] := \{1, 2, \dots, n\}$ and with $[\![n, m]\!] := \{n, n+1, \dots, m\}$. For a vector $\boldsymbol{x} \in \mathbb{R}^d$, we denote as x_i the i-th component of $\boldsymbol{x}$. For $a \in \mathbb{R}$, we define $(a)^+ := \max\{0, a\}$ and we extend the notation to vectors as $(\boldsymbol{x})^+ = ((x_1)^+, \dots, (x_d)^+)^\top$. Given a set $\mathcal{X} \subseteq \mathbb{R}^d$, we denote with $\Pi_{\mathcal{X}}$ the Euclidean-norm projection, i.e., $\Pi_{\mathcal{X}} \boldsymbol{x} \in \arg\min_{\boldsymbol{y} \in \mathcal{X}} \|\boldsymbol{y} - \boldsymbol{x}\|_2$ for any $\boldsymbol{x} \in \mathbb{R}^d$. For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, we denote with $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ their inner product. A function $f : \mathbb{R}^d \to \mathbb{R}$ is L_1-Lipschitz Continuous (L_1-LC) if $|f(\boldsymbol{x}) - f(\boldsymbol{x}')| \leq L_1 \|\boldsymbol{x} - \boldsymbol{x}'\|_2$ and L_2-Lipschitz Smooth (L_2-LS) if it is differentiable and $\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}')\|_2 \leq L_2 \|\boldsymbol{x} - \boldsymbol{x}'\|_2$ for every $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$.

Constrained Markov Decision Processes. A Constrained Markov Decision Process (CMDP, Altman, 1999) with $U \in \mathbb{N}$ constraints is represented by $\mathcal{M}_C := \left(\mathcal{S}, \mathcal{A}, p, r, \{c_i\}_{i \in [\![U]\!]}, \{b_i\}_{i \in [\![U]\!]}, \phi_0, \gamma\right)$, where $\mathcal{S} \subseteq \mathbb{R}^{d_{\mathcal{S}}}$ and $\mathcal{A} \subseteq \mathbb{R}^{d_{\mathcal{A}}}$ are the measurable state and action spaces; $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition model, where $p(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a})$ is the probability density of getting to state $\boldsymbol{s}' \in \mathcal{S}$ given that action $\boldsymbol{a} \in \mathcal{A}$ is taken in state $\boldsymbol{s} \in \mathcal{S}$; $r : \mathcal{S} \times \mathcal{A} \to [-1, 0]$ is the reward function, where $r(\boldsymbol{s}, \boldsymbol{a})$ is the instantaneous reward obtained by playing action $\boldsymbol{a}$ in state $\boldsymbol{s}$; $c_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the i-th cost function, where $c_i(\boldsymbol{s}, \boldsymbol{a})$ is the i-th instantaneous cost obtained by playing action $\boldsymbol{a}$ in state $\boldsymbol{s}$; $b_i \in [0, J_{\max}]$ is the threshold for the i-th cost for every $i \in [\![U]\!]$; $\phi_0 \in \Delta(\mathcal{S})$ is the initial state distribution; and $\gamma \in [0, 1]$ is the discount factor. A trajectory τ of length $T \in \mathbb{N} \cup \{+\infty\}$[1] is a sequence of T state-action pairs: $\tau = \left(\boldsymbol{s}_{\tau,0}, \boldsymbol{a}_{\tau,0}, \dots, \boldsymbol{s}_{\tau,T-1}, \boldsymbol{a}_{\tau,T-1}\right)$. The *discounted return* over a trajectory τ is $R(\tau) := \sum_{t=0}^{T-1} \gamma^t r(\boldsymbol{s}_{\tau,t}, \boldsymbol{a}_{\tau,t})$, while the i-th *discounted cumulative cost* is $C_i(\tau) := \sum_{t=0}^{T-1} \gamma^t c_i(\boldsymbol{s}_{\tau,t}, \boldsymbol{a}_{\tau,t})$. We define the additional cost function $c_0(\boldsymbol{s}, \boldsymbol{a}) := -r(\boldsymbol{s}, \boldsymbol{a}) \in [0, 1]$ and $C_0(\tau) := -R(\tau)$

[1] We admit $\gamma = 1$ just when $T < +\infty$.

just for presentation purposes. Note that, with $J_{\max} := \frac{1-\gamma^T}{1-\gamma}$, $R(\tau) \in [-J_{\max}, 0]$ and $C_i(\tau) \in [0, J_{\max}]$, for every $i \in [\![U]\!]$ and trajectory τ. Our goal is to minimize $\mathbb{E}[C_0(\tau)]$ subject to the constraints $\mathbb{E}[C_i(\tau)] \le b_i$ for every $i \in [\![U]\!]$.

Action-based Policy Gradients. Action-based (AB) PG methods focus on learning the parameters $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{d_\Theta}$ of a parametric stochastic policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$, where $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})$ represents the probability density of selecting action $\boldsymbol{a} \in \mathcal{A}$ being in state $\boldsymbol{s} \in \mathcal{S}$. At each step t of the interaction with the environment, the stochastic policy is employed to sample an action $\boldsymbol{a}_t \sim \pi_{\boldsymbol{\theta}_t}(\cdot|\boldsymbol{s}_t)$. To assess the performance of $\pi_{\boldsymbol{\theta}}$ w.r.t. the i-th cost function, with $i \in [\![0, U]\!]$, we employ the *AB performance index* $J_{\mathrm{A},i} : \Theta \to \mathbb{R}$, which is defined as $J_{\mathrm{A},i}(\boldsymbol{\theta}) := \mathbb{E}_{\tau \sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})}\left[C_i(\tau)\right]$, where $p_{\mathrm{A}}(\tau, \boldsymbol{\theta}) := \phi_0(\boldsymbol{s}_{\tau,0}) \prod_{t=0}^{T-1} \pi_{\boldsymbol{\theta}}(\boldsymbol{a}_{\tau,t}|\boldsymbol{s}_{\tau,t}) p(\boldsymbol{s}_{\tau,t+1}|\boldsymbol{s}_{\tau,t}, \boldsymbol{a}_{\tau,t})$ is the density of trajectory τ induced by policy $\pi_{\boldsymbol{\theta}}$.

Parameter-based Policy Gradients. Parameter-based (PB) PG methods focus on learning the parameters $\boldsymbol{\rho} \in \mathcal{R} \subseteq \mathbb{R}^{d_\mathcal{R}}$ of a parametric stochastic hyperpolicy $\nu_{\boldsymbol{\rho}} \in \Delta(\Theta)$. The hyperpolicy $\nu_{\boldsymbol{\rho}}$ is used to sample parameter configurations $\boldsymbol{\theta} \sim \nu_{\boldsymbol{\rho}}$ to be plugged into an underlying parametric policy $\pi_{\boldsymbol{\theta}}$, that will be then used for the interaction with the environment. Notice that $\pi_{\boldsymbol{\theta}}$ can also be deterministic. To assess the performance of $\nu_{\boldsymbol{\rho}}$ w.r.t. the i-th cost function, with $i \in [\![0, U]\!]$, we employ the *PB performance index* $J_{\mathrm{P},i} : \mathcal{R} \to \mathbb{R}$, which is defined as $J_{\mathrm{P},i}(\boldsymbol{\rho}) := \mathbb{E}_{\boldsymbol{\theta} \sim \nu_{\boldsymbol{\rho}}}\left[\mathbb{E}_{\tau \sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})}\left[C_i(\tau)\right]\right]$.

Constrained Optimization Problem. Having introduced the AB and PB performance indices, we formulate a *constrained optimization problem* (COP), which is agnostic w.r.t. the exploration paradigm:

$$\min_{\boldsymbol{v} \in \mathcal{V}} J_{\dagger,0}(\boldsymbol{v}) \quad \text{s.t.} \quad J_{\dagger,i}(\boldsymbol{v}) \le b_i, \quad \forall i \in [\![U]\!], \tag{1}$$

where $\dagger \in \{\mathrm{A}, \mathrm{P}\}$ and $\boldsymbol{v}$ is a generic parameter vector belonging to the parameter space $\mathcal{V}$. When $\dagger = \mathrm{A}$, we are considering the AB exploration paradigm, so $\mathcal{V} = \Theta$. On the other hand, when $\dagger = \mathrm{P}$, we are in the PB exploration paradigm, thus $\mathcal{V} = \mathcal{R}$.

3. Last-Iterate Global Convergence of C-PG

In this section, we present C-PG, a general primal-dual algorithm that optimizes a regularized version of the Lagrangian function (Section 3.1) associated with the COP of Equation (1). After having introduced the necessary assumptions (Section 3.2), we show that C-PG exhibit *dimension-free*[2] *last-iterate global* convergence guarantees (Section 3.3). For notational convenience, in the rest of this section, we use J_i in place of $J_{\dagger,i}$.

3.1. Regularized Lagrangian Approach

To solve the COP of Equation (1) we resort to the method of Lagrange multipliers (Bertsekas, 2014) introducing the Lagrangian function $\mathcal{L}_0(\boldsymbol{v}, \lambda) := J_0(\boldsymbol{v}) + \sum_{i=1}^{U} \lambda_i \left(J_i(\boldsymbol{v}) - b_i\right) = J_0(\boldsymbol{v}) + \langle \lambda, \mathbf{J}(\boldsymbol{v}) - \mathbf{b}\rangle$, where $\boldsymbol{v} \in \mathcal{V}$ is the primal variable and $\lambda \in \mathbb{R}^U_{\ge 0}$ are the Lagrangian multipliers or dual variable, $\mathbf{J} = (J_1, \dots, J_U)^\top$, and $\mathbf{b} = (b_1, \dots, b_U)^\top$. This allows rephrasing the COP in Equation (1) as a min-max optimization problem $\min_{\boldsymbol{v} \in \mathcal{V}} \max_{\lambda \in \mathbb{R}^U_{\ge 0}} \mathcal{L}_0(\boldsymbol{v}, \lambda)$ and we denote with $H_0(\boldsymbol{v}) := \max_{\lambda \in \mathbb{R}^U_{\ge 0}} \mathcal{L}_0(\boldsymbol{v}, \lambda)$ the *primal function* and its optimum with $H_0^* := \min_{\boldsymbol{v} \in \mathcal{V}} H_0(\boldsymbol{v})$. To obtain a *last-iterate* convergence guarantee, we make use of a regularization approach. Specifically, let $\omega > 0$ be a regularization parameter, we define the *ω-regularized Lagrangian function* as follows:

$$\mathcal{L}_\omega(\boldsymbol{v}, \lambda) := J_0(\boldsymbol{v}) + \sum_{i=1}^{U} \lambda_i \left(J_i(\boldsymbol{v}) - b_i\right) - \frac{\omega}{2} \|\lambda\|_2^2 = J_0(\boldsymbol{v}) + \langle \lambda, \mathbf{J}(\boldsymbol{v}) - \mathbf{b}\rangle - \frac{\omega}{2} \|\lambda\|_2^2 = \mathcal{L}_0(\boldsymbol{v}, \lambda) - \frac{\omega}{2} \|\lambda\|_2^2 .$$

The ridge regularization makes $\mathcal{L}_\omega(\boldsymbol{v}, \lambda)$ a strongly concave function of λ at the price of a bias that is quantified in Lemmas A.1, A.2, and A.3. Thus, we address the ω-regularized min-max optimization problem $\min_{\boldsymbol{v} \in \mathcal{V}} \max_{\lambda \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}, \lambda)$, where $\Lambda := \{\lambda \in \mathbb{R}^U_{\ge 0} : \|\lambda\|_2 \le \omega^{-1}\sqrt{U} J_{\max}\}$, in replacement of the original (non-regularized) one. We stress that this choice of Λ guarantees that the optimal Lagrange multipliers λ^*_ω lie within Λ. For this problem, we introduce

[2]The *dimension-free* property (Liu, Zhou, Kalathil, Kumar and Tian, 2021; Ding et al., 2020; Ding, Zhang, Duan, Başar and Jovanović, 2022; Ding et al., 2024) is achieved when the convergence rates do not depend on the cardinality of the state and/or action spaces.

the primal function $H_\omega(\boldsymbol{v}) := \max_{\boldsymbol{\lambda}\in\Lambda} \mathcal{L}_\omega(\boldsymbol{v},\boldsymbol{\lambda})$, that, thanks to the ridge regularization, admits the closed-form expression:

$$H_\omega(\boldsymbol{v}) = J_0(\boldsymbol{v}) + \frac{1}{2\omega}\sum_{i=1}^{U}\left(\left(J_i(\boldsymbol{v}) - b_i\right)^+\right)^2 = J_0(\boldsymbol{v}) + \frac{1}{2\omega}\|(\mathbf{J}(\boldsymbol{v}) - \mathbf{b})^+\|_2^2,$$

where the optimal values of the Lagrange multipliers are given by:

$$\boldsymbol{\lambda}^*(\boldsymbol{v}) = \Pi_\Lambda\left(\frac{1}{\omega}(\mathbf{J}(\boldsymbol{v}) - \mathbf{b})\right) = \frac{1}{\omega}(\mathbf{J}(\boldsymbol{v}) - \mathbf{b})^+,$$

that is guaranteed to have norm smaller than $\omega^{-1}\sqrt{U}J_{\max}$. Furthermore, we define $H_\omega^* := \min_{\boldsymbol{v}\in\mathcal{V}} H_\omega(\boldsymbol{v})$. C-PG updates the parameters $(\boldsymbol{v}_k, \boldsymbol{\lambda}_k)$ with an *alternate gradient descent-ascent* scheme for every iterate $k \in \mathbb{N}$:

$$\textbf{Primal Update}: \quad \boldsymbol{v}_{k+1} \leftarrow \Pi_{\mathcal{V}}\left(\boldsymbol{v}_k - \zeta_{\boldsymbol{v},k}\widehat{\nabla}_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)\right),$$

$$\textbf{Dual Update}: \quad \boldsymbol{\lambda}_{k+1} \leftarrow \Pi_{\Lambda}\left(\boldsymbol{\lambda}_k + \zeta_{\boldsymbol{\lambda},k}\widehat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}_\omega(\boldsymbol{v}_{k+1},\boldsymbol{\lambda}_k)\right),$$

where $\zeta_{\boldsymbol{v},k}, \zeta_{\boldsymbol{\lambda},k} > 0$ are the learning rates and $\widehat{\nabla}_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)$, $\widehat{\nabla}_{\boldsymbol{\lambda}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)$ are (unbiased) estimators of the gradients $\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k), \nabla_{\boldsymbol{\lambda}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)$ of the regularized Lagrangian function. Notice that C-PG performs *alternate* descent-ascent, as the update value for the dual variable is performed employing the *already updated* primal variable.

3.2. Assumptions

Before diving into the study of the convergence guarantees of C-PG, we list and motivate the assumptions necessary for our analysis.

Assumption 3.1 (Existence of Saddle Points)**.** *There exist $\boldsymbol{v}_0^* \in \mathcal{V}$ and $\boldsymbol{\lambda}_0^* \in \mathbb{R}_{\geq 0}^U$ such that $\mathcal{L}_0(\boldsymbol{v}_0^*,\boldsymbol{\lambda}_0^*) = \min_{\boldsymbol{v}\in\mathcal{V}}\max_{\boldsymbol{\lambda}\in\mathbb{R}_{\geq 0}^U}\mathcal{L}_0(\boldsymbol{v},\boldsymbol{\lambda})$.*

Assumption 3.1 ensures that the value of the min-max problem is attained by a pair of primal-dual values $\boldsymbol{v}_0^* \in \mathcal{V}$ and $\boldsymbol{\lambda}_0^* \in \mathbb{R}_{\geq 0}^U$ which, consequently, satisfy $\mathcal{L}_0(\boldsymbol{v}_0^*,\boldsymbol{\lambda}) \leq \mathcal{L}_0(\boldsymbol{v}_0^*,\boldsymbol{\lambda}_0^*) \leq \mathcal{L}_0(\boldsymbol{v},\boldsymbol{\lambda}_0^*)$ for every $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{\lambda} \in \mathbb{R}_{\geq 0}^U$. Analogous assumptions have been considered by Yang, Kiyavash and He (2020) and Ying et al. (2022). Thus, $(\boldsymbol{v}_0^*,\boldsymbol{\lambda}_0^*)$ is a saddle point of the Lagrangian function $\mathcal{L}_0$ and, consequently, *strong duality* holds. Alternatively, as commonly requested in CRL works, assuming *Slater's condition* combined with the requirement that the policy space covers all Markovian policies ensures strong duality (e.g., Paternain, Chamon, Calvo-Fullana and Ribeiro, 2019; Ding et al., 2020, 2024).[3]

Assumption 3.2 (Weak ψ-Gradient Domination)**.** *Let $\psi \in [1,2]$. There exist $\alpha_1 \in \mathbb{R}_{>0}$ and $\beta_1 \in \mathbb{R}_{\geq 0}$ such that, for every $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{\lambda} \in \Lambda$, it holds that:*

$$\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_0(\boldsymbol{v},\boldsymbol{\lambda})\right\|_2^\psi \geq \alpha_1\left(\mathcal{L}_0(\boldsymbol{v},\boldsymbol{\lambda}) - \min_{\boldsymbol{v}'\in\mathcal{V}}\mathcal{L}_0(\boldsymbol{v}',\boldsymbol{\lambda})\right) - \beta_1. \tag{2}$$

Assumption 3.2 is customary in the convergence analysis of policy gradient methods and it is usually enforced on the objective J_0 only (Yuan, Gower and Lazaric, 2022; Masiha, Salehkaleybar, He, Kiyavash and Thiran, 2022; Fatkhullin, Barakat, Kireeva and He, 2023). In particular, when $\beta_1 = 0$, we speak of strong ψ-gradient domination. In this form, for a generic exponent $\psi \in [1,2]$, this assumption has been employed by Masiha et al. (2022). Particular cases are $\psi = 1$, which corresponds to the standard weak *gradient domination* (GD), while for $\psi = 2$ we have the so-called *Polyak-Łojasiewicz* (PL) condition. Notice that Assumption 3.2 is enforced on the non-regularized Lagrangian function $\mathcal{L}_0$ (i.e., $\omega = 0$). However, it is easy to realize that it holds for the regularized one $\mathcal{L}_\omega$ by simply replacing $\mathcal{L}_0$ with $\mathcal{L}_\omega$ in Equation (2).

Remark 3.1 (When does Assumption 3.2 holds?)**.** *As remarked by Ding et al. (2024), the Lagrangian function, for a fixed value of $\boldsymbol{\lambda}$ can be regarded as the return of a new reward function $-C_0 - \langle\boldsymbol{\lambda},\mathbf{C}\rangle$, where $\mathbf{C} = (C_1,\dots,C_U)^\top$.*

[3] Assumption 3.1 combined with Slater's condition, i.e., the existence of a parametrization $\widetilde{\boldsymbol{v}} \in \mathcal{V}$ for which there exists $\xi > 0$ such that $J_i(\widetilde{\boldsymbol{v}}) - b < -\xi$ for all $i \in [\![U]\!]$ (strictly feasible), allows providing an upper bound to the Lagrange multipliers $\|\boldsymbol{\lambda}_0^*\|_2 \leq \xi^{-1}(J_0(\widetilde{\boldsymbol{v}}) - J_0(\boldsymbol{v}_0^*))$ using standard arguments (see Ying et al., 2022).

As a consequence, a sufficient condition for Assumption 3.2 is when the selected class of policies guarantees the ψ-gradient domination regardless *of the reward function. For instance, in tabular environments with natural policy parametrization, i.e., $\pi_{\boldsymbol{\theta}}(s) = \boldsymbol{\theta}_s$ for every $s \in \mathcal{S}$, the PL condition ($\psi = 2$ and $\beta_1 = 0$) holds (Bhandari and Russo, 2024). Moreover, in tabular environments with softmax policy, i.e., $\pi_{\boldsymbol{\theta}}(a|s) \propto \exp(\theta(s,a))$, GD ($\psi = 1$ and $\beta_1 = 0$) holds (Mei, Xiao, Szepesvari and Schuurmans, 2020). This enables a meaningful comparison of our results when resorting to softmax policies (e.g., Ding et al., 2020; Gladin et al., 2023; Ding et al., 2024). More in general, when (i) the Fisher information matrix induced by policy $\pi_{\boldsymbol{\theta}}$ is non-degenerate for every $\boldsymbol{\theta} \in \Theta$, i.e., $\mathbf{F}(\boldsymbol{\theta}) = \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s}) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})^\top] \geq \mu_F \mathbf{I}$ for some $\mu_F > 0$ and (ii) a compatible function approximation bias bound holds, i.e., $\mathbb{E}_{\pi_{\boldsymbol{\theta}^*}}[(A^{\pi_{\boldsymbol{\theta}}}(\boldsymbol{s},\boldsymbol{a}) - (1-\gamma)\boldsymbol{u}^\top \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s}))^2] \leq \epsilon_{bias}$ being $\boldsymbol{u} = \mathbf{F}(\boldsymbol{\theta})^\dagger \nabla_{\boldsymbol{\theta}} J_0(\boldsymbol{\theta})$ and the advantage function $A^{\pi_{\boldsymbol{\theta}}}$ computed w.r.t. reward $-c_0 - \langle \boldsymbol{\lambda}, \mathbf{c}\rangle$, the weak GD ($\psi = 1$) holds with $\alpha_1 = G\mu_F^{-1}$ and $\beta_1 = (1-\gamma)^{-1}\sqrt{\epsilon_{bias}}$, where G is such that $\|\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})\|_2 \leq G$ (Masiha et al., 2022).*

In principle, we could have enforced Assumption 3.2 on the primal function $H_\omega(\boldsymbol{v})$ only. However, this would come with two drawbacks: (i) the assumption would now depend explicitly on ω; (ii) the considerations of Remark 3.1 would no longer hold. Nevertheless, in Lemma A.4, we prove that Assumption 3.2 induces an analogous property on the primal function $H_\omega(\boldsymbol{v})$ in the regularized case.

Assumption 3.3 (Regularity of the Regularized Lagrangian $\mathcal{L}_0$)**.** *There exist $L_1, L_2, L_3 \in \mathbb{R}_{>0}$ such that, for every $\boldsymbol{v}, \boldsymbol{v}' \in \mathcal{V}$, and for every $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$, the following holds:*

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}_0(\cdot, \boldsymbol{\lambda}) \ L_1\text{-Lipschitz w.r.t. } \boldsymbol{v}: \qquad \left\|\nabla_{\boldsymbol{\lambda}} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \nabla_{\boldsymbol{\lambda}} \mathcal{L}_0(\boldsymbol{v}', \boldsymbol{\lambda})\right\|_2 \leq L_1 \left\|\boldsymbol{v} - \boldsymbol{v}'\right\|_2, \tag{3}$$

$$\mathcal{L}_0(\cdot, \boldsymbol{\lambda}) \ L_2\text{-Smooth w.r.t. } \boldsymbol{v}: \qquad \left\|\nabla_{\boldsymbol{v}} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \nabla_{\boldsymbol{v}} \mathcal{L}_0(\boldsymbol{v}', \boldsymbol{\lambda})\right\|_2 \leq L_2 \|\boldsymbol{v} - \boldsymbol{v}'\|_2, \tag{4}$$

$$\nabla_{\boldsymbol{v}} \mathcal{L}_0(\boldsymbol{v}, \cdot) \ L_3\text{-Lipschitz w.r.t } \boldsymbol{\lambda}: \qquad \left\|\nabla_{\boldsymbol{v}} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \nabla_{\boldsymbol{v}} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}')\right\|_2 \leq L_3 \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2. \tag{5}$$

Notice that, similarly to Assumption 3.2, we realize that if Assumption 3.3 holds for the non-regularized Lagrangian $\mathcal{L}_0$, it also holds (with the same constants) for the regularized one $\mathcal{L}_\omega$ for every $\omega > 0$. The regularity conditions of Assumption 3.3 are common in the literature (Yang et al., 2020) and mild when regarded from the policy optimization perspective. Equation (3) is satisfied whenever the constraint functions J_i are Lipschitz continuous w.r.t. $\boldsymbol{v}$. Indeed, $\left\|\nabla_{\boldsymbol{\lambda}} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \nabla_{\boldsymbol{\lambda}} \mathcal{L}_0(\boldsymbol{v}', \boldsymbol{\lambda})\right\|_2 = \left\|\mathbf{J}(\boldsymbol{v}) - \mathbf{J}(\boldsymbol{v}')\right\|_2$. Equation (4) is fulfilled when the objective function J_0 and the constraint functions J_i are smooth w.r.t. $\boldsymbol{v}$ and the Lagrange multipliers are bounded (guaranteed thanks to the projection Π_Λ), since $\left\|\nabla_{\boldsymbol{v}} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \nabla_{\boldsymbol{v}} \mathcal{L}_0(\boldsymbol{v}', \boldsymbol{\lambda})\right\|_2 \leq |\nabla_{\boldsymbol{v}} J_0(\boldsymbol{v}) - \nabla_{\boldsymbol{v}} J_0(\boldsymbol{v}')| + \sum_{i=1}^U \lambda_i |\nabla_{\boldsymbol{v}} J_i(\boldsymbol{v}) - \nabla_{\boldsymbol{v}} J_i(\boldsymbol{v})|$. Finally, Equation (5) is fulfilled whenever functions J_i admit bounded gradients, since $\left\|\nabla_{\boldsymbol{v}} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \nabla_{\boldsymbol{v}} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}')\right\|_2 \leq \|\nabla_{\boldsymbol{v}} \mathbf{J}(\boldsymbol{v})(\boldsymbol{\lambda} - \boldsymbol{\lambda}')\|_2$. It is worth noting that L_2 depends on the norm of the Lagrange multipliers and, consequently, due to the projection operator Π_Λ, we have that $L_2 = \mathcal{O}(\omega^{-1})$, whereas L_1 and L_3 are independent on ω.[4] Explicit conditions on the constitutive elements of the MDP and (hyper)policies to ensure Lipshitzness and smoothness of these quantities are reported in (Montenegro et al., 2024a, Appendix E) for both the AB and PB cases. These regularity properties enforced on $\mathcal{L}_\omega$ are inherited by the primal function H_ω which results to be $\left(L_2 + L_1^2\omega^{-1}\right)$-LS (Lemma A.7). Concerning the regularity of $\mathcal{L}_\omega$ w.r.t. $\boldsymbol{\lambda}$, we observe that it is a quadratic function and, therefore, it is ω-smooth and satisfies the PL condition, i.e., Assumption 3.2 with $\psi = 2$, $\beta_1 = 0$, and with $\alpha_1 = \omega$ (Lemma A.5).

Assumption 3.4 (Bounded Estimator Variance)**.** *For every $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{\lambda} \in \Lambda$, the estimators $\widehat{\nabla}_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})$ and $\widehat{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})$ are unbiased for $\nabla_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}) = \nabla_{\boldsymbol{v}} J_0(\boldsymbol{v}) + \sum_{i=1}^U \lambda_i \nabla_{\boldsymbol{v}} J_i(\boldsymbol{v})$ and $\nabla_{\boldsymbol{\lambda}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}) = \mathbf{J}(\boldsymbol{v}) - \mathbf{b} - \omega\boldsymbol{\lambda}$ with bounded variance, i.e., there exist $V_{\boldsymbol{v}}, V_{\boldsymbol{\lambda}} \in \mathbb{R}_{\geq 0}$ such that:*

$$\mathbb{V}\mathrm{ar}[\widehat{\nabla}_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})] \leq V_{\boldsymbol{v}}, \qquad \mathbb{V}\mathrm{ar}[\widehat{\nabla}_{\boldsymbol{\lambda}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})] \leq V_{\boldsymbol{\lambda}}.$$

Note that $V_{\boldsymbol{v}}$ typically depends on the Lagrange multipliers and, for standard sample-mean estimators, it is of order $V_{\boldsymbol{v}} = \mathcal{O}(\omega^{-2})$ thanks to the projection operator. In contrast, $V_{\boldsymbol{\lambda}}$ is usually not affected by ω since the term $\omega\boldsymbol{\lambda}$ is not estimated and, thus, it does not affect the variance of the sample mean estimator. The variance of such estimators can be easily controlled by leveraging on the properties of the score function as done in previous works (see Papini, Pirotta and Restelli 2022 and Montenegro et al. 2024a, Appendix E).

[4] We highlight the dependencies on ω since, as we shall see later, we will set $\omega = \mathcal{O}(\epsilon)$ having, consequently, an effect on the convergence rate.

3.3. Convergence Analysis

We are now ready to tackle the convergence analysis of C-PG to the global optimum of the COP of Equation (1). To this end, we study the *potential function* defined as $\mathcal{P}_k(\chi) := a_k + \chi b_k$, where $a_k := \mathbb{E}[H_\omega(\boldsymbol{v}_k) - H^*_\omega]$ and $b_k := \mathbb{E}[H_\omega(\boldsymbol{v}_k) - \mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k)]$, $\chi \in (0,1)$ will be specified later, and the expectation is taken w.r.t. the stochastic process generating samples. Since $a_k, b_k \geq 0$, intuitively, if $\mathcal{P}_k(\chi) \approx 0$ we have that both $a_k, b_k \approx 0$ and, consequently, convergence is achieved. Let us start relating $\mathcal{P}_k(\chi)$, with the solution of the COP in Equation (1).

Theorem 3.1 (Objective Function Gap and Constraint Violation)**.** *Let $\epsilon \in \mathbb{R}_{>0}$. Under Assumption 3.1, if $\mathcal{P}_k(\chi) \leq \epsilon$, it holds that:*

$$\mathbb{E}[J_0(\boldsymbol{v}_k) - J_0(\boldsymbol{v}^*_0)] \leq \epsilon + \frac{\omega}{2}\|\boldsymbol{\lambda}^*_0\|_2^2, \qquad \mathbb{E}[(J_i(\boldsymbol{v}_k) - b_i)^+] \leq 4\epsilon + \omega\|\boldsymbol{\lambda}^*_0\|_2, \quad \forall i \in [\![U]\!]. \tag{6}$$

Proof. Since $\mathcal{P}_k(\chi) \leq \epsilon$, it follows that $a_k \leq \epsilon$ and, consequently, $0 \leq \mathbb{E}[H_\omega(\boldsymbol{v}_k) - H^*_\omega] \leq \epsilon$. We start by bounding the norm of the dual variables:

$$\|\boldsymbol{\lambda}^*(\boldsymbol{v}_k)\|_2 \leq \|\boldsymbol{\lambda}^*_\omega\|_2 + \|\boldsymbol{\lambda}^*(\boldsymbol{v}_k) - \boldsymbol{\lambda}^*_\omega\|_2 \leq \|\boldsymbol{\lambda}^*_\omega\|_2 + \frac{4}{\omega}(H_\omega(\boldsymbol{v}_k) - H^*_\omega),$$

where we applied the triangular inequality and Lemma A.6, which proves that, for any $\boldsymbol{v} \in \mathcal{V}$, $H_\omega(\boldsymbol{v}) - H^*_\omega \geq \frac{\omega}{4}\|\boldsymbol{\lambda}^*(\boldsymbol{v}) - \boldsymbol{\lambda}^*_\omega\|_2$. The projection Π_Λ is such that $\boldsymbol{\lambda}^*(\boldsymbol{v}) = \Pi_\Lambda\left(\frac{1}{\omega}(\mathbf{J}(\boldsymbol{v}) - \mathbf{b})\right) = \frac{1}{\omega}(\mathbf{J}(\boldsymbol{v}) - \mathbf{b})^+$ and, consequently, we have:

$$\|(\mathbf{J}(\boldsymbol{v}_k) - \mathbf{b})^+\|_2 - \|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2 \leq 4(H_\omega(\boldsymbol{v}_k) - H^*_\omega).$$

By the last inequality, together with Lemma A.3, which states that:

$$0 \leq J_0(\boldsymbol{v}^*_0) - J_0(\boldsymbol{v}^*_\omega) \leq \omega\|\boldsymbol{\lambda}^*_0\|_2^2 \quad \text{and} \quad \|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2 \leq \omega\|\boldsymbol{\lambda}^*_0\|_2,$$

and applying the expectation on both sides, we have the following:

$$\mathbb{E}[\|(\mathbf{J}(\boldsymbol{v}_k) - \mathbf{b})^+\|_2] \leq \|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2 + 4\,\mathbb{E}[H_\omega(\boldsymbol{v}_k) - H^*_\omega] \leq \omega\|\boldsymbol{\lambda}^*_0\|_2 + 4\epsilon.$$

We obtain the constraint violation bound recalling that:

$$\mathbb{E}[\|(\mathbf{J}(\boldsymbol{v}_k) - \mathbf{b}))^+\|_2] \geq \left\|\mathbb{E}[(\mathbf{J}(\boldsymbol{v}_k) - \mathbf{b})^+]\right\|_2 \geq \|\,\mathbb{E}[(\mathbf{J}(\boldsymbol{v}_k) - \mathbf{b})^+]\|_\infty.$$

For the objective function bound, let us consider the following derivation. By definition of $H_\omega(\boldsymbol{v})$ and $\boldsymbol{\lambda}^*(\boldsymbol{v})$ we have:

$$J_0(\boldsymbol{v}_k) - J_0(\boldsymbol{v}^*_\omega) = H_\omega(\boldsymbol{v}_k) - H^*_\omega - \frac{\omega}{2}\left(\|\boldsymbol{\lambda}^*(\boldsymbol{v}_k)\|_2^2 - \|\boldsymbol{\lambda}^*_\omega\|_2^2\right).$$

Taking the expectation on both sides and upper bounding $\|\boldsymbol{\lambda}^*_\omega\|$ with $\|\boldsymbol{\lambda}^*_0\|$ from Lemma A.1, which states that $0 \leq \mathcal{L}_0(\boldsymbol{v}^*_0, \boldsymbol{\lambda}^*_0) - \mathcal{L}_0(\boldsymbol{v}^*_\omega, \boldsymbol{\lambda}^*_\omega) \leq \frac{\omega}{2}\left(\|\boldsymbol{\lambda}^*_0\|_2^2 - \|\boldsymbol{\lambda}^*_\omega\|_2^2\right)$, the following holds:

$$\begin{aligned}
\mathbb{E}[J_0(\boldsymbol{v}_k) - J_0(\boldsymbol{v}^*_\omega)] &= \mathbb{E}[H_\omega(\boldsymbol{v}_k) - H^*_\omega] - \frac{\omega}{2}\,\mathbb{E}[\|\boldsymbol{\lambda}^*(\boldsymbol{v}_k)\|_2^2 - \|\boldsymbol{\lambda}^*_\omega\|_2^2] \\
&\leq \mathbb{E}[H_\omega(\boldsymbol{v}_k) - H^*_\omega] + \frac{\omega}{2}\|\boldsymbol{\lambda}^*_\omega\|_2^2 \\
&\leq \epsilon + \frac{\omega}{2}\|\boldsymbol{\lambda}^*_0\|_2^2.
\end{aligned}$$

The result is obtained by applying Lemma A.3 (already stated in this proof) as follows:

$$\mathbb{E}[J_0(\boldsymbol{v}_k) - J_0(\boldsymbol{v}^*_0)] = \mathbb{E}[J_0(\boldsymbol{v}_k) - J_0(\boldsymbol{v}^*_\omega)] + \underbrace{J_0(\boldsymbol{v}^*_\omega) - J_0(\boldsymbol{v}^*_0)}_{\leq 0}.$$

□

Theorem 3.1 justifies the study of the potential $\mathcal{P}_k(\chi)$ as a technical tool to ensure convergence. Indeed, whenever $\mathcal{P}_k(\chi) \le \epsilon$ both (i) the objective function gap and (ii) the constraint violation scale linearly with ϵ and with the regularization parameter ω of the regularized Lagrangian $\mathcal{L}_\omega$ multiplied by the norm of the Lagrange multipliers of the non-regularized problem $\|\lambda_0^*\|_2$, which are finite under Assumption 3.1. This expression also suggests a choice of $\omega = \mathcal{O}(\epsilon)$ to enforce an overall ϵ error on both quantities. Note that, from Theorem 3.1, it is immediate to employ a *conservative constraint* ($b_i' \approx b_i - 4\epsilon - \omega\|\lambda_0^*\|_2$) to achieve zero constraint violation with no modification of the algorithm.

We are now ready to state the convergence guarantees for the potential function.

Theorem 3.2 (Convergence of $\mathcal{P}_K$)**.** *Under Assumptions 3.2, 3.3, 3.4, for $\chi < 1/5$, sufficiently small ϵ and ω, and a choice of* constant *learning rates $\zeta_{\upsilon}, \zeta_\lambda$, we have $\mathcal{P}_K(\chi) \le \epsilon + \beta_1/\alpha_1$ whenever:*[5]

- $K = \mathcal{O}(\omega^{-1}\log(\epsilon^{-1}))$ *if $\psi = 2$ and the gradients are exact (i.e., $V_{\upsilon} = V_\lambda = 0$);*
- $K = \mathcal{O}(\omega^{-1}\epsilon^{-\frac{2}{\psi}-1})$ *if $\psi \in [1,2)$ and the gradients are exact (i.e., $V_{\upsilon} = V_\lambda = 0$);*
- $K = \mathcal{O}(\omega^{-3}\epsilon^{-\frac{4}{\psi}+1})$ *if $\psi \in [1,2]$ and the gradients are estimated (i.e., $V_{\upsilon} = \mathcal{O}(\omega^{-2})$ and $V_\lambda = \mathcal{O}(1)$).*

Proof Sketch. The proof of Theorem 3.2 is quite technical, thus we report here just its sketch, which we divide into five parts.

Part I: bounding a_k. The *first part* of the proof consists of bounding $\mathbb{E}[a_{k+1} \mid \mathcal{F}_{k-1}] = \mathbb{E}[H_\omega(\upsilon_{k+1}) - H^* \mid \mathcal{F}_{k-1}]$, considering to be at a generic k^{th} iterate of C-PG with $\mathcal{F}_{k-1}$ a filtration up to iteration $k-1$. In particular, by exploiting the update rule of C-PG, via Lemma A.7 stating that H_ω is L_H-LS, and by selecting $\zeta_{\upsilon,k} \le L_H$, we can conclude that:

$$
\begin{aligned}
&\mathbb{E}\left[H_\omega(\upsilon_{k+1}) | \mathcal{F}_{k-1}\right] - H^* \\
&\le H_\omega(\upsilon_k) - H^* - \frac{\zeta_{\upsilon,k}}{2}\left\|\nabla_{\upsilon} H_\omega(\upsilon_k)\right\|_2^2 + \frac{\zeta_{\upsilon,k}}{2}\left\|\nabla_{\upsilon}\mathcal{L}_\omega(\upsilon_k,\lambda_k) - \nabla_{\upsilon} H_\omega(\upsilon_k)\right\|_2^2 + \frac{L_H}{2}\zeta_{\upsilon,k}^2 V_{\upsilon},
\end{aligned}
$$

where the constant V_{υ}, coming from Assumption 3.4, is such that $\mathbb{V}\mathrm{ar}[\nabla_{\upsilon}\mathcal{L}_\omega(\upsilon_k,\lambda_k)] \le V_{\upsilon}$.

Part II: bounding b_k. Similarly to what shown in the first part, the *second part* consists of bounding $\mathbb{E}[b_{k+1} \mid \mathcal{F}_{k-1}] = \mathbb{E}[H_\omega(\upsilon_{k-1}) - \mathcal{L}_\omega(\upsilon_{k+1},\lambda_{k+1}) \mid \mathcal{F}_{k-1}]$. To do so, we exploit Assumption 3.3 stating that $\mathcal{L}_\omega$ is L_2-LS. In particular, in Lemma A.5, we show that $\mathcal{L}_\omega$ is ω-LS and that it fulfills the PL condition with constant ω. From these observations, together with the update rule of C-PG and the selection $\zeta_{\lambda,k} \le 1/\omega$, we conclude that:

$$
\begin{aligned}
&\mathbb{E}\left[H_\omega(\upsilon_{k+1}) - \mathcal{L}_\omega(\upsilon_{k+1},\lambda_{k+1}) | \mathcal{F}_{k-1}\right] \\
&\le \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\left(H_\omega(\upsilon_k) - \mathcal{L}_\omega(\upsilon_k,\lambda_k)\right) + \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\left(\zeta_{\upsilon,k}\left(1 + \frac{L_2}{2}\zeta_{\upsilon,k}\right)\left\|\nabla_{\upsilon}\mathcal{L}_\omega(\upsilon_k,\lambda_k)\right\|_2^2 + \frac{L_2}{2}\zeta_{\upsilon,k}^2 V_{\upsilon}\right) \\
&\quad + \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\left(-\frac{\zeta_{\upsilon,k}}{2}\left\|\nabla_{\upsilon} H_\omega(\upsilon_k)\right\|_2^2 + \frac{\zeta_{\upsilon,k}}{2}\left\|\nabla_{\upsilon}\mathcal{L}_\omega(\upsilon_k,\lambda_k) - \nabla_{\upsilon} H_\omega(\upsilon_k)\right\|_2^2 + \frac{L_H}{2}\zeta_{\upsilon,k}^2 V_{\upsilon}\right) + \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda.
\end{aligned}
$$

Part III: bounding $\mathcal{P}_k(\chi)$. Having bounded separately a_k and b_k, and being $\mathcal{P}_k(\chi) = a_k + \chi b_k$, we can just put together the previously obtained results to have a bound on $\mathcal{P}_k(\chi)$. Moreover, exploiting Assumption 3.3 and by noticing that $\mathcal{L}_\omega$ satisfies the quadratic growth condition (since Lemma A.5 states that $\mathcal{L}_\omega$ satisfies the PL condition with ω as constant), we obtain the following inequality:

$$
\begin{aligned}
&a_{k+1} + \chi b_{k+1} \\
&\le a_k + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) b_k \\
&\quad + \left(2\zeta_{\upsilon,k}\left(1 + \frac{L_2}{2}\zeta_{\upsilon,k}\right)\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right.
\end{aligned}
$$

[5] In the context of this statement, the $\mathcal{O}(\cdot)$ notation preserves dependences on ϵ and ω only.

$$-\frac{\zeta_{\boldsymbol{\upsilon},k}}{2}\left(1+\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\Bigg)\mathbb{E}\left[\|\nabla_{\boldsymbol{\upsilon}}H_\omega(\boldsymbol{\upsilon}_k)\|_2^2\right]$$
$$+\left(2\zeta_{\boldsymbol{\upsilon},k}\left(1+\frac{L_2}{2}\zeta_{\boldsymbol{\upsilon},k}\right)\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)+\frac{\zeta_{\boldsymbol{\upsilon},k}}{2}\left(1+\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right)\frac{4L_3^2}{\omega}b_k$$
$$+\frac{\zeta_{\boldsymbol{\upsilon},k}^2}{2}\left(L_H+\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)(L_H+L_2)\right)V_{\boldsymbol{\upsilon}}+\chi\frac{\omega}{2}\zeta_{\lambda,k}^2V_\lambda.$$

Part IV: applying the ψ-gradient domination. Form the previously highlighted inequality, we aim at recovering a recursive equation in $\mathcal{P}_k(\chi)$. To this end, we apply Assumption 3.2, from which it follows that

$$\mathbb{E}\left[\|\nabla_{\boldsymbol{\upsilon}}H_\omega(\boldsymbol{\upsilon})\|_2^2\right]\geq\alpha_1^{\frac{2}{\psi}}\max\left\{0,\ \mathbb{E}\left[H_\omega(\boldsymbol{\upsilon})-\widetilde{H}^*\right]\right\}^{\frac{2}{\psi}},$$

where $\widetilde{H}^* := H^*+\beta_1/\alpha_1$. Now, exploiting this last result and enforcing $\chi\leq\min\{1/5,\ 1/(\max_{k\in[\![K]\!]}b_k)\}$, after many algebraic steps, we obtain the following inequality:

$$\widetilde{P}_{k+1}(\chi)\leq\widetilde{P}_k(\chi)-\widetilde{C}\max\left\{0,\ \widetilde{P}_k(\chi)\right\}^{\frac{2}{\psi}}+\widetilde{V},$$

where $\widetilde{P}_k(\chi):=a_k+\chi b_k-\beta_1/\alpha_1$, $\widetilde{C}:=2^{1-\frac{1}{\psi}}\frac{\zeta_{\boldsymbol{\upsilon},k}\alpha_1^{\frac{2}{\psi}}}{2}$, and $\widetilde{V}:=\frac{\zeta_{\boldsymbol{\upsilon},k}^2}{2}\left((1+2\chi)L_2+(1+\chi)\frac{L_1^2}{\omega}\right)V_{\boldsymbol{\upsilon}}+\chi\frac{\omega}{2}\zeta_{\lambda,k}^2V_\lambda$. We highlight that to get to this result, the learning rates have been selected as:

$$\zeta_{\boldsymbol{\upsilon},k}\leq\min\left\{\frac{1}{L_H},\frac{1}{L_2},\frac{\omega^2\chi\zeta_{\lambda,k}}{(1+\chi)\omega\alpha_1^{\frac{2}{\psi}}+4L_3^2(1+7\chi)}\right\}\quad\text{and}\quad\zeta_{\lambda,k}\leq\frac{1}{\omega}.$$

Part V: rates computation. Equipped with the recursive inequality reported in Part IV, we just have to compute the rates guaranteeing $\mathcal{P}_K(\chi)\leq\epsilon+\beta_1/\alpha_1$. In particular, we first analyze the *exact gradient* case, i.e., $\widetilde{V}=0$, for when $\psi=2$, and $\psi\in[1,2)$. Then, we do the same in the case of *estimated gradients*, i.e., $\widetilde{V}>0$. All the results are reported in Table 1. □

Some comments are in order. First, Theorem 3.2 holds for a specific choice of the constant $\chi\in(0,1/5)$ defining the potential function $\mathcal{P}_K(\chi)$. Second, the presented rates hold for sufficiently small values of ϵ and ω. This is just for presentation purposes, as the sample complexity[6] can only improve if we increase the values of ϵ and ω. Third, in the proof, an explicit expression of the learning rates is provided. Concerning their orders, for the case of exact gradients, we choose $\zeta_\lambda=\omega^{-1}$ and $\zeta_{\boldsymbol{\upsilon}}=\mathcal{O}(\omega)$, whereas for the estimated gradient case, we choose $\zeta_\lambda=\mathcal{O}(\omega\epsilon^{2/\psi})$ and $\zeta_{\boldsymbol{\upsilon}}=\mathcal{O}(\omega^3\epsilon^{2/\psi})$.

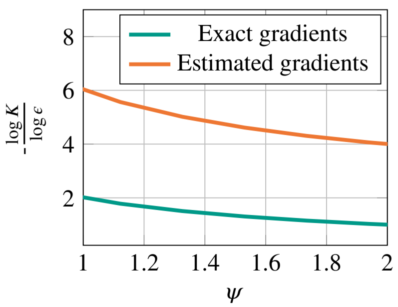


Figure 1: Plot of the exponents of ϵ^{-1} in the cases of Table 1.

Assuming ω to be a constant, we observe that both learning rates display the same dependence on ϵ and, consequently, they are in *single-time scale*. However, as we have seen in Theorem 3.1, in order to obtain guarantees on the original non-regularized problem, we have to set $\omega=\mathcal{O}(\epsilon)$, leading to a *two-time scales* algorithm. Fourth, we observe that, for both exact and estimated gradients, the sample complexity degrades as the constant ψ of the gradient domination moves from 2 to 1, delivering the smallest sample complexity when the PL condition holds. Finally, we highlight that C-PG jointly: (*i*) converges to the global optimum of the COP problem of Equation (1); (*ii*) delivers a *last-iterate* guarantee; (*iii*) has no dependence on the cardinality of the state or action spaces, making it completely *dimension-free*. Table 1 and Figure 1 summarize the results of Theorem 3.2.

[6]Theorem 3.2 provides an *iteration-complexity* guarantee. Concerning the *estimated gradient* case, this translates into a *sample complexity* guarantee since we are allowed to estimate gradients with a single sample.

	Exact Gradients			Estimated Gradients		
	ψ=1 (GD)	$\psi\in(1,2)$	ψ=2 (PL)	ψ=1 (GD)	$\psi\in(1,2)$	ψ=2 (PL)
Fixed ω	$\omega^{-1}\epsilon^{-1}$	$\omega^{-1}\epsilon^{-\frac{2}{\psi}+1}$	$\omega^{-1}\log(\epsilon^{-1})$	$\omega^{-3}\epsilon^{-3}\log(\epsilon^{-1})$	$\omega^{-3}\epsilon^{-\frac{4}{\psi}+1}\log(\epsilon^{-1})$	$\omega^{-3}\epsilon^{-1}\log(\epsilon^{-1})$
$\omega=\mathcal{O}(\epsilon)$	ϵ^{-2}	$\epsilon^{-\frac{2}{\psi}}$	$\epsilon^{-1}\log(\epsilon^{-1})$	$\epsilon^{-6}\log(\epsilon^{-1})$	$\epsilon^{-\frac{4}{\psi}-2}\log(\epsilon^{-1})$	$\epsilon^{-4}\log(\epsilon^{-1})$

Table 1
Summary of the sample complexity results of `C-PG` when either keeping ω fixed or setting it as $\omega = \mathcal{O}(\epsilon)$.

3.4. Action-based and Parameter-based Variants of `C-PG`

So far, we have focused on the exploration-agnostic formulation of the proposed method. In this section, we introduce its action-based and parameter-based variants, namely `C-PGAE` and `C-PGPE`. These variants differ in the considered cost functions, $J_{\mathrm{A},i}$ or $J_{\mathrm{P},i}$ for every $i \in [\![0, U]\!]$, and the estimators employed to update the optimization variables. We recall that the general Lagrangian function for the problem in Equation (1) is the following:

$$\mathcal{L}_{\dagger,\omega}(\boldsymbol{v}, \boldsymbol{\lambda}) := J_{\dagger,0}(\boldsymbol{v}) + \sum_{u=1}^{U} \lambda_u \left(J_{\dagger,u}(\boldsymbol{v}) - b_u\right) - \frac{\omega}{2} \|\boldsymbol{\lambda}\|_2^2,$$

being $\boldsymbol{v} \in \mathcal{V}$ a generic parameter vector to be optimized. As highlighted in Section 2, in the case of action-based exploration ($\dagger = \mathrm{A}$) $\boldsymbol{v}$ corresponds to the policy parameterization $\boldsymbol{v} = \boldsymbol{\theta} \in \Theta$, while in the case of parameter-based exploration ($\dagger = \mathrm{P}$) it coincides with the hyperpolicy parameterization $\boldsymbol{v} = \boldsymbol{\rho} \in \mathcal{R}$. In the following, we are going to consider the gradients w.r.t. both the parameterization $\boldsymbol{v}$ and the Lagrange multipliers $\boldsymbol{\lambda}$, having the following explicit forms in the exploration-agnostic setting:

$$\nabla_{\boldsymbol{v}}\mathcal{L}_{\dagger,\omega}(\boldsymbol{v}, \boldsymbol{\lambda}) = \nabla_{\boldsymbol{v}} J_{\dagger,0}(\boldsymbol{v}) + \sum_{u=1}^{U} \lambda_u \nabla_{\boldsymbol{v}} J_{\dagger,u}(\boldsymbol{v}) \quad \text{and} \quad \nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\dagger,\omega}(\boldsymbol{v}, \boldsymbol{\lambda}) = \mathbf{J}_{\dagger}(\boldsymbol{v}) + \mathbf{b} + \omega\boldsymbol{\lambda},$$

where $\mathbf{J}_{\dagger}(\boldsymbol{v}) := (J_{\dagger,1}(\boldsymbol{v}), \dots, J_{\dagger,U}(\boldsymbol{v}))^\top$ and $\mathbf{b} := (b_1, \dots, b_U)^\top$.

Action-based Exploration for `C-PG`. The action-based variant of `C-PG`, referred to as `C-PGAE`, aims at optimizing the parameters $\boldsymbol{\theta}$ of a parametric stochastic policy $\pi_{\boldsymbol{\theta}}$. In particular, for every $i \in [\![0, U]\!]$, we recall the definition of action-based cost functions: $J_{\mathrm{A},i}(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})}\left[C_i(\tau)\right]$, where $p_{\mathrm{A}}(\tau, \boldsymbol{\theta})$ is the probability density of trajectory τ induced by the policy $\pi_{\boldsymbol{\theta}}$.

Considering the gradient w.r.t. the parameters $\boldsymbol{\theta}$, the following holds:

$$\nabla_{\boldsymbol{\theta}} J_{\mathrm{A},i}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathop{\mathbb{E}}_{\tau \sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})}\left[C_i(\tau)\right] = \mathop{\mathbb{E}}_{\tau \sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})}\left[\nabla_{\boldsymbol{\theta}} \log p_{\mathrm{A}}(\tau, \boldsymbol{\theta}) C_i(\tau)\right].$$

As for standard PGs (Williams, 1992; Baxter and Bartlett, 2001), we can switch to its sample-based version to obtain an unbiased estimator of the gradient. In particular, we resort to a GPOMDP-like version (Baxter and Bartlett, 2001) for the proposed estimator:

$$\widehat{\nabla}_{\boldsymbol{\theta}} J_{\mathrm{A},i}(\boldsymbol{\theta}) := \frac{1}{N} \sum_{j=1}^{N} \sum_{t=0}^{T-1} \left(\sum_{l=0}^{t} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}_{\tau_j,l}, \boldsymbol{s}_{\tau_j,l}) \right) \gamma^t c_i(\boldsymbol{s}_{\tau_j,t}, \boldsymbol{a}_{\tau_j,t}),$$

where N, called batch size, is the number of independent trajectories $\{\tau_j\}_{j=1}^{N}$ such that $\tau_j \sim p_{\mathrm{A}}(\cdot \mid \boldsymbol{\theta})$. We just consider a GPOMDP-like version of the estimator, since the REINFORCE-like one would suffer from a higher variance as for standard PGs (Papini et al., 2022). Thus, considering that:

$$\mathcal{L}_{\mathrm{A},\omega}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J_{\mathrm{A},0}(\boldsymbol{\theta}) + \sum_{u=1}^{U} \lambda_u \left(J_{\mathrm{A},i}(\boldsymbol{\theta}) - b_u\right) - \frac{\omega}{2} \|\boldsymbol{\lambda}\|_2^2,$$

C-PGAE employs the following estimator to update the primal variable $\boldsymbol{\theta}$:

$$\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{A},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \widehat{\nabla}_{\boldsymbol{\theta}}J_{\mathrm{A},0}(\boldsymbol{\theta}) + \sum_{u=1}^{U}\lambda_u\widehat{\nabla}_{\boldsymbol{\theta}}J_{\mathrm{A},i}(\boldsymbol{\theta}).$$

If we now focus on the action-based gradient w.r.t. Lagrange multipliers, we have the following:

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mathrm{A},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \mathbf{J}_{\mathrm{A}}(\boldsymbol{\theta}) - \mathbf{b} - \omega\boldsymbol{\lambda},$$

where $\mathbf{J}_{\mathrm{A}}(\boldsymbol{\theta}) := (J_{\mathrm{A},1}(\boldsymbol{\theta}),\dots,J_{\mathrm{A},U}(\boldsymbol{\theta}))^{\top}$. Thus, its sample-based version used by C-PGAE to update the dual variable is the following:

$$\frac{\widehat{\partial}}{\partial\lambda_i}\mathcal{L}_{\mathrm{A},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \frac{1}{N}\sum_{j=1}^{N}C_i(\tau_j) - b_i - \omega\lambda_i,$$

where the N independent trajectories $\{\tau_j\}_{j=1}^{N}$ are such that $\tau_j \sim p_{\mathrm{A}}(\cdot \mid \boldsymbol{\theta})$.

Parameter-based Exploration for C-PG. The parameter-based variant of C-PG, referred to as C-PGPE, aims at optimizing the parameters $\boldsymbol{\rho}$ of a parametric stochastic hyperpolicy $\nu_{\boldsymbol{\rho}}$, used to sample the parameters $\boldsymbol{\theta}$ for an underlying parametric policy $\pi_{\boldsymbol{\theta}}$ (which can also be deterministic, as we shall see in the next section). In particular, for every $i \in [\![0,U]\!]$, we recall the definition of parameter-based cost functions: $J_{\mathrm{P},i}(\boldsymbol{\rho}) = \mathbb{E}_{\boldsymbol{\theta}\sim\nu_{\boldsymbol{\rho}}}\left[\mathbb{E}_{\tau\sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})}\left[C_i(\tau)\right]\right]$, where $p_{\mathrm{A}}(\tau,\boldsymbol{\theta})$ is the probability density of trajectory τ induced by the policy $\pi_{\boldsymbol{\theta}}$.

Considering the gradient w.r.t. the parameters $\boldsymbol{\theta}$, the following holds:

$$\nabla_{\boldsymbol{\rho}}J_{\mathrm{P},i}(\boldsymbol{\rho}) = \nabla_{\boldsymbol{\rho}}\mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim\nu_{\boldsymbol{\rho}}}\left[\mathop{\mathbb{E}}_{\tau\sim p_{\mathrm{A}}(\cdot|\boldsymbol{\theta})}\left[C_i(\tau)\right]\right] = \mathop{\mathbb{E}}_{\boldsymbol{\theta}\sim\nu_{\boldsymbol{\rho}}}\left[\nabla_{\boldsymbol{\rho}}\log\nu_{\boldsymbol{\rho}}(\boldsymbol{\theta})\mathop{\mathbb{E}}_{\tau\sim p_{\mathrm{A}}(\cdot,\boldsymbol{\theta})}\left[\nabla_{\boldsymbol{\theta}}\log p_{\mathrm{A}}(\tau|\boldsymbol{\theta})C_i(\tau)\right]\right].$$

As for the prototypical parameter-based method PGPE (Sehnke et al., 2010), we switch to its sample-based version to obtain an unbiased estimator of the gradient:

$$\widehat{\nabla}_{\boldsymbol{\rho}}J_{\mathrm{P},i}(\boldsymbol{\rho}) := \frac{1}{N}\sum_{j=1}^{N}\nabla_{\boldsymbol{\rho}}\log\nu_{\boldsymbol{\rho}}(\boldsymbol{\theta}_j)C_i(\tau_j),$$

where N, called batch size, is the number of independent parameter-trajectory pairs $\{(\boldsymbol{\theta}_j,\tau_j)\}_{j=1}^{N}$ such that $\tau_j \sim p_{\mathrm{A}}(\cdot \mid \boldsymbol{\theta}_j)$ and $\boldsymbol{\theta}_j \sim \nu_{\boldsymbol{\rho}}$. We highlight that, for parameter-based exploration, we sample N policy parameterizations $\boldsymbol{\theta}_j$ from the hyperpolicy $\nu_{\boldsymbol{\rho}}$, for each of which we sample a single trajectory $\tau_j \sim p_{\mathrm{A}}(\cdot \mid \boldsymbol{\theta}_j)$. That being said, considering that:

$$\mathcal{L}_{\mathrm{P},\omega}(\boldsymbol{\rho},\boldsymbol{\lambda}) = J_{\mathrm{P},0}(\boldsymbol{\rho}) + \sum_{u=1}^{U}\lambda_u\left(J_{\mathrm{P},i}(\boldsymbol{\rho}) - b_u\right) - \frac{\omega}{2}\left\|\boldsymbol{\lambda}\right\|_2^2,$$

C-PGPE employs the following estimator to update the primal variable $\boldsymbol{\rho}$:

$$\widehat{\nabla}_{\boldsymbol{\rho}}\mathcal{L}_{\mathrm{P},\omega}(\boldsymbol{\rho},\boldsymbol{\lambda}) = \widehat{\nabla}_{\boldsymbol{\rho}}J_{\mathrm{P},0}(\boldsymbol{\rho}) + \sum_{u=1}^{U}\lambda_u\widehat{\nabla}_{\boldsymbol{\rho}}J_{\mathrm{P},i}(\boldsymbol{\rho}).$$

If we now focus on the parameter-based gradient w.r.t. Lagrange multipliers, we have the following:

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}_{\mathrm{P},\omega}(\boldsymbol{\rho},\boldsymbol{\lambda}) = \mathbf{J}_{\mathrm{P}}(\boldsymbol{\rho}) - \mathbf{b} - \omega\boldsymbol{\lambda},$$

where $\mathbf{J}_{\mathrm{P}}(\boldsymbol{\rho}) := (J_{\mathrm{P},1}(\boldsymbol{\rho}),\dots,J_{\mathrm{P},U}(\boldsymbol{\rho}))^{\top}$. Thus, its sample-based version used by C-PGPE to update the dual variable is the following:

$$\frac{\widehat{\partial}}{\partial\lambda_i}\mathcal{L}_{\mathrm{P},\omega}(\boldsymbol{\rho},\boldsymbol{\lambda}) = \frac{1}{N}\sum_{j=1}^{N}C_i(\tau_j) - b_i - \omega\lambda_i,$$

where the N independent trajectories $\{\tau_j\}_{j=1}^{N}$ are such that $\tau_j \sim p_{\mathrm{A}}(\cdot \mid \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j \sim \nu_{\boldsymbol{\rho}}$ for every $j \in [\![N]\!]$.

4. Deterministic Policy Deployment of C-PG

In this section, we analyze the convergence guarantees of C-PG towards an optimal deterministic policy. To this end, we focus on the setting of *white noise (hyper)policies* (Montenegro et al., 2024a). Specifically, we restrict both action-based and parameter-based exploration strategies as stochastic perturbations of an underlying parametric deterministic policy $\mu_{\boldsymbol{\theta}}$. In this framework, stochastic policies $\pi_{\boldsymbol{\theta}}$ are modeled as perturbations of the actions prescribed by $\mu_{\boldsymbol{\theta}}$, while stochastic hyperpolicies $\nu_{\boldsymbol{\rho}}$ are interpreted as perturbations of the parameters $\boldsymbol{\theta}$ of the deterministic policy. Leveraging this structure, we study the behavior of the C-PG algorithm when it learns using stochastic (hyper)policies and subsequently *deploys* their deterministic counterpart by *switching off* the stochasticity at the end of training. Our analysis establishes the sample complexity required by C-PG to guarantee that the deployed deterministic policy, regardless of the exploration paradigm used during training, is an optimal feasible one.

We begin by presenting the noise model used in our analysis (Section 4.1), followed by a description of the deterministic policy deployment process in CMDPs (Section 4.2). We then state the assumptions required for convergence (Section 4.3), and finally provide the sample complexity required by C-PG to converge in the last iterate to an optimal deterministic policy when the noise is *switched off* at the end of training (Section 4.4).

4.1. White Noise Exploration

While deterministic policies are desirable in real-world applications (see Section 1), learning them directly typically requires off-policy actor-critic architectures (Silver et al., 2014; Lillicrap et al., 2015; Xiong et al., 2022), which pose significant challenges for convergence analysis even in unconstrained settings. In this part, we introduce a specific noise model that enables us to restrict both action-based and parameter-based exploration strategies as stochastic perturbations of deterministic policies. This formulation allows us to quantify the performance gap induced by a given parameterization of a stochastic (hyper)policy w.r.t. its associated deterministic policy.

We begin by defining deterministic policies and the related performance and cost functions. A parametric deterministic policy is a function $\mu_{\boldsymbol{\theta}} : \mathcal{S} \to \mathcal{A}$, where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{d_\Theta}$ is the parameter vector. For every $i \in [\![0, U]\!]$, the performance and cost functions $J_{\mathrm{D},i} : \Theta \to \mathbb{R}$ related to a deterministic policy are:

$$J_{\mathrm{D},i}(\boldsymbol{\theta}) := \mathop{\mathbb{E}}_{\tau \sim p_{\mathrm{D}}(\cdot|\boldsymbol{\theta})} \left[C_i(\tau) \right],$$

where $p_{\mathrm{D}}(\tau; \boldsymbol{\theta}) := \phi_0(\boldsymbol{s}_0) \prod_{t=0}^{T-1} p(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \mu_{\boldsymbol{\theta}}(\boldsymbol{s}_t))$ is the probability density of trajectory τ induced by $\mu_{\boldsymbol{\theta}}$. Using these definitions, we introduce the deterministic regularized Lagrangian function employed by C-PG as:

$$\mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = J_{\mathrm{D},0}(\boldsymbol{\theta}) + \sum_{i=1}^{U} \left(\lambda_i (J_{\mathrm{D},i}(\boldsymbol{\theta}) - b_i) \right) - \frac{\omega}{2} \|\boldsymbol{\lambda}\|_2^2.$$

We can now redefine both AB and PB exploration on top of deterministic policies (Montenegro et al., 2024a). Considering AB exploration, stochasticity is injected at the action level, perturbing the deterministic policy's output at each environmental interaction step. In PB exploration, noise is applied directly to the policy parameters before execution, resulting in a fixed perturbed version of the underlying deterministic policy for an entire trajectory. Next, we formally present how we intend a perturbation in the action or parameter spaces.

Definition 4.1 (White Noise)**.** *Let $d \in \mathbb{N}$ and $\sigma \in \mathbb{R}_{>0}$. A probability distribution $\Phi_d \in \Delta\left(\mathbb{R}^d\right)$ is defined as* white noise *if it satisfies the following conditions:*

$$\mathbb{E}_{\epsilon \sim \Phi_d}[\boldsymbol{\epsilon}] = \mathbf{0}_d, \quad \mathbb{E}_{\epsilon \sim \Phi_d}[\|\boldsymbol{\epsilon}\|_2^2] \leq d\sigma^2, \tag{7}$$

where $\mathbf{0}_d \in \mathbb{R}^d$ is a d-dimensional vector of all zero components.

Definition 4.1 includes zero-mean Gaussian distributions $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_d, \sigma\boldsymbol{\Lambda})$ with $\lambda_{\max}(\boldsymbol{\Lambda}) = 1$, ensuring that $\mathbb{E}[\|\boldsymbol{\epsilon}\|_2^2] = \sigma^2 \mathrm{tr}(\boldsymbol{\Lambda}) \leq d\sigma^2$. We stress that this noise has to be considered *white* across exploration steps. We can now redefine action-based and parameter-based explorations as white noise perturbations of the actions or the parameters of an underlying parametric deterministic policy $\mu_{\boldsymbol{\theta}}$.

Action-Based (AB) Exploration. Considering AB PG methods (see Section 2), we consider a *parametric stochastic policy* $\pi_{\boldsymbol{\theta}}$ as built upon an underlying deterministic policy $\mu_{\boldsymbol{\theta}}$ by perturbing each action suggested by $\mu_{\boldsymbol{\theta}}$ with a white noise random vector. Formally, we consider the following definition of white noise policies.

Definition 4.2 (White Noise Policies)**.** *Let $\boldsymbol{\theta} \in \Theta$ and $\mu_{\boldsymbol{\theta}} : \mathcal{S} \to \mathcal{A}$ be a parametric deterministic policy. Given a white noise distribution $\Phi_{d_{\mathcal{A}}}$ (Definition 4.1), a white-noise-based policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \to \Delta(\mathcal{A})$ is defined such that, for every state $\boldsymbol{s} \in \mathcal{S}$, the action $\boldsymbol{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s})$ satisfies $\boldsymbol{a} = \mu_{\boldsymbol{\theta}}(\boldsymbol{s}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \Phi_{d_{\mathcal{A}}}$ which is sampled independently at every step (i.e., whenever an action is sampled).*

We highlight that Definition 4.2 further justifies the name for AB exploration since the exploration is carried out at the action level.

Next, for every $i \in [\![0, U]\!]$ we redefine the cost functions $J_{\mathrm{A},i}$ leveraging the introduced characterization of white noise policies. To this end, we need to introduce the concept of *non stationary* deterministic cost functions. Let $\underline{\boldsymbol{\epsilon}} = (\boldsymbol{\epsilon}_t)_{t=0}^{T-1}$ be a sequence of independently sampled white noise vectors satisfying Definition 4.1. Let $\underline{\boldsymbol{\mu}} = (\boldsymbol{\mu}_t)_{t=0}^{T-1}$ be a non stationary deterministic policy where, at time step t, the deterministic policy $\boldsymbol{\mu}_t : \mathcal{S} \to \mathcal{A}$ is played, with $\boldsymbol{\mu}_t = \mu_{\boldsymbol{\theta}} + \boldsymbol{\epsilon}_t$. For every $i \in [\![0, U]\!]$, we introduce the cost functions for this kind of policy: $J_{\mathrm{D},i}(\underline{\boldsymbol{\mu}}) := \mathbb{E}_{\tau \sim p_{\mathrm{D}}(\cdot|\underline{\boldsymbol{\mu}})}[C_i(\tau)]$, where $p_{\mathrm{D}}(\tau, \underline{\boldsymbol{\mu}})$ is the density of a trajectory τ induced by the non stationary deterministic policy $\underline{\boldsymbol{\mu}}$. Equipped with this new definition, we can reintroduce the AB cost functions $J_{\mathrm{A},i}$ which admit the following definition, together with the one already provided in Section 2, when the considered stochastic policy $\pi_{\boldsymbol{\theta}}$ complies with Definition 4.2:

$$J_{\mathrm{A},i}(\boldsymbol{\theta}) := \mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \sim \Phi_{d_{\mathcal{A}}}^T} \left[J_{\mathrm{D},i}(\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + \underline{\boldsymbol{\epsilon}}) \right],$$

where $\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + \underline{\boldsymbol{\epsilon}} = (\mu_{\boldsymbol{\theta}} + \boldsymbol{\epsilon}_t)_{t=0}^{T-1}$ and $\Phi_{d_{\mathcal{A}}}$ is a white noise distribution compliant with Definition 4.1..

Parameter-Based (PB) Exploration. Considering PB PG methods (see Section 2), we consider a *parametric stochastic hyperpolicy* $\nu_{\boldsymbol{\theta}}$ as built upon an underlying deterministic policy $\mu_{\boldsymbol{\theta}}$ by perturbing the parameter vector $\boldsymbol{\theta}$ with a white noise random vector. Formally, we consider the following definition of white-noise hyperpolicies.

Definition 4.3 (White Noise Hyperpolicies)**.** *Let $\boldsymbol{\theta} \in \Theta$ and $\mu_{\boldsymbol{\theta}} : \mathcal{S} \to \mathcal{A}$ be a parametric deterministic policy. Given a white noise distribution Φ_{d_Θ} (Definition 4.1), a white-noise-based hyperpolicy $\nu_{\boldsymbol{\theta}} \in \Delta(\Theta)$ is defined such that, for every parameter $\boldsymbol{\theta} \in \Theta$, the perturbed parameter $\boldsymbol{\theta}' \sim \nu_{\boldsymbol{\theta}}$ satisfies $\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}$, independently for every trajectory.*

As previously done for action-based exploration, we stress that this definition further justifies the name of PB exploration, since the exploration is carried out at parameter level. Moreover, we let the reader note that the noise $\boldsymbol{\epsilon}$ is sampled once at the *beginning* of each trajectory, meaning that the resulting policy $\mu_{\boldsymbol{\theta}+\boldsymbol{\epsilon}}$ remains deterministic throughout the entire trajectory collection phase.

Next, for every $i \in [\![0, U]\!]$ we redefine the cost functions $J_{\mathrm{P},i}$ leveraging the introduced characterization of white noise hyperpolicies. We reintroduce the PB cost functions, which admit the following definition, together with the one already provided in Section 2, when the considered stochastic hyperpolicy $\nu_{\boldsymbol{\theta}}$ complies with Definition 4.3:

$$J_{\mathrm{P},i}(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}} \left[J_{\mathrm{D},i}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) \right],$$

where Φ_{d_Θ} is a white noise distribution compliant with Definition 4.1.

In the remaining part of this section, we will consider an *exploration-agnostic* setting, in which we denote with $\dagger \in \{\mathrm{A}, \mathrm{P}\}$ the two different exploration approaches. We highlight that the problem formulation and all the theoretical results of Section 3 still hold.

4.2. Deploying Deterministic Policies in CMDPs

In this part, we analyze the effect of switching off the stochasticity on the regularized Lagrangian employed by `C-PG` to solve the COP in Equation (1) when dealing with stochastic policies and hyperpolicies that satisfy Definitions 4.2 and 4.3, respectively.

Before presenting such results, we introduce two assumptions enforcing the regularity of the deterministic objectives $J_{\mathrm{D},i}$ w.r.t. the parameters $\boldsymbol{\theta}$ and the non stationary deterministic policies $\underline{\boldsymbol{\mu}}$ associated with AB exploration.

Assumption 4.1 ($J_{\mathrm{D},i}$ Regularity w.r.t. $\boldsymbol{\theta}$). *For every $i \in [\![0, U]\!]$, there exist $L_{1D,i}, L_{2D,i} \in \mathbb{R}_{>0}$ such that, for every $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, the following conditions hold:*

$$\left|J_{D,i}(\boldsymbol{\theta}) - J_{D,i}(\boldsymbol{\theta}')\right| \leq L_{1D,i} \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2 \quad \text{and} \quad \left\|\nabla_{\boldsymbol{\theta}} J_{D,i}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} J_{D,i}(\boldsymbol{\theta}')\right\|_2 \leq L_{2D,i} \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2.$$

Moreover, we denote $L_{1D,\max} := \max_{i\in[\![0,U]\!]} L_{1D,i}$ and $L_{2D,\max} := \max_{i\in[\![0,U]\!]} L_{2D,i}$.

Assumption 4.2 ($J_{\mathrm{D},i}$ Regularity w.r.t. $\underline{\boldsymbol{\mu}}$). *For every $i \in [\![0, U]\!]$, there exist $L_{1\underline{\boldsymbol{\mu}},i}, L_{2\underline{\boldsymbol{\mu}},i} \in \mathbb{R}_{>0}$ such that, for every pair of non stationary deterministic policies $\underline{\boldsymbol{\mu}}, \underline{\boldsymbol{\mu}}'$, the following conditions hold:*

$$\left|J_{D,i}(\underline{\boldsymbol{\mu}}) - J_{D,i}(\underline{\boldsymbol{\mu}}')\right| \leq L_{1\underline{\boldsymbol{\mu}},i} \sum_{t=0}^{T-1} \sup_{s\in\mathcal{S}} \left\|\boldsymbol{\mu}_t(s) - \boldsymbol{\mu}'_t(s)\right\|_2 \quad \text{and}$$

$$\left\|\nabla_{\underline{\boldsymbol{\mu}}} J_{D,i}(\underline{\boldsymbol{\mu}}) - \nabla_{\underline{\boldsymbol{\mu}}} J_{D,i}(\underline{\boldsymbol{\mu}}')\right\|_2 \leq L_{2\underline{\boldsymbol{\mu}},i} \sum_{t=0}^{T-1} \sup_{s\in\mathcal{S}} \left\|\boldsymbol{\mu}_t(s) - \boldsymbol{\mu}'_t(s)\right\|_2.$$

Moreover, we denote $L_{1\underline{\boldsymbol{\mu}},\max} := \max_{i\in[\![0,U]\!]} L_{1\underline{\boldsymbol{\mu}},i}$ and $L_{2\underline{\boldsymbol{\mu}},\max} := \max_{i\in[\![0,U]\!]} L_{2\underline{\boldsymbol{\mu}},i}$.

We stress that these assumptions will be crucial for presenting the core result of this section regarding the effects on $\mathcal{L}_{\dagger,\omega}$ when switching off the stochasticity in the context of white noise exploration (see Section 4.1). Additionally, we let the reader note that Assumption 4.1 induces both $J_{\mathrm{A},i}$ and $J_{\mathrm{P},i}$ to enjoy the same regularity condition stated in such an assumption when considering (hyper)policies complying with Definitions 4.2 and 4.3 (Montenegro et al., 2024a). Moreover, the Lipschitz constants are fully characterized in (Montenegro et al., 2024a).

We are now ready to analyze the effect of switching off the stochasticity in PB and AB exploration on the regularized Lagrangian $\mathcal{L}_{\dagger,\omega}$ employed by the `C-PG` method.

Theorem 4.1. *Considering (hyper)policies complying with Definitions 4.2 (AB) or 4.3 (PB), under Assumptions 4.1 (PB) or 4.2 (AB), the following results hold:*

i. (Uniform Bound) for every $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\lambda} \in \mathbb{R}^U_{\geq 0}$:

$$\left|\mathcal{L}_{D,\omega}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}, \boldsymbol{\lambda})\right| \leq \left(1 + \|\boldsymbol{\lambda}\|_1\right) L_{1\dagger}\sigma\sqrt{d_\dagger}.$$

ii. ($\mathcal{L}_{D,\omega}$ Upper Bound) let $(\boldsymbol{\theta}^*_{D,\omega}, \boldsymbol{\lambda}^*_{D,\omega})$ be a saddle point of $\mathcal{L}_{D,\omega}$ and let $(\boldsymbol{\theta}^*_{\dagger,\omega}, \boldsymbol{\lambda}^*_{\dagger,\omega})$ be a saddle point of $\mathcal{L}_{\dagger,\omega}$. Then:

$$\mathcal{L}_{D,\omega}(\boldsymbol{\theta}^*_{\dagger,\omega}, \boldsymbol{\lambda}^*_{\dagger,\omega}) - \mathcal{L}_{D,\omega}(\boldsymbol{\theta}^*_{D,\omega}, \boldsymbol{\lambda}^*_{D,\omega}) \leq 2\left(1 + \|\boldsymbol{\lambda}^*_{\dagger,\omega}\|_1\right) L_{1\dagger}\sigma\sqrt{d_\dagger}.$$

Where $L_{1P} := L_{1D,\max}$, $L_{1A} := L_{1\underline{\boldsymbol{\mu}},\max}$, $d_P := d_\Theta$, and $d_A := d_{\mathcal{A}}$.

Proof. We start the derivation by recalling the explicit form of $\left|\mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}, \boldsymbol{\lambda})\right|$:

$$\begin{aligned}
&\left|\mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}, \boldsymbol{\lambda})\right| \\
&= \left| J_{\mathrm{D},0}(\boldsymbol{\theta}) + \sum_{i=1}^{U} \lambda_i \left(J_{\mathrm{D},i}(\boldsymbol{\theta}) - b_i\right) - \frac{\omega}{2}\|\boldsymbol{\lambda}\|_2^2 - J_{\dagger,0}(\boldsymbol{\theta}) - \sum_{i=1}^{U} \lambda_i \left(J_{\dagger,i}(\boldsymbol{\theta}) - b_i\right) + \frac{\omega}{2}\|\boldsymbol{\lambda}\|_2^2 \right| \\
&\leq \left|J_{\mathrm{D},0}(\boldsymbol{\theta}) - J_{\dagger,0}(\boldsymbol{\theta})\right| + \sum_{i=1}^{U} \lambda_i \left|J_{\mathrm{D},i}(\boldsymbol{\theta}) - J_{\dagger,i}(\boldsymbol{\theta})\right|,
\end{aligned}$$

where the last line follows by simply having applied the triangular inequality.

To continue the proof, we need to resort to Theorems 5.1 (PB) and 5.2 (AB) by (Montenegro et al., 2024a). These state that under Assumptions 4.1 (PB) or 4.2 (AB), when dealing with an (hyper)policy complying with Definitions 4.2 (AB) and 4.3 (PB), the following holds:

$$\left|J_{\mathrm{D},i}(\boldsymbol{\theta}) - J_{\dagger,i}(\boldsymbol{\theta})\right| \leq L_{1\dagger,i}\sigma\sqrt{d_\dagger},$$

where $L_{1\mathrm{P},i} := L_{1\mathrm{D},i}$, $L_{1\mathrm{A},i} := L_{1\underline{\boldsymbol{\mu}},i}$, $d_{\mathrm{P}} = d_{\Theta}$, and $d_{\mathrm{A}} = d_{\mathcal{A}}$.

By leveraging this result, the following holds:

$$\begin{aligned}\left|\mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda}) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta},\boldsymbol{\lambda})\right| &\leq \left|J_{\mathrm{D},0}(\boldsymbol{\theta}) - J_{\dagger,0}(\boldsymbol{\theta})\right| + \sum_{i=1}^{U}\lambda_i\left|J_{\mathrm{D},i}(\boldsymbol{\theta}) - J_{\dagger,i}(\boldsymbol{\theta})\right| \\ &\leq \left(1+\sum_{i=1}^{U}\lambda_i\right)L_{1\dagger}\sigma\sqrt{d_{\dagger}} \\ &= \left(1+\|\boldsymbol{\lambda}\|_1\right)L_{1\dagger}\sigma\sqrt{d_{\dagger}},\end{aligned}$$

being $L_{1\mathrm{P}} := L_{1\mathrm{D},\max}$ and $L_{1\mathrm{A}} := L_{1\underline{\boldsymbol{\mu}},\max}$, which concludes the first part of the proof.

We can now face the second part of the proof. In particular, let $(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega})$ be a saddle point of $\mathcal{L}_{\dagger,\omega}$ and let $(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\mathrm{D},\omega})$ be a saddle point of $\mathcal{L}_{\mathrm{D},\omega}$. Before going on with the derivation, we recall that a saddle point by definition satisfies the following property:

$$\mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}) \leq \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\mathrm{D},\omega}) \leq \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda}^*_{\mathrm{D},\omega}),$$

for every $\boldsymbol{\theta}\in\Theta$ and $\boldsymbol{\lambda}\in\mathbb{R}^U_{\geq 0}$. That being said, the following holds:

$$\begin{aligned}&\mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\mathrm{D},\omega}) \\ &\leq \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) \\ &= \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) \pm \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) \\ &\leq \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) + \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) \\ &\leq \left|\mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega})\right| + \left|\mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\dagger,\omega})\right| \\ &\leq 2\left(1+\left\|\boldsymbol{\lambda}^*_{\dagger,\omega}\right\|_1\right)L_{1\dagger}\sigma\sqrt{d_{\dagger}},\end{aligned}$$

where we have just exploited the previously recalled property of saddle points and, in the last line, the result proved in the first part of this proof. □

Some comments are in order. Theorem 4.1 quantifies two sources of error: (i) is the gap $|\mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda}) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta},\boldsymbol{\lambda})|$ incurred when *switching off* the stochasticity of the (hyper)policy; (ii) is the error $\mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\dagger,\omega},\boldsymbol{\lambda}^*_{\dagger,\omega}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega},\boldsymbol{\lambda}^*_{\mathrm{D},\omega})$ arising when *deploying* the parameters of the learned stochastic (hyper)policy. We highlight that both error terms scale linearly with the stochasticity level σ in the (hyper)policy, and with the regularity constants introduced in Assumptions 4.1 (PB) and 4.2 (AB). In addition, the losses depend on the ℓ_1-norm of the Lagrange multipliers. In particular, the second bound depends on the ℓ_1-norm of the Lagrange multiplier at the saddle point of the regularized stochastic Lagrangian $\mathcal{L}_{\dagger,\omega}$. They also depend on the problem dimensionality, denoted $d_{\dagger}$, which corresponds to the parameter space dimensionality d_{Θ} in the PB case and to the action space dimensionality $d_{\mathcal{A}}$ in the AB case. We further note that AB exploration embeds an additional dependence on the interaction horizon T within the constant $L_{1\underline{\boldsymbol{\mu}},\max}$. Finally, this result allows us to recover the well-known trade-off between PB and AB exploration strategies (Metelli, Papini, Faccio and Restelli, 2018; Montenegro et al., 2024a): the former may suffer in high-dimensional parameter spaces (large d_{Θ}), whereas the latter may struggle with high-dimensional action spaces or long interaction horizons (large $d_{\mathcal{A}}$ or large T).

The analysis presented here will play a central role for establishing the sample complexity of `C-PG` when learning an optimal feasible deterministic policy via stochastic (hyper)policies and subsequently deploying their deterministic counterpart by switching off the stochasticity at the end of training.

4.3. Conditions for Convergence

In order to establish the convergence guarantees of `C-PG` to the optimal feasible deterministic policy, achieved by switching off the stochasticity after learning an optimal stochastic (hyper)policy, we proceed as follows. First, we leverage Theorem 3.2 to quantify the sample complexity required for learning an optimal stochastic (hyper)policy in the last iterate of C-PG. Then, we leverage Theorem 4.1 to characterize the loss incurred in terms of potential function $\mathcal{P}_K(\chi)$ when transitioning from a stochastic (hyper)policy to its deterministic counterpart by setting $\sigma = 0$.

To apply Theorem 3.2, it is necessary to verify the set of assumptions introduced in Section 3.2. In this part, we revisit those assumptions and, when possible, we aim to minimize their number by showing that, under the noise model introduced in Section 4.1 to represent both the AB and PB exploration paradigms, most of the required conditions can be inherited from analogous regularity properties imposed on the underlying deterministic-policy-dependent quantities.

Saddle Point Existence. Assumption 3.1 enforces the existence of a saddle point $(\boldsymbol{\theta}^*_{\dagger,\omega}, \boldsymbol{\lambda}^*_{\dagger,\omega})$ for the *stochastic* Lagrangian $\mathcal{L}_{\dagger,\omega}$. This assumption is only needed in Theorem 3.1 to map the *stochastic* (i.e., associated to $\mathcal{L}_{\dagger,\omega}$) potential function $\mathcal{P}_{\dagger,K}(\chi)$ to the $J_{\dagger,i}$ terms, demonstrating it is a useful tool in quantifying the sample complexity to ensure last-iterate global convergence of C-PG. Similarly, it is possible to show the same mapping between $\mathcal{P}_{\mathrm{D},K}(\chi)$ to the $J_{\mathrm{D},i}$ terms as done in Theorem 3.1, by applying the same theorem straightforwardly under the existence of a saddle point $(\boldsymbol{\theta}^*_{\mathrm{D},\omega}, \boldsymbol{\lambda}^*_{\mathrm{D},\omega})$ for the *deterministic* Lagrangian $\mathcal{L}_{\mathrm{D},\omega}$. Since the *stochastic* saddle point $(\boldsymbol{\theta}^*_{\dagger,\omega}, \boldsymbol{\lambda}^*_{\dagger,\omega})$ is not related to the *deterministic* one $(\boldsymbol{\theta}^*_{\mathrm{D},\omega}, \boldsymbol{\lambda}^*_{\mathrm{D},\omega})$, we need to assume the existence of both.

Weak ψ-Gradient Domination. Assumption 3.2 is a core component of the theoretical analysis of C-PG, as it enables regularization solely w.r.t. the Lagrange multipliers $\boldsymbol{\lambda}$, as further discussed in Section 3.2. Rather than directly assuming weak ψ-gradient domination on the stochastic Lagrangian $\mathcal{L}_{\dagger,0}$, we demonstrate that this property can be inherited from the corresponding assumption on the deterministic Lagrangian $\mathcal{L}_{\mathrm{D},0}$, for both exploration paradigms.

Assumption 4.3 (Weak ψ-Gradient Domination on $\mathcal{L}_{\mathrm{D},0}$)**.** *Let $\psi \in [1,2]$. There exist $\alpha_D > 0$ and $\beta_D \geq 0$ such that, for every $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\lambda} \in \mathbb{R}^U_{\geq 0}$, it holds that:*

$$\left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{D,0}(\boldsymbol{\theta}, \boldsymbol{\lambda})\right\|_2^{\psi} \geq \alpha_D \left(\mathcal{L}_{D,0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{D,0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) \right) - \beta_D.$$

This leads to a characterization analogous to that of Assumption 3.2, but imposed on the deterministic Lagrangian $\mathcal{L}_{\mathrm{D},0}$ rather than directly on the stochastic one $\mathcal{L}_{\dagger,0}$, that just inherits this property, as we later show. Before establishing this inheritance result, we introduce an additional assumption that is required only in the case of AB exploration.

Assumption 4.4 (Regularity of $\mu_{\boldsymbol{\theta}}$)**.** *There exists $L_{1\mu} \in \mathbb{R}_{\geq 0}$ such that, for every $\mathbf{s} \in \mathcal{S}$ and $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, the following holds:*

$$\left\|\mu_{\boldsymbol{\theta}}(\mathbf{s}) - \mu_{\boldsymbol{\theta}'}(\mathbf{s})\right\|_2 \leq L_{1\mu} \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2.$$

This assumption only requires that the deterministic policy $\mu_{\boldsymbol{\theta}}$ is $L_{1\mu}$-LC w.r.t. its parameters, which is needed to inherit regularity properties from the deterministic objectives $J_{\mathrm{D},i}$ in the AB exploration paradigm (Montenegro et al., 2024a). We are now ready to state the inheritance of the weak ψ-gradient domination.

Theorem 4.2 (Inherited Weak ψ-Gradient Domination on $\mathcal{L}_{\dagger,0}$)**.** *Consider an (hyper)policy complying with Definitions 4.2 (AB) or 4.3 (PB). Under Assumptions 4.1 (PB) or 4.2 (AB), 4.3, and 4.4 (AB), for any $\psi \in [1,2]$, $\boldsymbol{\theta} \in \Theta$, and $\boldsymbol{\lambda} \in \mathbb{R}^U_{\geq 0}$, the following holds:*

$$\left\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}, \boldsymbol{\lambda})\right\|_2^{\psi} \geq \alpha_D \left(\mathcal{L}_{\dagger,0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) \right) - \beta_{\dagger}(\sigma, \psi),$$

where:

$$\beta_P(\sigma,\psi) := \beta_D + \left(2\alpha_D L_{1D,\max} + (1+\|\boldsymbol{\lambda}\|_1)^{\psi-1} L_{2D,\max}^{\psi} \sigma^{\psi-1} d_{\Theta}^{\psi/2-1}\right)(1+\|\boldsymbol{\lambda}\|_1)\sigma\sqrt{d_{\Theta}},$$

and:

$$\beta_A(\sigma,\psi) := \beta_D + \left(2\alpha_D L_{1\underline{\mu},\max} + (1+\|\boldsymbol{\lambda}\|_1)^{\psi-1} L_{1\mu}^{\psi} L_{2D,\max}^{\psi} \sigma^{\psi-1} T^{\psi/2} d_{\Theta}^{\psi/2-1}\right)(1+\|\boldsymbol{\lambda}\|_1)\sigma\sqrt{d_{\mathcal{A}}}.$$

Proof. We begin the proof by considering the term $\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}', \boldsymbol{\lambda})$. Given that we consider an (hyper)policy complying with Definitions 4.2 (AB) or 4.3 (PB), and being under Assumptions 4.1 (PB) and 4.2 (AB), we can apply Theorem 4.1, stating that:

$$\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \geq -(1 + \|\boldsymbol{\lambda}\|_1) L_{1\dagger} \sigma \sqrt{d_\dagger},$$

where for $\dagger = \mathrm{A}$ we have $L_{1\mathrm{A}} = L_{1\boldsymbol{\mu},\max}$ and $d_{\mathrm{A}} = d_{\mathcal{A}}$, and for $\dagger = \mathrm{P}$ we have $L_{1\mathrm{P}} = L_{1\mathrm{D},\max}$ and $d_{\mathrm{P}} = d_\Theta$. That being said, the following derivation holds:

$$\begin{aligned}
&\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) \\
&= \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) \pm \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \\
&\geq -(1 + \|\boldsymbol{\lambda}\|_1) L_{1\dagger} \sigma \sqrt{d_\dagger} + \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) \pm \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) \\
&\geq \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) - (1 + \|\boldsymbol{\lambda}\|_1) L_{1\dagger} \sigma \sqrt{d_\dagger} + \min_{\boldsymbol{\theta} \in \Theta} \left(\mathcal{L}_{\dagger,0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) - \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) \right) \\
&\geq \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) - 2(1 + \|\boldsymbol{\lambda}\|_1) L_{1\dagger} \sigma \sqrt{d_\dagger},
\end{aligned}$$

where we applied Theorem 4.1 twice.

Thus, starting from the statement of Assumption 4.3, the following holds:

$$\begin{aligned}
\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda})\|_2^{\psi} &\geq \alpha_{\mathrm{D}} \left(\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) \right) - \beta_{\mathrm{D}} \\
&\geq \alpha_{\mathrm{D}} \left(\mathcal{L}_{\dagger,0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}' \in \Theta} \mathcal{L}_{\dagger,0}(\boldsymbol{\theta}', \boldsymbol{\lambda}) \right) - \beta_{\mathrm{D}} - 2\alpha_{\mathrm{D}} (1 + \|\boldsymbol{\lambda}\|_1) L_{1\dagger} \sigma \sqrt{d_\dagger}. \qquad (8)
\end{aligned}$$

To continue the proof, we need to consider separately the two exploration paradigms, in order to properly exploit the corresponding noise model.

PB Exploration. According to Definition 4.3, we can rewrite the PB Lagrangian as:

$$\mathcal{L}_{\mathrm{P},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}} \left[\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta} + \boldsymbol{\epsilon}, \boldsymbol{\lambda}) \right].$$

Thus, given $\alpha \in [0, 1]$ and defining $\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\epsilon}} := \alpha \boldsymbol{\theta} + (1 - \alpha)(\boldsymbol{\theta} + \boldsymbol{\epsilon})$, the following holds:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{P},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}} \left[\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta} + \boldsymbol{\epsilon}, \boldsymbol{\lambda}) \right] = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}} \left[\boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_{\mathrm{D},0}(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\epsilon}}, \boldsymbol{\lambda}) \right],$$

where we have simply applied the Taylor expansion centered in $\boldsymbol{\epsilon} = \mathbf{0}_{d_\Theta}$. Now, by applying the Euclidean norm, we have the following:

$$\begin{aligned}
\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{P},0}(\boldsymbol{\theta}, \boldsymbol{\lambda})\|_2 &= \left\| \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) + \mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}} \left[\boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_{\mathrm{D},0}(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\epsilon}}, \boldsymbol{\lambda}) \right] \right\|_2 \\
&\geq \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda})\|_2 - \left\| \mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}} \left[\boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_{\mathrm{D},0}(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\epsilon}}, \boldsymbol{\lambda}) \right] \right\|_2 \\
&\geq \|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda})\|_2 - \mathop{\mathbb{E}}_{\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}} \left[\left\| \boldsymbol{\epsilon}^\top \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_{\mathrm{D},0}(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\epsilon}}, \boldsymbol{\lambda}) \right\|_2 \right],
\end{aligned}$$

which follows by applying the triangular and Jensen's inequalities. Now, by applying the Cauchy-Schwartz inequality as $\mathbb{E}_{\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}}[\|\boldsymbol{\epsilon}\|_2] \leq \sqrt{\mathbb{E}_{\boldsymbol{\epsilon} \sim \Phi_{d_\Theta}}[\|\boldsymbol{\epsilon}\|_2^2]} \leq \sigma \sqrt{d_\Theta}$ and by exploiting the fact that, under Assumption 4.1:

$$\left\| \nabla_{\boldsymbol{\theta}}^2 \mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta}, \boldsymbol{\lambda}) \right\|_2 \leq \left\| \nabla_{\boldsymbol{\theta}}^2 J_{\mathrm{D},0}(\boldsymbol{\theta}) \right\|_2 + \sum_{i=1}^{U} \lambda_i \left\| \nabla_{\boldsymbol{\theta}}^2 J_{\mathrm{D},i}(\boldsymbol{\theta}) \right\|_2 \leq L_{2\mathrm{D},\max} \left(1 + \sum_{i=1}^{U} \lambda_i \right) = \left(1 + \|\boldsymbol{\lambda}\|_1 \right) L_{2\mathrm{D},\max},$$

we can conclude the following:

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2 &\le \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{P},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2 + \mathop{\mathbb{E}}_{\boldsymbol{\epsilon}\sim\Phi_{d_\Theta}}\left[\left\|\boldsymbol{\epsilon}^\top\nabla^2_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{D},0}(\widetilde{\boldsymbol{\theta}}_{\boldsymbol{\epsilon}},\boldsymbol{\lambda})\right\|_2\right]\\ &\le \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{P},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2 + \left(1+\|\boldsymbol{\lambda}\|_1\right)L_{2\mathrm{D},\max}\sigma\sqrt{d_\Theta}.\end{aligned}$$

Now, considering that $\psi\in[1,2]$, by exploiting the superadditivity of $(\cdot)^\psi$, we have:

$$\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2^\psi \le \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{P},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2^\psi + \left(\left(1+\|\boldsymbol{\lambda}\|_1\right)L_{2\mathrm{D},\max}\sigma\sqrt{d_\Theta}\right)^\psi.$$

Combining this last result with Equation (8), we conclude the inheritance of the weak ψ-GD from $\mathcal{L}_{\mathrm{D},\omega}$ in the PB case:

$$\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{P},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2^\psi \ge \alpha_{\mathrm{D}}\left(\mathcal{L}_{\mathrm{P},0}(\boldsymbol{\theta},\boldsymbol{\lambda}) - \min_{\boldsymbol{\theta}'\in\Theta}\mathcal{L}_{\mathrm{P},0}(\boldsymbol{\theta}',\boldsymbol{\lambda})\right) - \beta_{\mathrm{P}}(\sigma,\psi),$$

where:

$$\beta_{\mathrm{P}}(\sigma,\psi) := \beta_{\mathrm{D}} + \left(2\alpha_{\mathrm{D}}L_{1\mathrm{D},\max} + (1+\|\boldsymbol{\lambda}\|_1)^{\psi-1}L^\psi_{2\mathrm{D},\max}\sigma^{\psi-1}d_\Theta^{\psi/2-1}\right)(1+\|\boldsymbol{\lambda}\|_1)\sigma\sqrt{d_\Theta}.$$

AB Exploration. According to Definition 4.2, we can rewrite the AB Lagrangian as:

$$\mathcal{L}_{\mathrm{A},0}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \mathop{\mathbb{E}}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}\left[\mathcal{L}_{\mathrm{D},0}(\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}+\underline{\boldsymbol{\epsilon}},\boldsymbol{\lambda})\right].$$

Thus, given $\alpha\in[0,1]$ and defining $\underline{\widetilde{\boldsymbol{\mu}}}_{\boldsymbol{\theta}} := \alpha\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + (1-\alpha)(\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}+\underline{\boldsymbol{\epsilon}})$, the following holds:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{A},0}(\boldsymbol{\theta},\boldsymbol{\lambda}) &= \mathop{\mathbb{E}}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}\left[\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{D},0}(\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}+\underline{\boldsymbol{\epsilon}},\boldsymbol{\lambda})\right]\\ &= \mathop{\mathbb{E}}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}\left[\nabla_{\underline{\boldsymbol{\mu}}}\mathcal{L}_{\mathrm{D},0}(\underline{\boldsymbol{\mu}},\boldsymbol{\lambda})|_{\underline{\boldsymbol{\mu}}=\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}+\underline{\boldsymbol{\epsilon}}}\nabla_{\boldsymbol{\theta}}(\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}+\underline{\boldsymbol{\epsilon}})\right]\\ &= \mathop{\mathbb{E}}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}\left[\nabla_{\underline{\boldsymbol{\mu}}}\mathcal{L}_{\mathrm{D},0}(\underline{\boldsymbol{\mu}},\boldsymbol{\lambda})|_{\underline{\boldsymbol{\mu}}=\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}} + \underline{\boldsymbol{\epsilon}}^\top\nabla^2_{\underline{\boldsymbol{\mu}}}\mathcal{L}_{\mathrm{D},0}(\underline{\boldsymbol{\mu}},\boldsymbol{\lambda})|_{\underline{\boldsymbol{\mu}}=\underline{\widetilde{\boldsymbol{\mu}}}_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\right]\\ &= \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta},\boldsymbol{\lambda}) + \mathop{\mathbb{E}}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}\left[\underline{\boldsymbol{\epsilon}}^\top\nabla^2_{\underline{\boldsymbol{\mu}}}\mathcal{L}_{\mathrm{D},0}(\underline{\boldsymbol{\mu}},\boldsymbol{\lambda})|_{\underline{\boldsymbol{\mu}}=\underline{\widetilde{\boldsymbol{\mu}}}_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\right].\end{aligned}$$

where we simply applied the chain rule and the Taylor expansion centered in $\underline{\boldsymbol{\epsilon}} = \mathbf{0}_{d_\mathcal{A}T}$. Now, by applying the Euclidean norm, we have:

$$\begin{aligned}\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{A},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2 &= \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta},\boldsymbol{\lambda}) + \mathop{\mathbb{E}}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}\left[\underline{\boldsymbol{\epsilon}}^\top\nabla^2_{\underline{\boldsymbol{\mu}}}\mathcal{L}_{\mathrm{D},0}(\underline{\boldsymbol{\mu}},\boldsymbol{\lambda})|_{\underline{\boldsymbol{\mu}}=\underline{\widetilde{\boldsymbol{\mu}}}_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\right]\right\|_2\\ &\ge \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2 - \left\|\mathop{\mathbb{E}}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}\left[\underline{\boldsymbol{\epsilon}}^\top\nabla^2_{\underline{\boldsymbol{\mu}}}\mathcal{L}_{\mathrm{D},0}(\underline{\boldsymbol{\mu}},\boldsymbol{\lambda})|_{\underline{\boldsymbol{\mu}}=\underline{\widetilde{\boldsymbol{\mu}}}_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\right]\right\|_2\\ &\ge \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2 - \mathop{\mathbb{E}}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}\left[\left\|\underline{\boldsymbol{\epsilon}}^\top\nabla^2_{\underline{\boldsymbol{\mu}}}\mathcal{L}_{\mathrm{D},0}(\underline{\boldsymbol{\mu}},\boldsymbol{\lambda})|_{\underline{\boldsymbol{\mu}}=\underline{\widetilde{\boldsymbol{\mu}}}_{\boldsymbol{\theta}}}\nabla_{\boldsymbol{\theta}}\underline{\boldsymbol{\mu}}_{\boldsymbol{\theta}}\right\|_2\right],\end{aligned}$$

which follows from just applying the triangular and Jensen's inequalities. Now, by applying the Cauchy-Schwartz inequality as $\mathbb{E}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}[\|\underline{\boldsymbol{\epsilon}}\|_2] \le \sqrt{\mathbb{E}_{\underline{\boldsymbol{\epsilon}}\sim\Phi^T_{d_\mathcal{A}}}[\|\underline{\boldsymbol{\epsilon}}\|_2^2]} \le \sigma\sqrt{Td_\mathcal{A}}$ and exploiting Assumptions 4.2 and 4.4, by following the same procedure of the PB case, we obtain:

$$\left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{A},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2 \ge \left\|\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{D},0}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2 - (1+\|\boldsymbol{\lambda}\|_1)L_{1\mu}L_{2\underline{\boldsymbol{\mu}},\max}\sigma\sqrt{Td_\mathcal{A}},$$

and thus:

$$\|\nabla_{\theta}\mathcal{L}_{\mathrm{A},0}(\theta,\lambda)\|_2^{\psi} \geq \alpha_{\mathrm{D}}\left(\mathcal{L}_{\mathrm{A},0}(\theta,\lambda) - \min_{\theta'\in\Theta}\mathcal{L}_{\mathrm{A},0}(\theta',\lambda)\right) - \beta_{\mathrm{A}}(\sigma,\psi),$$

where

$$\beta_{\mathrm{A}}(\sigma,\psi) := \beta_{\mathrm{D}} + \left(2\alpha_{\mathrm{D}}L_{1\underline{\mu},\max} + (1+\|\lambda\|_1)^{\psi-1}L_{1\mu}^{\psi}L_{2\mathrm{D},\max}^{\psi}\sigma^{\psi-1}T^{\psi/2}d_{\Theta}^{\psi/2-1}\right)(1+\|\lambda\|_1)\sigma\sqrt{d_{\mathcal{A}}},$$

showing the inheritance of the weak ψ-GD from $\mathcal{L}_{\mathrm{D},\omega}$ in the AB case too. □

We highlight that the weak ψ-gradient domination property is inherited from the deterministic Lagrangian $\mathcal{L}_{\mathrm{D},0}$ with the *same* multiplicative constant α_{D}. Moreover, by setting $\sigma = 0$, one exactly recovers the result stated in Assumption 4.3. Finally, under the adopted noise model, this property follows directly from regularity assumptions on the deterministic cost functions $J_{\mathrm{D},i}$ (Assumptions 4.1 and 4.2) and on the deterministic policy μ_{θ} (Assumption 4.4), together with the bounds on the loss incurred when switching off the noise (see Theorem 4.1).

Regularity of $\mathcal{L}_{\dagger,0}$. Assumption 3.3 imposes regularity conditions on the stochastic Lagrangian $\mathcal{L}_{\dagger,0}$, which are standard in the literature on the convergence of primal-dual methods (Yang et al., 2020), as previously discussed in Section 3. As with the other conditions required to apply Theorem 3.2, these regularity properties can also be inherited from Assumption 4.1, as formalized in the following theorem.

Theorem 4.3 (Inherited Regularity of $\mathcal{L}_{\dagger,0}$). *Consider a (hyper)policy complying with Definitions 4.2 (AB) or 4.3 (PB). Under Assumption 4.1, for every $\lambda,\lambda' \in \mathbb{R}^U_{\geq 0}$ and $\theta,\theta' \in \Theta$ the following conditions hold:*

$$\begin{aligned}
&\|\nabla_{\lambda}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\lambda}\mathcal{L}_{\dagger,0}(\theta',\lambda)\|_2 \leq \sqrt{U}L_{1D,\max}\|\theta-\theta'\|_2,\\
&\|\nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta',\lambda)\|_2 \leq (1+\|\lambda\|_1)L_{2D,\max}\|\theta-\theta'\|_2,\\
&\|\nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda')\|_2 \leq L_{1D,\max}\|\lambda-\lambda'\|_2.
\end{aligned}$$

Proof. We start by proving that $\mathcal{L}_{\dagger,0}(\cdot,\lambda)$ is LS w.r.t. the parameters θ, thus, for every $\lambda \in \mathbb{R}^U_{\geq 0}$ and $\theta,\theta' \in \Theta$, we aim to find $L_2 \in \mathbb{R}_{\geq 0}$ such that:

$$\|\nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta',\lambda)\|_2 \leq L_2\|\theta-\theta'\|_2.$$

This can be easily done by expanding the $\nabla_{\theta}\mathcal{L}_{\dagger,0}$ terms:

$$\begin{aligned}
\|\nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta',\lambda)\|_2 &= \left\|\nabla_{\theta}J_{\dagger,0}(\theta) - \nabla_{\theta}J_{\dagger,0}(\theta') + \sum_{i=1}^{U}\lambda_i\left(\nabla_{\theta}J_{\dagger,i}(\theta) - \nabla_{\theta}J_{\dagger,i}(\theta')\right)\right\|_2\\
&\leq \|\nabla_{\theta}J_{\dagger,0}(\theta) - \nabla_{\theta}J_{\dagger,0}(\theta')\|_2 + \sum_{i=1}^{U}\lambda_i\|\nabla_{\theta}J_{\dagger,i}(\theta) - \nabla_{\theta}J_{\dagger,i}(\theta')\|_2,
\end{aligned}$$

by the triangular inequality. Now, we recover Lemmas D.3 and D.7 by Montenegro et al. (2024a), stating that under Assumption 4.1 both $J_{\mathrm{P},i}$ and $J_{\mathrm{A},i}$ are LS with the same constant of $J_{\mathrm{D},i}$. In our setting, it means that $J_{\dagger,i}$ is $L_{2\mathrm{D},\max}$-LS, for every $i \in [\![0,U]\!]$. That being said, the following holds:

$$\begin{aligned}
\|\nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta',\lambda)\|_2 &\leq \|\nabla_{\theta}J_{\dagger,0}(\theta) - \nabla_{\theta}J_{\dagger,0}(\theta')\|_2 + \sum_{i=1}^{U}\lambda_i\|\nabla_{\theta}J_{\dagger,i}(\theta) - \nabla_{\theta}J_{\dagger,i}(\theta')\|_2\\
&\leq \left(1+\sum_{i=1}^{U}\lambda_i\right)L_{2\mathrm{D},\max}\|\theta-\theta'\|_2\\
&= (1+\|\lambda\|_1)L_{2\mathrm{D},\max}\|\theta-\theta'\|_2,
\end{aligned}$$

thus having quantified the smoothness constant.

We can now proceed by proving that $\nabla_{\lambda}\mathcal{L}_{\dagger,0}(\cdot,\lambda)$ is LC, so we have to find a constant $L_1 \in \mathbb{R}_{\geq 0}$ such that, for every $\lambda \in \mathbb{R}^U_{\geq 0}$ and $\theta, \theta' \in \Theta$:

$$\left\|\nabla_{\lambda}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\lambda}\mathcal{L}_{\dagger,0}(\theta',\lambda)\right\|_2 \leq L_1 \left\|\theta - \theta'\right\|_2.$$

As done in the previous case, we expand the $\nabla_{\lambda}\mathcal{L}_{\dagger,0}$ terms:

$$\left\|\nabla_{\lambda}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\lambda}\mathcal{L}_{\dagger,0}(\theta',\lambda)\right\|_2 = \left\|\mathbf{J}_{\dagger}(\theta) - \mathbf{J}_{\dagger}(\theta')\right\|_2 = \sqrt{\sum_{i=0}^{U}\left(J_{\dagger,i}(\theta) - J_{\dagger,i}(\theta')\right)^2} \leq \sqrt{U} L_{1\mathrm{D},\max} \left\|\theta - \theta'\right\|_2,$$

where we just applied the same reasoning of Lemmas D.3 and D.7 by Montenegro et al. (2024a) to state that $J_{\dagger,i}$ is $L_{1\mathrm{D},\max}$-LS, for every $i \in [\![0, U]\!]$.

Finally, we prove that $\nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\cdot)$ is LC, finding a constant $L_3 \in \mathbb{R}_{\geq 0}$ such that, for every $\lambda, \lambda' \in \mathbb{R}^U_{\geq 0}$:

$$\left\|\nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda')\right\|_2 \leq L_3 \left\|\lambda - \lambda'\right\|_2.$$

As done before, we expand the $\nabla_{\theta}\mathcal{L}_{\dagger,0}$ terms:

$$\left\|\nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda) - \nabla_{\theta}\mathcal{L}_{\dagger,0}(\theta,\lambda')\right\|_2 = \left\|\nabla_{\theta}\mathbf{J}_{\dagger}(\theta)(\lambda - \lambda')\right\|_2 \leq L_{1\mathrm{D},\max} \left\|\lambda - \lambda'\right\|_2,$$

where we just exploited the fact that under Assumption 4.1 $J_{\mathrm{D},i}$ is $L_{1\mathrm{D},\max}$-LC for every $i \in [\![0, U]\!]$. □

We highlight that, as previously discussed in Section 3, the only constant that depends on $\mathcal{O}(\omega^{-1})$ when considering the learning process of `C-PG` is the smoothness constant of $\mathcal{L}_{\dagger,0}(\cdot,\lambda)$. This dependence arises from the term $\|\lambda\|_1$, which is upper bounded by $\mathcal{O}(\omega^{-1})$ due to the regularization employed in `C-PG`.

Bounded Estimators' Variances. The last condition for convergence we have to discuss is the one of Assumption 3.4, requiring that the unbiased estimators $\widehat{\nabla}_{\theta}\mathcal{L}_{\dagger,\omega}$ and $\widehat{\nabla}_{\lambda}\mathcal{L}_{\dagger,\omega}$ have bounded variance. Under the specific noise model at hand, this property holds under the following assumption.

Assumption 4.5 (Bounded Scores of Φ)**.** *Let $\Phi \in \Delta(\mathbb{R}^d)$ be a white noise complying with Definition 4.1 with variance bound $\sigma \in \mathbb{R}_{>0}$ and density ϕ. ϕ is differentiable in its argument and there exists universal constant $c \in \mathbb{R}_{>0}$ such that:*

$$\mathop{\mathbb{E}}_{\epsilon\sim\Phi}\left[\|\nabla_{\epsilon}\log\phi(\epsilon)\|_2^2\right] \leq c d\sigma^{-2} \quad \textit{and} \quad \mathop{\mathbb{E}}_{\epsilon\sim\Phi}\left[\|\nabla_{\epsilon}^2\log\phi(\epsilon)\|_2\right] \leq c\sigma^{-2}.$$

Intuitively, this assumption is equivalent to the more common ones requiring the boundedness of the expected norms of the score function and its gradient (Papini et al., 2022; Yuan et al., 2022). Note that a zero-mean Gaussian noise $\Phi = \mathcal{N}(\mathbf{0}_d, \mathbf{\Sigma})$ fulfills Assumption 4.5. Indeed, one has $\nabla_{\epsilon}\log\phi(\epsilon) = \mathbf{\Sigma}^{-1}\epsilon$ and $\nabla_{\epsilon}^2\log\phi(\epsilon) = \mathbf{\Sigma}^{-1}$. Thus, $\mathbb{E}[\|\nabla_{\epsilon}\log\phi(\epsilon)\|_2^2] = \mathrm{tr}(\mathbf{\Sigma}^{-1}) \leq d\lambda_{\min}(\mathbf{\Sigma})^{-1}$ and $\mathbb{E}[\|\nabla_{\epsilon}^2\log\phi(\epsilon)\|_2] = \lambda_{\min}(\mathbf{\Sigma})^{-1}$. In particular, for an isotropic Gaussian $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, we have $\lambda_{\min}(\mathbf{\Sigma}) = \sigma^2$, fulfilling Assumption 4.5 with $c = 1$.

Under Assumption 4.5, the variances of the estimators employed in `C-PG`, which are described in Section 3.4, are bounded, as shown by the following lemma.

Lemma 4.4 (Bounded Estimators' Variances)**.** *Consider a (hyper)policy complying with Definitions 4.2 (AB) and 4.3 (PB). Under Assumptions 4.4 (just for AB) and 4.5, the following conditions hold:*

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\lambda}\mathcal{L}_{\dagger,\omega}(\theta,\lambda)\right] \leq \frac{U(1-\gamma^T)^2}{N(1-\gamma)^2} =: V_{\lambda} \quad \textit{and} \quad \mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\theta}\mathcal{L}_{\dagger,\omega}(\theta,\lambda)\right] \leq \frac{Z_{\dagger,\theta}(1+\|\lambda\|_1)^2}{N\sigma^2} =: V_{\dagger,\theta},$$

where $Z_{P,\theta} := c d_{\Theta}\left(\frac{1-\gamma^T}{1-\gamma}\right)^2$ and $Z_{A,\theta} := c d_{\mathcal{A}} L_{1\mu}^2\left(\frac{1-\gamma^T}{1-\gamma}\right)^3$.

Proof. We start by bounding the variance of $\widehat{\nabla}_{\lambda}\mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \widehat{\mathbf{J}}_{\dagger}(\boldsymbol{\theta}) - \mathbf{b} - \omega\boldsymbol{\lambda}$, where $\widehat{\mathbf{J}}_{\dagger}(\boldsymbol{\theta}) = (\widehat{J}_{\dagger,0}(\boldsymbol{\theta}),\dots,\widehat{J}_{\dagger,U}(\boldsymbol{\theta}))^{\top}$ and $\widehat{J}_{\dagger,i}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{j=0}^{N-1} C_i(\tau_j)$ with $\tau_j \sim p_{\mathrm{A}}(\cdot,\boldsymbol{\theta})$ (in AB exploration) or $\tau_j \sim p_{\mathrm{D}}(\cdot,\boldsymbol{\theta}_j)$ and $\boldsymbol{\theta}_j \sim \nu_{\boldsymbol{\theta}}$ (in PB exploration). Thus, we can notice that the variance arises just from $\mathbf{J}_{\dagger}(\boldsymbol{\theta})$. Defining $\mathbf{C}(\tau) := (C_1(\tau),\dots,C_U(\tau))^{\top}$, where $C_i(\tau) = \sum_{t=0}^{T-1}\gamma^t c_i(\boldsymbol{s}_{\tau,t},\boldsymbol{a}_{\tau,t}) \le \frac{1-\gamma^T}{1-\gamma}$, the following holds:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\lambda}\mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta},\boldsymbol{\lambda})\right] = \frac{1}{N}\mathbb{V}\mathrm{ar}\left[\mathbf{C}(\tau_1)\right] = \frac{1}{N}\mathbb{E}\left[\|\mathbf{C}(\tau_1)\|_2^2\right] \le \frac{U(1-\gamma^T)^2}{N(1-\gamma)^2}.$$

We now bound the variance of $\widehat{\nabla}\mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta},\boldsymbol{\lambda})$, for which we distinguish the PB and the AB cases. Starting from PB exploration, we can express the estimator at hand as:

$$\begin{aligned}\left\|\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{P},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda})\right\|_2 &\le \frac{1}{N}\sum_{j=0}^{N-1}\left\|\nabla_{\boldsymbol{\theta}}\log\nu_{\boldsymbol{\theta}}(\boldsymbol{\theta}_j)\right\|_2\left\|C_0(\tau_j)+\sum_{i=1}^{U}\lambda_i C_i(\tau_j)\right\|_2\\ &\le \frac{(1+\|\boldsymbol{\lambda}\|_1)(1-\gamma^T)}{N(1-\gamma)}\sum_{j=0}^{N-1}\left\|\nabla_{\boldsymbol{\theta}}\log\nu_{\boldsymbol{\theta}}(\boldsymbol{\theta}_j)\right\|_2.\end{aligned}$$

Thus, we have the following:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{P},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda})\right] &= \frac{1}{N}\mathbb{V}\mathrm{ar}\left[\nabla_{\boldsymbol{\theta}}\log\nu_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1)\left(C_0(\tau_1)+\sum_{i=1}^{U}\lambda_i C_i(\tau_1)\right)\right]\\ &= \frac{1}{N}\mathbb{E}\left[\left\|\nabla_{\boldsymbol{\theta}}\log\nu_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1)\left(C_0(\tau_1)+\sum_{i=1}^{U}\lambda_i C_i(\tau_1)\right)\right\|_2^2\right]\\ &\le \frac{(1+\|\boldsymbol{\lambda}\|_1)^2(1-\gamma^T)^2}{N(1-\gamma)^2}\mathbb{E}\left[\|\nabla_{\boldsymbol{\theta}}\log\nu_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1)\|_2^2\right].\end{aligned}$$

We now recover Lemma E.4 by Montenegro et al. (2024a), stating that under Assumption 4.5 $\mathbb{E}[\|\nabla_{\boldsymbol{\theta}}\log\nu_{\boldsymbol{\theta}}(\boldsymbol{\theta}')\|_2^2] \le c d_{\Theta}\sigma^{-2}$ for every $\boldsymbol{\theta},\boldsymbol{\theta}' \in \Theta$. Thus, we can conclude the following:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{P},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda})\right] \le \frac{c d_{\Theta}(1+\|\boldsymbol{\lambda}\|_1)^2(1-\gamma^T)^2}{N\sigma^2(1-\gamma)^2}.$$

Switching to the AB case, we can express the estimator as:

$$\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{A},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda}) = \frac{1}{N}\sum_{j=0}^{N-1}\sum_{t=0}^{T-1}\left(\sum_{l=0}^{t}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_{\tau_j,l}\mid\boldsymbol{s}_{\tau_j,l})\right)\gamma^t\left(c_0(\boldsymbol{s}_{\tau_j,t},\boldsymbol{a}_{\tau_j,t})+\sum_{i=1}^{U}\lambda_i c_i(\boldsymbol{s}_{\tau_j,t},\boldsymbol{a}_{\tau_j,t})\right).$$

Thus, we have the following:

$$\begin{aligned}&\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{L}_{\mathrm{A},\omega}(\boldsymbol{\theta},\boldsymbol{\lambda})\right]\\ &= \frac{1}{N}\mathbb{V}\mathrm{ar}\left[\sum_{t=0}^{T-1}\left(\sum_{l=0}^{t}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_{\tau_1,l}\mid\boldsymbol{s}_{\tau_1,l})\right)\gamma^t\left(c_0(\boldsymbol{s}_{\tau_1,t},\boldsymbol{a}_{\tau_1,t})+\sum_{i=1}^{U}\lambda_i c_i(\boldsymbol{s}_{\tau_1,t},\boldsymbol{a}_{\tau_1,t})\right)\right]\\ &= \frac{1}{N}\mathbb{E}\left[\left\|\sum_{t=0}^{T-1}\left(\sum_{l=0}^{t}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_{\tau_1,l}\mid\boldsymbol{s}_{\tau_1,l})\right)\gamma^t\left(c_0(\boldsymbol{s}_{\tau_1,t},\boldsymbol{a}_{\tau_1,t})+\sum_{i=1}^{U}\lambda_i c_i(\boldsymbol{s}_{\tau_1,t},\boldsymbol{a}_{\tau_1,t})\right)\right\|_2^2\right]\\ &\le \frac{1}{N}\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t\left(c_0(\boldsymbol{s}_{\tau_1,t},\boldsymbol{a}_{\tau_1,t})+\sum_{i=1}^{U}\lambda_i c_i(\boldsymbol{s}_{\tau_1,t},\boldsymbol{a}_{\tau_1,t})\right)^2\left(\sum_{t=0}^{T-1}\gamma^t\sum_{l=0}^{t}\left\|\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_{\tau_1,l}\mid\boldsymbol{s}_{\tau_1,l})\right\|_2^2\right)\right]\\ &\le \frac{(1+\|\boldsymbol{\lambda}\|_1)^2(1-\gamma^T)}{N(1-\gamma)}\mathbb{E}\left[\sum_{t=0}^{T-1}\gamma^t\sum_{l=0}^{t}\left\|\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_{\tau_1,l}\mid\boldsymbol{s}_{\tau_1,l})\right\|_2^2\right].\end{aligned}$$

We now recover the result of Lemma E.3 by Montenegro et al. (2024b) stating that, under Assumptions 4.4 and 4.5, it holds that $\mathbb{E}[\|\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\tau)\|_2^2] \le c d_{\mathcal{A}} L_{1\mu}^2 \sigma^{-2}$ for every $\boldsymbol{\theta} \in \Theta$ and trajectory τ. Thus, we conclude the following:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{A},\omega}(\boldsymbol{\theta}, \boldsymbol{\lambda})\right] \le \frac{c d_{\mathcal{A}} L_{1\mu}^2 (1 + \|\boldsymbol{\lambda}\|_1)^2 (1-\gamma^T)^3}{N\sigma^2 (1-\gamma)^3}.$$

□

As previously discussed in Section 3, we highlight that while $V_{\boldsymbol{\lambda}} = \mathcal{O}(1)$, $V_{\dagger,\boldsymbol{\theta}} = \mathcal{O}(\sigma^{-2}\omega^{-2})$ when considering the learning process of C-PG. The dependence of the latter from ω^{-2} arises from the regularization and projection Π_Λ, which ensures that the Lagrange multipliers are bounded by $\mathcal{O}(\omega^{-1})$.

4.4. Convergence Analysis

We are now almost ready to present the convergence of C-PG to the optimal feasible *deterministic* policy obtained by switching off the stochasticity (i.e., setting $\sigma = 0$) at the end of the learning process. Recall that in Theorem 3.2 we analyzed the convergence of the *stochastic* (i.e., associated with $\mathcal{L}_{\dagger,\omega}$) potential function $\mathcal{P}_{\dagger,k}(\chi)$, after establishing in Theorem 3.1 that it serves as a tool to assess the last-iterate global convergence of C-PG. Before proceeding, we formally introduce the *deterministic* (i.e., associated with $\mathcal{L}_{\mathrm{D},\omega}$) potential function. Considering $\omega > 0$, $\chi \in [0,1]$, and to be at a generic $k \in \mathbb{N}$ iterate of C-PG, $\mathcal{P}_{\mathrm{D},k}(\chi)$ i sdefined as:

$$\mathcal{P}_{\mathrm{D},k}(\chi) := \mathbb{E}\left[\max_{\boldsymbol{\lambda}\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega}, \boldsymbol{\lambda}^*_{\mathrm{D},\omega})\right] + \chi\, \mathbb{E}\left[\max_{\boldsymbol{\lambda}\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k)\right],$$

where $(\boldsymbol{\theta}^*_{\mathrm{D},\omega}, \boldsymbol{\lambda}^*_{\mathrm{D},\omega})$ is the saddle point of $\mathcal{L}_{\mathrm{D},\omega}$.

In this part, we first quantify the loss $|\mathcal{P}_{\dagger,k}(\chi) - \mathcal{P}_{\mathrm{D},k}(\chi)|$ between the stochastic and deterministic potential functions at a generic iterate $k \in [\![K]\!]$ of C-PG. We then leverage this result to derive the sample complexity required to ensure that the deterministic policy deployed in the last iterate of C-PG is an optimal feasible solution for the problem at hand.

Theorem 4.5. *Suppose to run* C-PG *for k iterations. Considering (hyper)policies complying with Definitions 4.2 (AB) or 4.3 (PB), under Assumptions 4.1 (PB) or 4.2 (AB), for any $\chi \in [0,1]$, the following result holds:*

$$\left|\mathcal{P}_{\dagger,k}(\chi) - \mathcal{P}_{D,k}(\chi)\right| \le 4(1+\Lambda_{\max}) L_{1\dagger} \sigma \sqrt{d_\dagger},$$

where $\Lambda_{\max} := \omega^{-1} U J_{\max}$, $L_{1P} := L_{1D,\max}$, $L_{1A} := L_{1\underline{\boldsymbol{\mu}},\max}$, $d_P := d_\Theta$, and $d_A := d_{\mathcal{A}}$.

Proof. Considering to be at iteration k of C-PG, we aim to find an upper bound to the quantity $|\mathcal{P}_{\dagger,k}(\chi) - \mathcal{P}_{\mathrm{D},k}(\chi)|$. Let $(\boldsymbol{\theta}^*_{\mathrm{D},\omega}, \boldsymbol{\lambda}^*_{\mathrm{D},\omega})$ be a saddle point of $\mathcal{L}_{\mathrm{D},\omega}$ and let $(\boldsymbol{\theta}^*_{\dagger,\omega}, \boldsymbol{\lambda}^*_{\dagger,\omega})$ be a saddle point of $\mathcal{L}_{\dagger,\omega}$. Moreover, since we are considering the learning process of C-PG, the norm of the Lagrange multipliers is bounded by $\|\boldsymbol{\lambda}\|_1 \le \sqrt{U}\|\boldsymbol{\lambda}\|_2 \le \omega^{-1} U J_{\max} =: \Lambda_{\max}$ due to the regularization. The first thing we do is to quantify the quantity $|\mathcal{P}_{\dagger,k}(\chi) - \mathcal{P}_{\mathrm{D},k}(\chi)|$:

$$\begin{aligned}
&\left|\mathcal{P}_{\dagger,k}(\chi) - \mathcal{P}_{\mathrm{D},k}(\chi)\right| \\
&= \Bigg|\mathbb{E}\left[\max_{\boldsymbol{\lambda}\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}^*_{\dagger,\omega}, \boldsymbol{\lambda}^*_{\dagger,\omega})\right] + \chi\, \mathbb{E}\left[\max_{\boldsymbol{\lambda}\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k)\right] \\
&\quad - \mathbb{E}\left[\max_{\boldsymbol{\lambda}\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega}, \boldsymbol{\lambda}^*_{\mathrm{D},\omega})\right] - \chi\, \mathbb{E}\left[\max_{\boldsymbol{\lambda}\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}) - \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k)\right]\Bigg|.
\end{aligned}$$

For the sake of readability, we introduce the following quantities:

$$\begin{aligned}
\mathsf{A} &:= \max_{\boldsymbol{\lambda}\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}) - \max_{\boldsymbol{\lambda}\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}), \\
\mathsf{B} &:= \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}^*_{\mathrm{D},\omega}, \boldsymbol{\lambda}^*_{\mathrm{D},\omega}) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}^*_{\dagger,\omega}, \boldsymbol{\lambda}^*_{\dagger,\omega}), \\
\mathsf{C} &:= \mathcal{L}_{\mathrm{D},\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k) - \mathcal{L}_{\dagger,\omega}(\boldsymbol{\theta}_k, \boldsymbol{\lambda}_k).
\end{aligned}$$

The quantity $|\mathcal{P}_{\dagger,k}(\chi) - \mathcal{P}_{\mathrm{D},k}(\chi)|$ we aim to bound, can be expressed as:

$$\left|\mathcal{P}_{\dagger,k}(\chi) - \mathcal{P}_{\mathrm{D},k}(\chi)\right| = |(1+\chi)\,\mathbb{E}[\mathsf{A}] + \mathbb{E}[\mathsf{B}] + \chi\,\mathbb{E}[\mathsf{C}]|\,.$$

By simply applying the triangular and Jensen's inequalities, it holds that:

$$\left|\mathcal{P}_{\dagger,k}(\chi) - \mathcal{P}_{\mathrm{D},k}(\chi)\right| \le (1+\chi)\,\mathbb{E}[|\mathsf{A}|] + \mathbb{E}[|\mathsf{B}|] + \chi\,\mathbb{E}[|\mathsf{C}|]\,.$$

Thus, we can now focus on bounding the terms A, B, and C. Starting from A:

$$\begin{aligned}
|\mathsf{A}| &= \left|\max_{\lambda\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\dagger,\omega}(\theta_k,\lambda) - \max_{\lambda\in\mathbb{R}^U_{\ge 0}} \mathcal{L}_{\mathrm{D},\omega}(\theta_k,\lambda)\right| \\
&\le \left|\max_{\lambda\in\mathbb{R}^U_{\ge 0}} \left(\mathcal{L}_{\dagger,\omega}(\theta_k,\lambda) - \mathcal{L}_{\mathrm{D},\omega}(\theta_k,\lambda)\right)\right| \\
&\le (1+\Lambda_{\max})L_{1\dagger}\sigma\sqrt{d_\dagger},
\end{aligned}$$

where in the last step we applied the result from Theorem 4.1. We can now focus on the term B:

$$\begin{aligned}
|\mathsf{B}| &= \left|\mathcal{L}_{\mathrm{D},\omega}(\theta^*_{\mathrm{D},\omega},\lambda^*_{\mathrm{D},\omega}) - \mathcal{L}_{\dagger,\omega}(\theta^*_{\dagger,\omega},\lambda^*_{\dagger,\omega})\right| \\
&\le \left|\mathcal{L}_{\mathrm{D},\omega}(\theta^*_{\dagger,\omega},\lambda^*_{\mathrm{D},\omega}) - \mathcal{L}_{\dagger,\omega}(\theta^*_{\dagger,\omega},\lambda^*_{\mathrm{D},\omega})\right| \\
&\le (1+\Lambda_{\max})L_{1\dagger}\sigma\sqrt{d_\dagger},
\end{aligned}$$

where we first exploited the properties of saddle points, then we leveraged Theorem 4.1. Finally, we can focus on the term C:

$$|\mathsf{C}| = \left|\mathcal{L}_{\mathrm{D},\omega}(\theta_k,\lambda_k) - \mathcal{L}_{\dagger,\omega}(\theta_k,\lambda_k)\right| \le (1+\Lambda_{\max})L_{1\dagger}\sigma\sqrt{d_\dagger},$$

again by applying Theorem 4.1.

Putting all together, and bounding $\chi \le 1$, we conclude the following:

$$\left|\mathcal{P}_{\dagger,k}(\chi) - \mathcal{P}_{\mathrm{D},k}(\chi)\right| \le 4(1+\Lambda_{\max})L_{1\dagger}\sigma\sqrt{d_\dagger}.$$

□

Notice that the result from Theorem 4.5 has exactly the same form as the one from Theorem 4.1, except for the fact that here, since we are dealing exclusively with regularized Lagrangian functions, we directly exploited the fact that the Lagrange multipliers have a norm bounded by $\Lambda_{\max}$.

Theorem 4.6 (Sample Complexity for Deterministic Deployment)**.** *Suppose to run* `C-PG` *for K iterations employing a (hyper)policy complying with Definitions 4.2 (AB) or 4.3 (PB). Suppose to be under Assumptions 4.1 (PB) or 4.2 (AB), 4.3, 4.4 (AB), and 4.5. For $\psi \in [1,2]$, $\chi < 1/5$, sufficiently small ϵ and ω, and a choice of* constant *learning rates $\zeta_\lambda = \mathcal{O}(\omega\sigma^2\epsilon^{2/\psi})$ and $\zeta_\theta = \omega\zeta_\lambda$, whenever $K = \mathcal{O}(\omega^{-3}\sigma^{-2}\epsilon^{-\frac{4}{\psi}+1})$ and the gradients are estimated, we have that:*

$$\mathcal{P}_{D,K}(\chi) \le \epsilon + \frac{\beta_\dagger(\sigma,\psi)}{\alpha_D} + 4(1+\Lambda_{\max})L_{1\dagger}\sigma\sqrt{d_\dagger},$$

where $\beta_\dagger(\sigma,\psi)$ is quantified in Theorem 4.2, $\Lambda_{\max} := \omega^{-1}UJ_{\max}$, $L_{1P} := L_{1D,\max}$, $L_{1A} := L_{1\underline{\mu},\max}$, $d_P := d_\Theta$, and $d_A := d_{\mathcal{A}}$.

	Deterministic Deployment with Estimated Gradients		
	ψ=1 (GD)	$\psi\in$(1,2)	ψ=2 (PL)
Fixed ω and σ	$\omega^{-3}\sigma^{-2}\epsilon^{-3}\log(\epsilon^{-1})$	$\omega^{-3}\sigma^{-2}\epsilon^{1-4/\psi}\log(\epsilon^{-1})$	$\omega^{-3}\sigma^{-2}\epsilon^{-1}\log(\epsilon^{-1})$
ω=$\mathcal{O}(\epsilon)$ and ϵ=$\mathcal{O}(\epsilon)$	$\epsilon^{-8}\log(\epsilon^{-1})$	$\epsilon^{-4-4/\psi}\log(\epsilon^{-1})$	$\epsilon^{-6}\log(\epsilon^{-1})$

Table 2
Summary of the sample complexity results for the deterministic deployment C-PG when considering estimated gradients.

Proof Sketch. Under the considered set of assumptions, we recover the results of Theorems 4.2, 4.3, and 4.5 and of Lemma 4.4, matching the conditions needed to establish the sample complexity exhibited in Theorem 3.2 for ensuring that $\mathcal{P}_{\dagger,K}(\chi) \le \epsilon + \frac{\beta_\dagger(\sigma.\psi)}{\alpha_{\mathrm{D}}}$, where we employed the coefficients of the inherited weak ψ-GD (Theorem 4.2).

In particular, recovering "Part V: Rates Computation" of the proof of Theorem 3.2 with $\psi \in [1,2]$ and inexact gradients, considering $L_1 = \mathcal{O}(1)$ and $L_2 = \mathcal{O}(\omega^{-1})$ (Theorem 4.3), while $V_{\dagger,\theta} = \mathcal{O}(\omega^{-2}\sigma^{-2})$ and $V_\lambda = \mathcal{O}(1)$ (Lemma 4.4), and selecting $\zeta_\lambda = \mathcal{O}(\omega\sigma^2\epsilon^{2/\psi})$ and $\zeta_\theta = \mathcal{O}(\omega^3\sigma^2\epsilon^{2/\psi})$, we have that, with a sample complexity of order:

$$K \le \mathcal{O}\left(\frac{\log\frac{1}{\epsilon}}{\omega^3\sigma^2\epsilon^{-1+4/\psi}}\right),$$

it is guaranteed that $\mathcal{P}_{\dagger,K}(\chi) \le \epsilon + \frac{\beta_\dagger(\sigma,\psi)}{\alpha_{\mathrm{D}}}$. Now, exploiting the result of Theorem 4.5, we have that the same sample complexity ensures that:

$$\mathcal{P}_{\mathrm{D},K}(\chi) \le \epsilon + \frac{\beta_\dagger(\sigma,\psi)}{\alpha_{\mathrm{D}}} + 4(1+\Lambda_{\max})L_{1\dagger}\sigma\sqrt{d_\dagger}.$$

□

We highlight that all the remarks made for Theorem 3.2 also apply in this case. Moreover, the sample complexity achieved here matches that of Theorem 3.2, with the exception of an additional σ^{-2} factor. This term arises from the characterization of the constant $V_{\dagger,\theta} = \mathcal{O}(\omega^{-2}\sigma^{-2})$ in Lemma 4.4, due to the specific noise model employed to define the AB and PB exploration paradigms.

We also note that Theorem 3.1 remains applicable in this setting, as it holds for general potential functions $\mathcal{P}_{\dagger,k}(\chi)$, including the deterministic one $\mathcal{P}_{\mathrm{D},k}(\chi)$. As discussed in Section 3, and consistently with the analysis without deterministic deployment, Theorem 3.1 suggests choosing $\omega = \mathcal{O}(\epsilon)$. In the current context, Theorem 4.6 further suggests setting $\sigma = \mathcal{O}(\epsilon)$ in order to ensure that $\mathcal{P}_{\mathrm{D},K}(\chi) \le \mathcal{O}(\epsilon) + c$, where c is a constant.

Therefore, by letting both ω and σ scale as $\mathcal{O}(\epsilon)$, the resulting sample complexity becomes $K = \mathcal{O}(\epsilon^{4-4/\psi}\log\epsilon^{-1})$, as summarized in Table 2. We emphasize that this result aligns with the known sample complexity bounds for policy gradient methods under the same noise model in the unconstrained setting (Montenegro et al., 2024a).

Nonetheless, we remark that the theoretical recommendation of setting $\sigma = \mathcal{O}(\epsilon)$ may be impractical in real applications. In practice, using such a low level of stochasticity may lead to slower convergence, even though it results in a more accurate deterministic deployment (Montenegro, Mussi, Metelli and Papini, 2025).

Finally, by retrieving the specific quantities of the AB and PB exploration paradigms, we recover the known trade-off between them (Metelli et al., 2018): AB may suffer from long interaction horizons or high-dimensional action spaces (large T or $d_{\mathcal{A}}$), whereas PB may suffer high-dimensional parameter spaces (large d_Θ).

5. Related Work

In this section, we review related work primarily focusing on policy optimization methods for CRL, convergence guarantees of primal-dual approaches, and the learning of deterministic policies in constrained environments.

Policy Optimization Approaches for Constrained Reinforcement Learning. Policy Optimization based algorithms for Constrained Reinforcement Learning mostly follow *primal*-only or *primal-dual* approaches. *Primal*-only algorithms (Dalal, Dvijotham, Vecerik, Hester, Paduraru and Tassa, 2018; Chow, Nachum, Duéñez-Guzmán and Ghavamzadeh, 2018; Yu, Yang, Kolar and Wang, 2019; Liu, Ding and Liu, 2020; Xu, Liang and Lan, 2021) avoid considering dual variables by focusing on the design of the objective function and by designing the update rules for the policy at hand incorporating the constraint satisfaction part.

The main benefit of employing primal-only algorithms lies in the fact that there is no need to consider another variable to learn, and therefore, no need to tune its learning rate. However, few of the existing methods establish global convergence to an optimal feasible solution. For instance, Xu et al. (2021) propose CRPO, an algorithm employing an *unconstrained* policy maximization update taking into account the reward when all the constraints are satisfied, while leveraging on-policy minimization updates in the direction of violated constraint functions. Moreover, it exhibits average global convergence guarantees for the tabular setting. On the other hand, *primal-dual* algorithms (Chow et al., 2017; Achiam et al., 2017; Tessler et al., 2019; Stooke et al., 2020; Ding et al., 2020, 2021; Bai et al., 2022; Ying et al., 2022; Bai et al., 2023; Gladin et al., 2023; Ding et al., 2024) are the most commonly used and investigated. Indeed, the effectiveness of using the primal-dual approach is justified by Paternain et al. (2019), which states that this kind of approach has zero duality gap under Slater's condition when optimizing over the space of all the possible stochastic policies. Among the reported works, Stooke et al. (2020) propose PID Lagrangian, a method to update the dual variable, smoothing the oscillations around the threshold value of the costs during the learning. The practical strength of such a method is that it can be paired with any of the existing policy optimization methods. The other cited works are treated in detail in the next paragraph.

Lagrangian-based Policy Search Convergence Guarantees. A lot of research effort has been spent on studying the convergence guarantees of primal-dual policy optimization methods. In this field, the goal is to ensure last-iterate convergence guarantees with rates that are dimension-free, i.e., independent of the state and action spaces' dimensions, and work with multiple constraints. In the rest of this paragraph, we talk about single time-scale algorithms when the methods at hand prescribe the usage of the same step sizes for both the primal and dual variables' updates. Vázquez-Abad, Krishnamurthy, Martin and Baltcheva (2002) and Bhatnagar and Lakshmanan (2012) propose primal-dual policy gradient-based methods built upon distinct time scales and relying on nested loops. Such methods only show *asymptotic* convergence guarantees. Chow et al. (2017) propose two primal-dual methods ensuring *asymptotic* convergence guarantees. The peculiarity of those methods lies in the fact that their notion of CMDP encapsulates risk-based constraints, introducing an additional learning variable. Their algorithms have guarantees of *asymptotic* convergence to stationary points. The recent works by Zheng, You and Mallada (2022) and Moskovitz, O'Donoghue, Veeriah, Flennerhag, Singh and Zahavy (2023) also propose methods ensuring *asymptotic* global convergence guarantees. These methods exploit occupancy-measure iterates rather than policy iterates. Ding et al. (2020) propose NPG-PD, which relies on a natural policy gradient approach and, under Slater's assumption, ensures dimension-dependent *average*-iterate global convergence guarantees in the single-constrained setting with a single time-scale and with exact gradients. This work has been extended by Ding et al. (2022), striking dimension-free rates, but still just guaranteeing *average*-iterate convergence with exact gradients. However, sample-based versions of NPG-PD achieving, under additional assumptions, the same convergence rates are provided by the authors. Another work ensuring an *average*-iterate rate is the one by Liu et al. (2021). The latter exhibits a convergence rate of order $\widetilde{\mathcal{O}}(\epsilon^{-1})$, considering to act in tabular CMDPs with softmax policies and having access to exact gradients and to a generative model. Liu et al. (2021) propose also a sample-based version of their algorithm, keeping the same setting previously described, which ensures a convergence rate on *average* of order $\widetilde{\mathcal{O}}(\epsilon^{-3})$. Both Ying et al. (2022) and Gladin et al. (2023) propose algorithms involving regularization. The proposed methods rely on natural policy-based subroutines and show dimension-dependent *last-iterate* global convergence guarantees, relying on two time-scales. These methods work also with multiple constraints. Additionally, Ding et al. (2024) propose RPG-PD and OPG-PD, exhibiting *last-iterate* global convergence guarantees under Slater's condition in a single-constraint setting. The former is a regularized version of the algorithm proposed by Ding et al. (2020), showing *last-iterate* global convergence at a sublinear rate. The latter leverages on the optimistic gradient method (Hsieh, Iutzeler, Malick and Mertikopoulos, 2019) to unlock a faster linear convergence rate. These methods show single time-scale dimension-dependent rates and both leverage on exact gradients. However, for RPG-PD there exists an *inexact* version showing, under additional assumptions on the statistical and transfer errors and the relative condition number (Ding et al., 2024, Assumption 2), the same guarantees of the exact one. Finally, Mondal and Aggarwal (2024) introduce PDR-ANPG a primal dual-based regularized accelerated natural policy gradient algorithm

Algorithm	Dimension-free	Setting	Exploration Type	Single time-scale	Gradients	Assumptions	Sample Complexity	Iteration Complexity
Dual Descent (Ying et al., 2022)	✗	$U \geq 1$ $T = \infty$ Softmax param.	AB	✗	Inexact	Slater Sufficient Exploration	$\mathcal{O}\left(\epsilon^{-2}\log^2\epsilon^{-1}\right)$	$\mathcal{O}\left(\log^2\epsilon^{-1}\right)$
Cutting-Plane (Gladin et al., 2023)	✗	$U \geq 1$ $T = \infty$ Softmax param.	AB	✗	Inexact	Slater Uniform Ergodicity Oracle	$\mathcal{O}\left(\epsilon^{-4}\log^3\epsilon^{-1}\right)$	$\mathcal{O}\left(\log^3\epsilon^{-1}\right)$
Exact RPG-PD (Ding et al., 2024)	✗	$U = 1$ $T = \infty$ Softmax param.	AB	✓	Exact	Slater	-	$\mathcal{O}\left(\epsilon^{-6}\log^2\epsilon^{-1}\right)$
Inexact RPG-PD (Ding et al., 2024)	✗	$U = 1$ $T = \infty$ Softmax param.	AB	✓	Inexact	Slater Stat. Err. Bounded Transf. Err. Bounded Cond. Num. $< +\infty$	$\mathcal{O}\left(\epsilon^{-6}\log^2\epsilon^{-1}\right)$	$\mathcal{O}\left(\epsilon^{-6}\log^2\epsilon^{-1}\right)$
OPG-PD (Ding et al., 2024)	✗	$U = 1$ $T = \infty$ Softmax param.	AB	✓	Exact	Slater	-	$\mathcal{O}\left(\log^2\epsilon^{-1}\right)$
PDR-ANPG (Mondal and Aggarwal, 2024)	✗	$U = 1$ $T = \infty$ General param.	AB	✗	Exact	Slater NPG Oracle $\epsilon_{\text{bias}} > 0$ FIM Positive Definite	$\mathcal{O}(\epsilon^{-2}\min\{\epsilon^{-2}, (\epsilon_{\text{bias}})^{-1/3}\}\log^2\epsilon^{-1})$	-
C-PG (This work)	✓	$U \geq 1$ $T \in \mathbb{N} \cup \{\infty\}$ General param.	AB and PB	✗	Exact Inexact	Asm. 3.1, 3.2, 3.3, 3.4	Table 1	Table 1
Lower Bound (Vaswani et al., 2022)	✗	$U = 1$ $T = \infty$	-	-	Inexact	Slater	$\Omega\left(\epsilon^{-2}\right)$	-

Table 3
Comparison among primal-dual methods ensuring last-iterate global convergence guarantees.

that utilizes entropy and quadratic regularizers in the Lagrangian function and a natural policy gradient (NPG) oracle as a subroutine. This method, which is not single time scale, operates in the setting of CMDPs with $|\mathcal{S}| = +\infty$, $|\mathcal{A}| < +\infty$, and $U = 1$. Under the Slater condition, considering a fisher information matrix (FIM) to be positive definite, and considering a general policy parameterization with a policy-class error $\epsilon_{\text{bias}} > 0$, PDR-ANPG exhibits a sample complexity of order $\mathcal{O}(\epsilon^{-2}\min\{\epsilon^{-2}, (\epsilon_{\text{bias}}^{-1/3})\}\log^2\epsilon^{-1})$. It is worth noticing that all the mentioned works just consider the *action-based* exploration approach for policy optimization, while the *parameter-based* one remains unexplored. For the sake of clarity, Table 3 shows a detailed comparison between our approach and the other presented methods exhibiting *last-iterate* global convergence guarantees.

Furthermore, Vaswani, Yang and Szepesvári (2022) have recently proposed a dimension-dependent lower bound for the sample complexity of $\Omega\left(\epsilon^{-2}\right)$, assuming to be under the Slater condition and considering single-constrained CMDPs with finite state and action spaces.

Learning Deterministic Policies in CMDPs. While policy gradient methods have been extensively studied in the context of CRL, most existing approaches focus exclusively on stochastic policies, whereas the study of deterministic policies for CMDPs has received comparatively little attention. Deterministic policies, however, are crucial for real-world applications where reliability, safety, and predictability are essential. Despite this, very few works have tackled the problem of learning deterministic policies in CRL settings, particularly in continuous-state and continuous-action CMDPs. A recent contribution in this direction is presented by Rozada et al. (2025), who introduce the *Deterministic Policy Gradient Primal-Dual* (D-PGPD) algorithm, a novel method designed to directly learn deterministic policies in CMDPs. The proposed approach leverages an *entropy-regularized Lagrangian* formulation, where the primal update performs a proximal-point-type ascent step solving a quadratic-regularized maximization sub-problem, while the dual update performs a gradient descent step solving a quadratic-regularized minimization sub-problem. The theoretical contribution of this work is the proof that the exact version of D-PGPD achieves asymptotically an ϵ-optimal solution in $\mathcal{O}(\epsilon^{-6})$ iterations, making it one of the first primal-dual methods that directly optimize deterministic policies in CRL. Moreover, an approximated version of D-PGPD (namely AD-PGPD), incorporating function approximation, achieves the same convergence rate under the assumption that the approximation error satisfies $\epsilon_{\text{approx}} = \mathcal{O}(\epsilon^4)$. These rates hold under the assumption that the model of the environment is known. A key limitation of the proposed approach is that it considers only a single constraint, restricting its applicability to more complex multi-constrained CMDPs commonly found in real-world applications. Additionally, unlike stochastic policy-based CRL approaches, where exploration is driven by the inherent randomness of the policy, D-PGPD learns a deterministic policy directly, relying entirely on the environment to provide the required exploration. While this design ensures stable and consistent policy execution, it also presents a major drawback: the lack of explicit exploration mechanisms significantly limits the applicability of D-PGPD in practice, as it may struggle in environments where intrinsic randomness is insufficient to ensure adequate state-action space coverage. To extend D-PGPD to a model-free setting, the authors propose a sample-based version of AD-PGPD, which leverages rollouts for policy evaluation. However, this approach introduces a significant complexity increase: the model-free algorithm requires $\mathcal{O}(\epsilon^{-18})$ rollouts to compute an ϵ-optimal policy, making it substantially less practical for large-scale real-world tasks.

6. Numerical Validation

In this section, we empirically validate the theoretical results established throughout the paper. Further experimental details are reported in Appendix D.[7]

Before proceeding, we clarify that C-PGPE and C-PGAE refer to the PB and AB variants of C-PG, respectively. We denote by DC-PGPE and DC-PGAE their corresponding versions with deterministic deployment, obtained by switching off the stochasticity (i.e., setting $\sigma = 0$). Accordingly, the curves shown for DC-PGPE and DC-PGAE correspond to the performances and costs of the deterministic policies encountered during training by C-PGPE and C-PGAE, respectively.

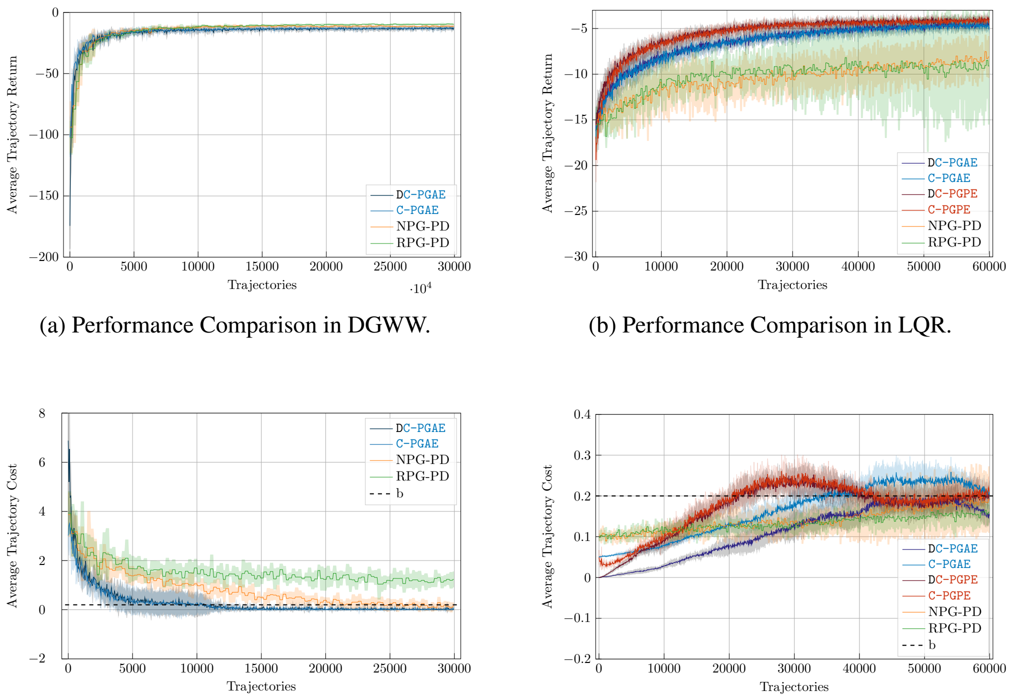
The section is organized as follows. Section 6.1 presents comparisons between the proposed algorithms and state-of-the-art baselines, Section 6.2 investigates the effects of deterministic deployment on policies learned via C-PG methods, and Section 6.3 provides a sensitivity analysis on the impact of the regularization parameter ω.

6.1. Comparison Against Baselines

Comparison in DGWW. We compare our proposal C-PGAE against the sample-based versions of NPG-PD (Ding et al., 2020, Appendix H) and RPG-PD (Ding et al., 2024, Appendix C.9). The environment in which the methods are tested is the Discrete Grid World with Walls (DGWW, see Appendix D) with a horizon of $T = 100$. In this experiment, all the methods aim to learn the parameters of a tabular softmax policy with 196 parameters, maximizing the trajectory reward while considering a single constraint on the average trajectory cost, for which we set a threshold $b = 0.2$. All the methods were run for $K = 3000$ iterations with a batch size of $N = 10$ trajectories per iteration, and with constant learning rates. In particular, for both C-PGAE and NPG-PD, we employed $\zeta_{\theta} = 0.01$ and $\zeta_{\lambda} = 0.1$, while for RPG-PD we selected $\zeta_{\theta} = 0.01$ and $\zeta_{\lambda} = 0.01$. For C-PGAE and RPG-PD, we used a regularization constant $\omega = 10^{-4}$. All the details about the experimental setting are summarized in Table 5. We would like to stress that, as prescribed by the respective convergence theorems, we chose a two-timescale learning rate approach for C-PGAE and a single-timescale one for RPG-PD. Figures 2a and 2c show the performance curves (i.e., the one associated with the objective function and the one for the costs). As can be noticed, C-PGAE manages to strike the objective of the constrained optimization problem with less trajectories. Indeed, the sample-based NPG-PD requires to estimate the value and the action-value functions for all the states and state-action pairs, resulting in analyzing $|\mathcal{S}|+|\mathcal{S}||\mathcal{A}|$ additional trajectories w.r.t. C-PGAE for every iteration of the algorithm. The sample-based RPG-PD also requires additional trajectories to be analyzed, which, in practice, for a correct learning behavior, result to be the same in number as the extra ones analyzed by NPG-PD. In this environment, DC-PGAE exhibits almost the same behavior of C-PGAE, thus meaning that the encountered stochastic policies do not meet a significant loss in performances and costs when switching off the stochasticity.

Comparison in LQR. We compare our proposals C-PGAE and C-PGPE against the continuous sample-based version of NPG-PD (Ding et al., 2022, Algorithm 1) with works with generic policy parameterizations. In the following, we refer to this version of NPG-PD as NPG-PD2. Moreover, we added a ridge-regularized version of NPG-PD2, which we call RPG-PD2, to resemble the type of regularization we employed for our proposed methods. For all the regularized methods (i.e., C-PGAE, C-PGPE, and RPG-PD2) we selected as regularization constant $\omega = 10^{-4}$. The setting for this experiment considers a bidimensional LQR environment with a single cost over the provided actions (see Appendix D) and with a fixed horizon $T = 50$. Here, the methods aim at maximizing the average reward over trajectories, while keeping the average cost over trajectories under the threshold $b = 0.9$. In particular, C-PGAE learns the parameters of a linear gaussian policy with a variance $\sigma^2_{\mathrm{A}} = 10^{-3}$ and employing a learning rate schedule governed by the Adam scheduler (Kingma and Ba, 2015) with $\zeta_{\theta,0} = 0.001$ and $\zeta_{\lambda,0} = 0.01$. C-PGPE learns the parameters of a Gaussian

[7] The code to reproduce the experiments is available at `https://github.com/MontenegroAlessandro/MagicRL/tree/constraints`.



(a) Performance Comparison in DGWW.

(b) Performance Comparison in LQR.

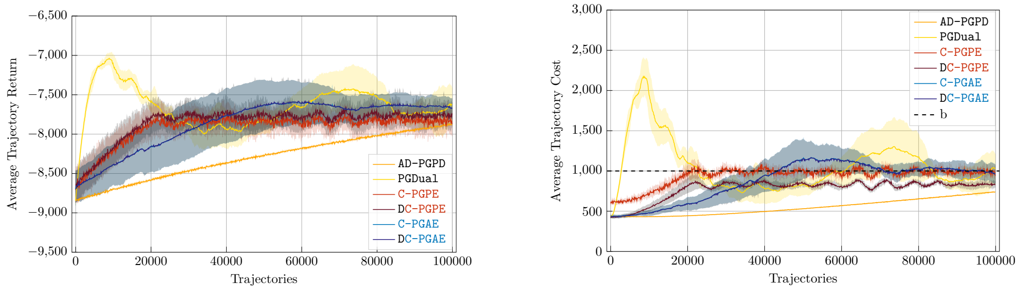
(c) Cost Comparison in DGWW.

(d) Cost Comparison in LQR.

Figure 2: Average return and cost curves in the *CostLQR* and *DGWW* environments (5 runs, mean ± 95% C.I.).

hyperpolicy, with a variance $\sigma_{\mathrm{P}}^2 = 10^{-3}$, which samples the parameters of a deterministic linear policy. It employs a learning rate schedule also governed by Adam with $\zeta_{\rho,0} = 0.001$ and $\zeta_{\lambda,0} = 0.01$. Both C-PGAE and C-PGPE were run for $K = 6000$ iterations with a batch of $N = 100$ trajectories per iteration. NPG-PD2 and RPG-PD2 are both actor-critic methods which were run for $K = 1000$ iterations with a batch size of $N = 600$ trajectories per iteration. In particular, among the trajectories of the reported batch size, $N_1 = 500$ was used for the inner critic-loop, while $N_2 = 100$ for performance and cost estimations. The inner loop step size was selected as a constant, as prescribed by the original algorithm, and with a value $\alpha = 10^{-5}$. Furthermore, since such methods were designed for infinite-horizon discounted environments, we tested them on the same LQR as for C-PGAE and C-PGPE, but leaving $T = 1000$ and $\gamma = 0.98$ (the effective horizon is $(1-\gamma)^{-1} = 50$). The step sizes for the primal and dual variables updates were governed by Adam with $\zeta_{\theta,0} = 0.003$ and $\zeta_{\lambda,0} = 0.01$. As for C-PGAE, both NPG-PD2 and RPG-PD2 aimed at learning the parameters of a linear Gaussian policy, with variance $\sigma_{\mathrm{A}}^2 = 10^{-3}$. All the details about this experiment are summarized in Table 6. Figures 2b and 2d report the learning curves for the average return and the cost over trajectories. As can be seen, our methods manage to solve the constrained optimization problem at hand by leveraging on less trajectories. Indeed, NPG-PD2 and RPG-PD2 suffer the inner critic loop, which adds additional trajectories to be analyzed per iteration (in this specific case $N_1 = 500$). We stress that the actor-critic methods were very sensible to the hyperparameter selection, especially to the length and the step size of the inner loop. When considering DC-PGPE and DC-PGAE, i.e., the deterministic policy curves associated with C-PGPE and C-PGAE respectively, we observe that their overall behavior remains comparable. However, a notable difference arises in the cost curve under AB exploration, where a significant reduction in the incurred costs leads the resulting deterministic policies to consistently satisfy the cost constraint.

Comparison in RobotWorld. We evaluate C-PGAE and C-PGPE against the sample-based versions of AD-PGPD (Rozada et al., 2025) and PGDual (Zhao and You, 2021; Brunke, Greeff, Hall, Yuan, Zhou, Panerati and Schoellig, 2022) in the *RobotWorld* environment (see Appendix D), which is a modification of the *CostLQR* one with quadratic reward and cost functions and where the agents are allowed to control both velocity and acceleration. The C-PG algorithms operate with a finite horizon of $T = 100$ and a discount factor of $\gamma = 1$, while AD-PGPD and PGDual use an infinite horizon $T = 1000$ with $\gamma = 0.99$. Also in this case, the aim is to maximize the performance



(a) Performance Comparison.

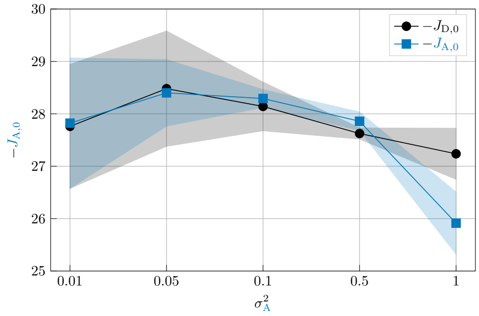
(b) Cost Comparison.

Figure 3: Average return and cost curves in *RobotWorld* (5 runs, mean ± 95% C.I.).

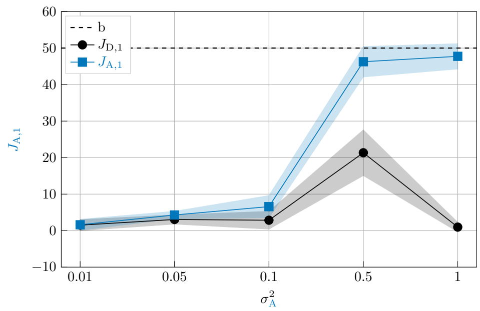
function while keeping the unique cost function under the threshold $b = 1000$. We highlight that this experiment is the same presented in (Rozada et al., 2025) to which we added our methods. That being said, both AD-PGPD and PGDual as prescribed in (Rozada et al., 2025). C-PGPE employs a linear gaussian hyperpolicy with a variance $\sigma_{\mathrm{P}} = 10^{-6}$, collecting a batch of $N = 100$ trajectories per iteration, using a regularization parameter $\omega = 10^{-4}$, and learning rate schedules governed by Adam (Kingma and Ba, 2015) with initial values $\zeta_{\theta,0} = 5 \cdot 10^{-6}$ and $\zeta_{\lambda,0} = 5 \cdot 10^{-3}$. C-PGAE employs a linear gaussian stochastic policy with a variance $\sigma_{\mathrm{A}} = 5 \cdot 10^{-2}$, collecting a batch of $N = 100$ trajectories per iteration, using a regularization parameter $\omega = 10^{-4}$, and learning rate schedules governed by Adam (Kingma and Ba, 2015) with initial values $\zeta_{\theta,0} = 5 \cdot 10^{-6}$ and $\zeta_{\lambda,0} = 10^{-4}$. Both C-PGPE and C-PGAE were run for $K = 10000$ iterations. All the hyperparameters are further presented in Table 7. Figure 3 presents the learning curves for performance and cost across different algorithms. The results highlight that C-PG-based methods consistently achieve better constraint satisfaction while maintaining competitive performance. Notably, C-PGAE and C-PGPE and their deterministic deployment counterparts DC-PGAE and DC-PGPE show faster convergence compared to AD-PGPD and PGDual, which exhibit significant variance and instability in both performance and cost. Furthermore, deterministic variants of C-PG demonstrate a lower constraint violation than the stochastic counterpart, especially when dealing with PB exploration.

6.2. Deterministic Deployment Study

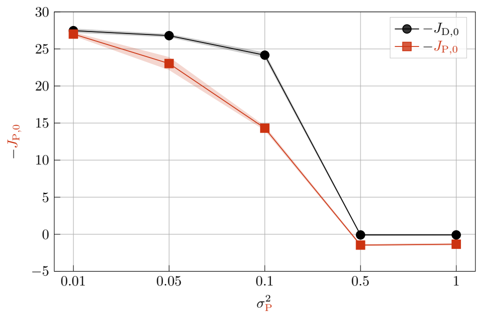
In this experiment, we empirically analyze the deterministic deployment of both C-PGAE and C-PGPE when learning by employing a fixed stochasticity ($\sigma > 0$) and then deploying a deterministic policy switching off the stochasticity ($\sigma = 0$) of the last parameterization encountered while learning. In particular, here we consider the *CostSwimmer-v4* and the *CostHopper-v4* environments (see Appendix D for details) with $T = 100$ and $\gamma = 1$. The employed version for the environments resemble the original one from the MuJoCo control suite (Todorov, Erez and Tassa, 2012), but introducing a cost function representing the energy associated with the control action. In this set of experiments, we study the difference in performance and cost when switching from a stochastic (hyper)policy to a deterministic policy at the end of the learning. In this case, we averaged the last 100 iterates to evaluate the actual deterministic deployment. Additionally, we conducted this deployment loss study for diverse values of stochasticity σ. Both C-PGPE and C-PGAE were run for $K = 3000$ iterations with a batch size of $N = 100$ trajectories collected per iteration. The learning rates were governed by Adam (Kingma and Ba, 2015), the regularization parameter was set to $\omega = 10^{-4}$, and both the methods employed linear Gaussian (hyper)policies with variances $\sigma^2 \in \{0.01, 0.05, 0.1, 0.5, 1\}$. Further details on the setting employed for this set of experiments are presented in Tables 8 and 9. In Figures 4 and 5 it is possible to note that, as the stochasticity parameter σ grows, the distance of $J_{\mathrm{A},1}(\boldsymbol{\theta}_K)$ and $J_{\mathrm{P},1}(\boldsymbol{\theta}_K)$ from $J_{\mathrm{D},1}(\boldsymbol{\theta}_K)$ increases, while the distance between $J_{\mathrm{A},0}(\boldsymbol{\theta}_K)$ and $J_{\mathrm{P},0}(\boldsymbol{\theta}_K)$ from $J_{\mathrm{D},0}(\boldsymbol{\theta}_K)$ shows the same straightforward behavior only in AB exploration—in PB exploration, this is respected only when the learned policy has meaningful performance values. Furthermore, the impact of different exploration paradigms on the learning curves can be observed. In C-PGAE, where noise is injected at each time step, the variance is significantly higher compared to C-PGPE, where noise is sampled only once at the beginning of each trajectory. This distinction results in more stable learning dynamics for C-PGPE. However, C-PGPE also exhibits sensitivity to the magnitude of the injected noise, which can negatively affect its learning capabilities when the noise level is too high. Finally, we highlight that empirically there exists an



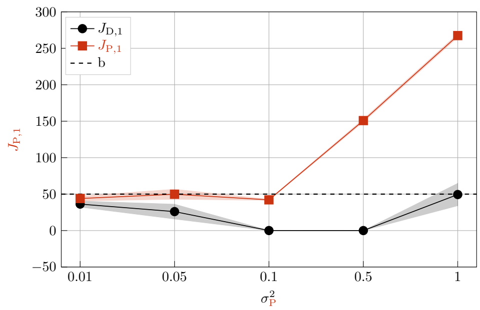
(a) Performance Comparison on C-PGAE.



(b) Cost Comparison on C-PGAE.



(c) Performance Comparison on C-PGPE.



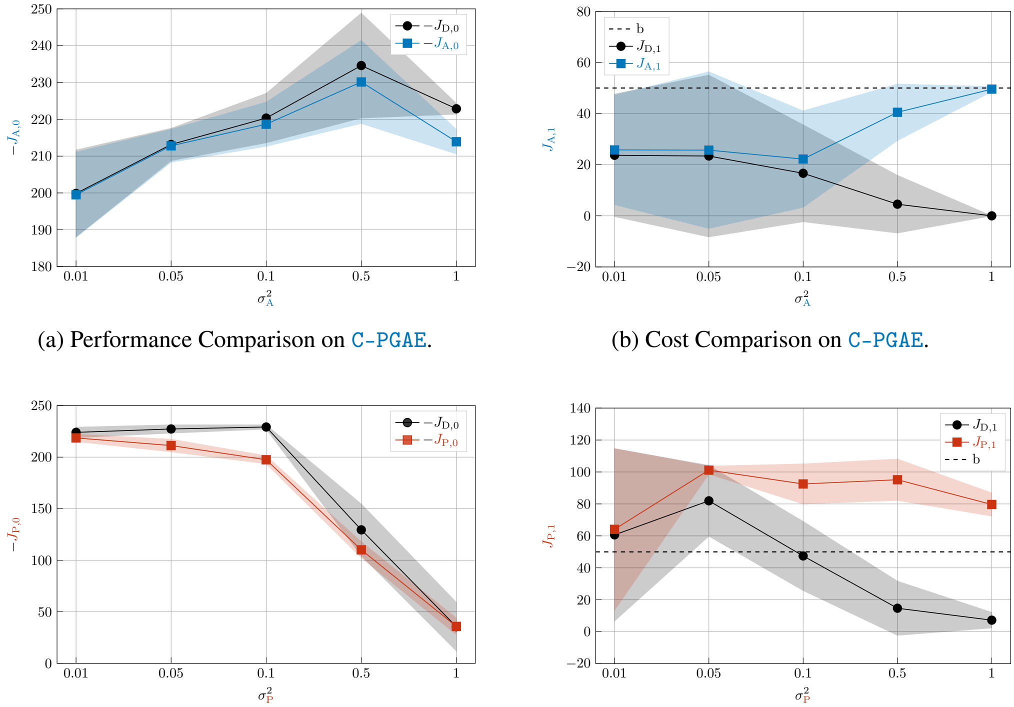
(d) Cost Comparison on C-PGPE.

Figure 4: Deterministic Deployment Study on *CostSwimmer-v4* (5 runs, mean ± 95% C.I.).

"optimal" value for the stochasticity σ leading to a parameterization $\boldsymbol{\theta}_K$ resulting in a deterministic policy maximizing the performance while staying below the cost threshold.

6.3. Regularization Sensitivity Study

In this last experiment, we study the sensitivity of C-PGAE and C-PGPE w.r.t. the regularization term ω. We tested the algorithms on a bidimensional *CostLQR* environment (see Appendix D for details). For the environment at hand, we considered a horizon $T = 50$. We run both algorithms for $K = 10000$ iterations, with a batch size of $N = 100$ trajectories per iteration, and with a varying regularization term such that $\omega \in \{0, 10^{-4}, 10^{-2}\}$. We considered a single constraint on the average trajectory cost, for which we set a threshold $b = 0.2$. For the step size schedules, we employed Adam (Kingma and Ba, 2015) with initial rates $\zeta_{\theta,0} = 10^{-3}$ and $\zeta_{\lambda,0} = 10^{-2}$. Moreover, in this specific experiment, C-PGAE employed a linear gaussian policy with variance $\sigma_A^2 = 10^{-3}$. On the other hand, C-PGPE employed a linear Gaussian hyperpolicy with variance $\sigma_P^2 = 10^{-3}$ over a linear deterministic policy. The experimental setting is summarized in Table 10. Figures 6 and 7 and show the Lagrangian curves, the performance ones (i.e., the one associated with the objective function), and the cost-related ones. From the shown curves it is possible to notice that, for both C-PGAE and C-PGPE, a higher regularization ($\omega = 10^{-2}$) corresponds to a higher bias w.r.t. the constraint satisfaction. This bias is compliant with what is shown by Theorem 3.1. Indeed, the higher the regularization, the stricter the constraint threshold should be made. These considerations also hold for both DC-PGPE and DC-PGAE. In particular, we highlight that, while for PB exploration the stochastic and deterministic curves are almost the same, for AB exploration the deterministic curve related to the cost is always under the cost threshold. Finally, we report in Figure 8 the evolution of the values of the Lagrangian multipliers λ during the learning. As expected from the theory, for both C-PGAE and C-PGPE a higher regularization leads to smaller values of λ. Moreover, we empirically notice that C-PGAE reaches higher values of λ w.r.t. the ones seen by C-PGPE.



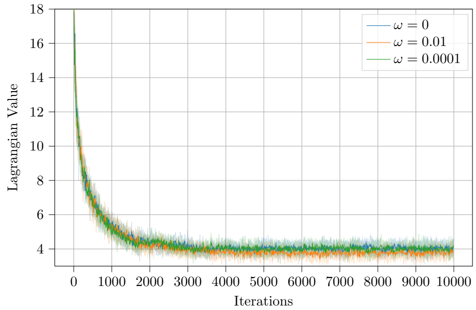
(a) Performance Comparison on C-PGAE. (b) Cost Comparison on C-PGAE.

(c) Performance Comparison on C-PGPE. (d) Cost Comparison on C-PGPE.

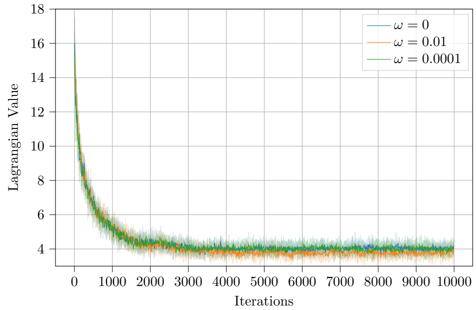
Figure 5: Deterministic Deployment Study on *CostHopper-v4* (5 runs, mean ± 95% C.I.).

7. Conclusion

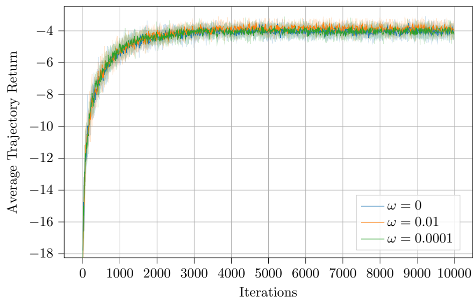
In this work, we proposed a general framework for addressing continuous CRL problems via *primal-dual policy-based* algorithms, employing an alternating ascent–descent scheme. Our *exploration-agnostic* algorithm, C-PG, provides *dimension-free, global, last-iterate* convergence guarantees under the standard weak gradient domination assumption. Furthermore, we reinterpreted both *action-based* and *parameter-based* exploration paradigms as white-noise perturbations applied to parametric deterministic policies, either at the action level or at the parameter one. Under this noise model, we established all the conditions required to ensure last-iterate convergence of C-PG, and we proved that C-PG converges to an optimal deterministic policy when trained via a stochastic (hyper)policy and the stochasticity is subsequently switched off at the end of the learning phase. We validated our theoretical findings by comparing our methods against state-of-the-art baselines and demonstrating their effectiveness, particularly in the deployment of deterministic policies. Future research should aim to improve *sample complexity* of C-PG, with the goal of matching the lower bounds established by Vaswani et al. (2022). Another promising direction is the development of *single time-scale algorithms* that retain the same convergence guarantees. Finally, our analysis assumes a fixed level of stochasticity σ, which must be set on the order of $\mathcal{O}(\epsilon)$ to guarantee convergence to an optimal deterministic policy under the adopted noise model. However, this assumption may be impractical in real-world scenarios, where stochasticity is often either learned or gradually annealed during training. This gap between theory and practice has recently been addressed by Montenegro et al. (2025) in the unconstrained setting; extending this line of work to the constrained case remains an important direction for future research.



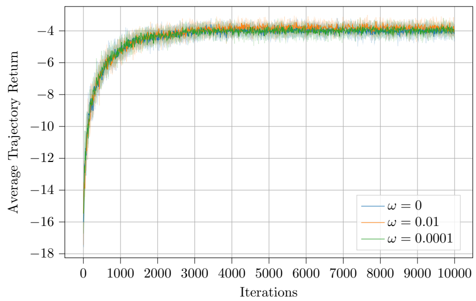
(a) Lagrangian Curves (C-PGPE).



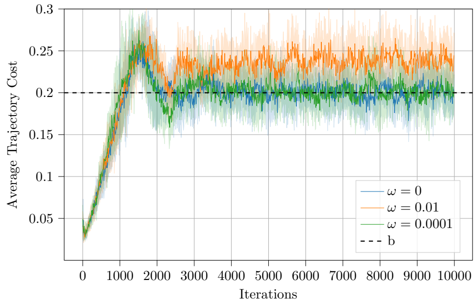
(b) Lagrangian Curves (DC-PGPE).



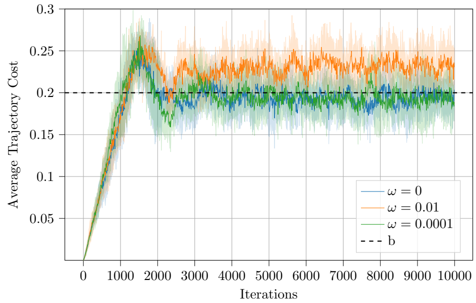
(c) Performance Curves (C-PGPE).



(d) Performance Curves (DC-PGPE).

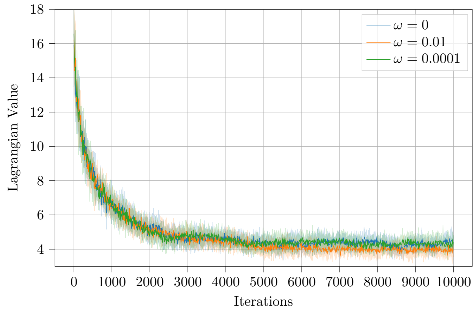


(e) Cost Curves (C-PGPE).

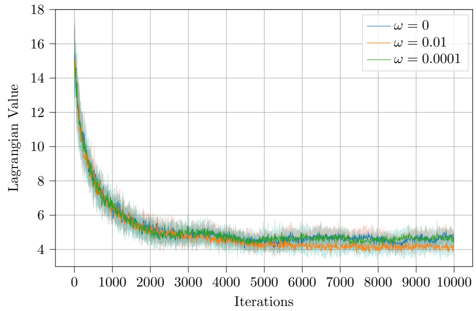


(f) Cost Curves (DC-PGPE).

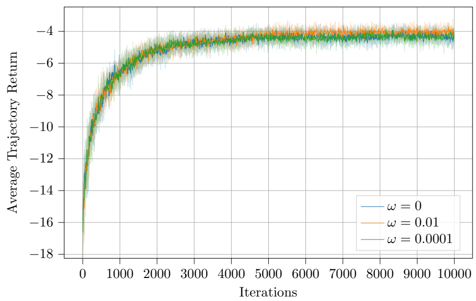
Figure 6: Sensitivity study on *CostLQR* of C-PGPE with regularization values $\omega \in \{0, 10^{-2}, 10^{-4}\}$ (5 runs, mean $\pm$ 95% C.I.).



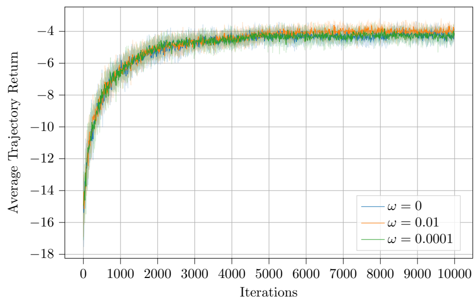
(a) Lagrangian Curves (C-PGAE).



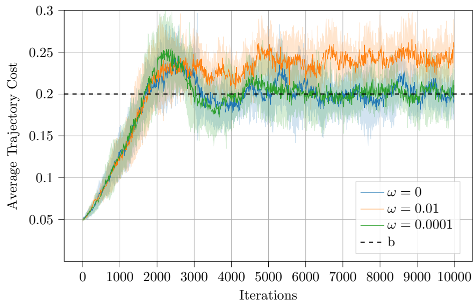
(b) Lagrangian Curves (DC-PGAE).



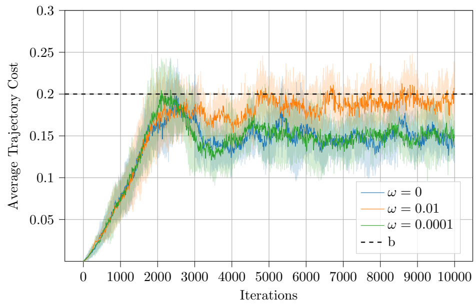
(c) Performance Curves (C-PGAE).



(d) Performance Curves (DC-PGAE).

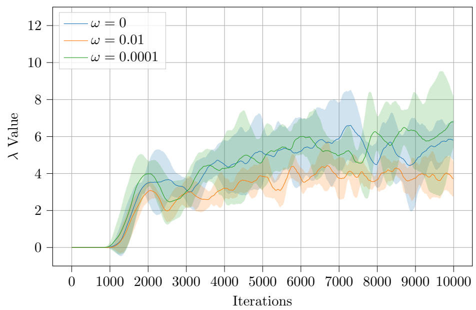


(e) Cost Curves (C-PGAE).

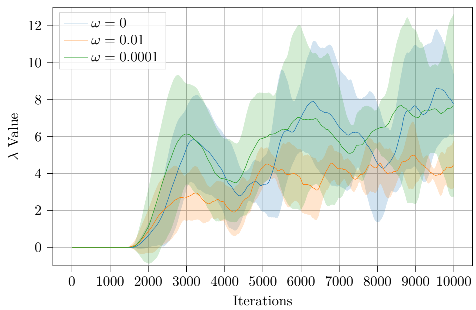


(f) Cost Curves (DC-PGAE).

Figure 7: Sensitivity study on *CostLQR* of C-PGAE with regularization values $\omega \in \{0, 10^{-2}, 10^{-4}\}$ (5 runs, mean $\pm$ 95% C.I.).



(a) Lambda Curves (C-PGPE).



(b) Lambda Curves (C-PGAE).

Figure 8: λ curves for C-PGPE and C-PGAE on *CostLQR* with C-PGAE with regularization values $\omega \in \{0, 10^{-2}, 10^{-4}\}$ (5 runs, mean $\pm$ 95% C.I.).

A. Omitted Proofs

Lemma A.1 (Regularization Bias on Saddle Points - 1). *Under Assumption 3.1, for every $\omega \geq 0$, let $(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*)$ be a saddle point of $\mathcal{L}_\omega$, it holds that:*

$$0 \leq \mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_0^*) - \mathcal{L}_0(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*) \leq \frac{\omega}{2}\left(\|\boldsymbol{\lambda}_0^*\|_2^2 - \|\boldsymbol{\lambda}_\omega^*\|_2^2\right).$$

Proof. From the fact that $(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*)$ is a saddle point of $\mathcal{L}_\omega$, we have for every $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{\lambda} \in \Lambda$:

$$\mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}_\omega^*) \geq \mathcal{L}_\omega(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*) \geq \mathcal{L}_\omega(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}) \tag{9}$$

$$\iff \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}_\omega^*) - \frac{\omega}{2}\|\boldsymbol{\lambda}_\omega^*\|_2^2 \geq \mathcal{L}_0(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*) - \frac{\omega}{2}\|\boldsymbol{\lambda}_\omega^*\|_2^2 \geq \mathcal{L}_0(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}) - \frac{\omega}{2}\|\boldsymbol{\lambda}\|_2^2 \tag{10}$$

$$\iff \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}_\omega^*) \geq \mathcal{L}_0(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*) \geq \mathcal{L}_0(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}) + \frac{\omega}{2}\left(\|\boldsymbol{\lambda}_\omega^*\|_2^2 - \|\boldsymbol{\lambda}\|_2^2\right). \tag{11}$$

From the fact that $(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_0^*)$ is a saddle point of $\mathcal{L}_0$, we have for every $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{\lambda} \in \Lambda$:

$$\mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}_0^*) \geq \mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_0^*) \geq \mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}). \tag{12}$$

By setting $(\boldsymbol{v}, \boldsymbol{\lambda}) \leftarrow (\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*)$ in Equation (12) and $(\boldsymbol{v}, \boldsymbol{\lambda}) \leftarrow (\boldsymbol{v}_0^*, \boldsymbol{\lambda}_0^*)$ in Equation (11), we obtain:

$$\mathcal{L}_0(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_0^*) \geq \mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_0^*) \geq \mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_\omega^*) \tag{13}$$

$$= \mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_\omega^*) \geq \mathcal{L}_0(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*) \geq \mathcal{L}_0(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_0^*) + \frac{\omega}{2}\left(\|\boldsymbol{\lambda}_\omega^*\|_2^2 - \left\|\boldsymbol{\lambda}_0^*\right\|_2^2\right) \tag{14}$$

$$\geq \mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_0^*) + \frac{\omega}{2}\left(\|\boldsymbol{\lambda}_\omega^*\|_2^2 - \left\|\boldsymbol{\lambda}_0^*\right\|_2^2\right), \tag{15}$$

thus:

$$\mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_0^*) \geq \mathcal{L}_0(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*) \geq \mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_0^*) + \frac{\omega}{2}\left(\|\boldsymbol{\lambda}_\omega^*\|_2^2 - \left\|\boldsymbol{\lambda}_0^*\right\|_2^2\right). \tag{16}$$

□

Lemma A.2 (Regularization Bias on Saddle Points - 2). *Under Assumption 3.1, for every $\omega \geq 0$, it holds that:*

$$0 \leq \min_{\boldsymbol{v} \in \mathcal{V}} \max_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \min_{\boldsymbol{v} \in \mathcal{V}} \max_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}) \leq \frac{\omega}{2}\|\boldsymbol{\lambda}_0^*\|_2^2.$$

Proof. The first inequality follows from the observation that $\mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) \geq \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})$ for every $\omega \geq 0$. For the second inequality, let us denote as $(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*)$ the saddle point for $\mathcal{L}_\omega$ and let $\Lambda^* = \left\{\boldsymbol{\lambda}_0^*, \boldsymbol{\lambda}_\omega^*\right\}$. We have:

$$\mathcal{L}_0(\boldsymbol{v}_0^*, \boldsymbol{\lambda}_0^*) - \mathcal{L}_\omega(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*) = \min_{\boldsymbol{v} \in \mathcal{V}} \max_{\boldsymbol{\lambda} \in \Lambda^*} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \min_{\boldsymbol{v} \in \mathcal{V}} \max_{\boldsymbol{\lambda} \in \Lambda^*} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}) \tag{17}$$

$$\leq \max_{\boldsymbol{v} \in \mathcal{V}} \left| \max_{\boldsymbol{\lambda} \in \Lambda^*} \mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \max_{\boldsymbol{\lambda} \in \Lambda^*} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}) \right| \tag{18}$$

$$= \max_{\boldsymbol{v} \in \mathcal{V}, \boldsymbol{\lambda} \in \Lambda^*} \left|\mathcal{L}_0(\boldsymbol{v}, \boldsymbol{\lambda}) - \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})\right| \tag{19}$$

$$= \frac{\omega}{2} \max\left\{\left\|\boldsymbol{\lambda}_0^*\right\|_2^2;\ \|\boldsymbol{\lambda}_\omega^*\|_2^2\right\} \tag{20}$$

$$= \frac{\omega}{2}\left\|\boldsymbol{\lambda}_0^*\right\|_2^2, \tag{21}$$

where we used Lemma A.1 to conclude that $\left\|\boldsymbol{\lambda}_0^*\right\|_2^2 \geq \|\boldsymbol{\lambda}_\omega^*\|_2^2$. □

Lemma A.3 (Objective bound and Constraint violation). *Under Assumption 3.1, for every $\omega \geq 0$, letting $(\boldsymbol{v}^*_\omega, \boldsymbol{\lambda}^*_\omega)$ be a saddle point of $\mathcal{L}_\omega$, it holds that:*

$$0 \leq J_0(\boldsymbol{v}^*_0) - J_0(\boldsymbol{v}^*_\omega) \leq \omega \|\boldsymbol{\lambda}^*_0\|_2^2, \tag{22}$$
$$\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2 \leq \omega \|\boldsymbol{\lambda}^*_0\|_2. \tag{23}$$

Proof. Since $(\boldsymbol{v}^*_0, \boldsymbol{\lambda}^*_0)$ is a saddle point of $\mathcal{L}_0$, it holds that $\boldsymbol{v}^*_0$ is feasible and, consequently, $\mathcal{L}_0(\boldsymbol{v}^*_0, \boldsymbol{\lambda}^*_0) = J_0(\boldsymbol{v}^*_0)$. Moreover, let $\omega > 0$: since $(\boldsymbol{v}^*_\omega, \boldsymbol{\lambda}^*_\omega)$ is a saddle point of $\mathcal{L}_\omega$ it holds that $\boldsymbol{\lambda}^*_\omega = \boldsymbol{\lambda}^*(\boldsymbol{v}^*_\omega) = \Pi_\Lambda \left(\frac{1}{\omega}(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b}) \right) = \frac{1}{\omega}(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+$, since $\frac{1}{\omega}\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2 \leq \omega^{-1}\sqrt{U} J_{\max}$. Thus, we have:

$$\mathcal{L}_0(\boldsymbol{v}^*_\omega, \boldsymbol{\lambda}^*_\omega) = J_0(\boldsymbol{v}^*_\omega) + \langle \boldsymbol{\lambda}^*_\omega, \mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b} \rangle = J_0(\boldsymbol{v}^*_\omega) + \frac{1}{\omega}\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2^2. \tag{24}$$

From Lemma A.1, we have:

$$0 \leq J_0(\boldsymbol{v}^*_0) - J_0(\boldsymbol{v}^*_\omega) - \frac{1}{\omega}\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2^2 \leq \frac{\omega}{2}\|\boldsymbol{\lambda}^*_0\|_2^2 - \frac{1}{2\omega}\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2^2. \tag{25}$$

By summing $\frac{1}{\omega}\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2^2$ to all members, we have:

$$\frac{1}{\omega}\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2^2 \leq J_0(\boldsymbol{v}^*_0) - J_0(\boldsymbol{v}^*_\omega) \leq \frac{\omega}{2}\|\boldsymbol{\lambda}^*_0\|_2^2 + \frac{1}{2\omega}\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2^2. \tag{26}$$

Now taking the first and last member, we conclude:

$$\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2^2 \leq \omega^2 \|\boldsymbol{\lambda}^*_0\|_2^2. \tag{27}$$

Since $\frac{1}{\omega}\|(\mathbf{J}(\boldsymbol{v}^*_\omega) - \mathbf{b})^+\|_2^2 \geq 0$ and plugging the latter inequality into the third member of (26) we obtain:

$$0 \leq J_0(\boldsymbol{v}^*_0) - J_0(\boldsymbol{v}^*_\omega) \leq \omega \|\boldsymbol{\lambda}^*_0\|_2^2. \tag{28}$$

□

Lemma A.4 (Weak ψ-Gradient Domination on $H_\omega(\boldsymbol{v})$). *Under Assumption 3.2, if $\omega > 0$, for every $\boldsymbol{v} \in \mathcal{V}$ and $\boldsymbol{\lambda} \in \Lambda$, it holds that:*

$$\left\|\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v})\right\|_2^\psi \geq \alpha_1 \left(H_\omega(\boldsymbol{v}) - \min_{\boldsymbol{v}' \in \mathcal{V}} H_\omega(\boldsymbol{v}') \right) - \beta_1. \tag{29}$$

Proof. If $\omega > 0$, the dual variable exist finite since the maximization problem over $\boldsymbol{\lambda}$ is concave:

$$\boldsymbol{\lambda}^*(\boldsymbol{v}) = \arg\max_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}).$$

Thus, we have from Lemma A.7 that$\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}) = \nabla_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*(\boldsymbol{v})}$ and by Assumption 3.2 we have the following:

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v})\right\|_2 &= \left\|\nabla_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*(\boldsymbol{v})}\right\|_2 && (30)\\
&\geq \alpha_1 \left(\mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}^*(\boldsymbol{v})) - \min_{\boldsymbol{v}' \in \mathcal{V}} \mathcal{L}_\omega(\boldsymbol{v}', \boldsymbol{\lambda}^*(\boldsymbol{v})) \right) - \beta_1 && (31)\\
&\geq \alpha_1 \left(H_\omega(\boldsymbol{v}) - \min_{\boldsymbol{v}' \in \mathcal{V}} \max_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}', \boldsymbol{\lambda}) \right) - \beta_1 && (32)\\
&= \alpha_1 \left(H_\omega(\boldsymbol{v}) - H^*_\omega \right) - \beta_1. && (33)
\end{aligned}$$

□

Lemma A.5. *Let $\omega > 0$ and $\boldsymbol{v} \in \mathcal{V}$. The following statements hold:*

- $\mathcal{L}_\omega(\boldsymbol{v}, \cdot)$ *is ω-smooth, i.e., for every $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$ it holds that:*

$$\left|\nabla_{\boldsymbol{\lambda}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}') - \nabla_{\boldsymbol{\lambda}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})\right| \le \omega \left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\right\|_2^2$$

- $\mathcal{L}_\omega(\boldsymbol{v}, \cdot)$ *satisfies the PL condition, i.e., for every $\boldsymbol{\lambda} \in \Lambda$ it holds that:*

$$\|\nabla_{\boldsymbol{\lambda}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})\|_2^2 \ge \omega \left(\max_{\boldsymbol{\lambda}' \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}') - \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})\right).$$

- $\mathcal{L}_\omega(\boldsymbol{v}, \cdot)$ *satisfies the* error bound *(EB) condition, i.e., for every $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$ it holds that:*

$$\|\nabla_{\boldsymbol{\lambda}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})\| \ge \frac{\omega}{2} \|\boldsymbol{\lambda}^*(\boldsymbol{v}) - \boldsymbol{\lambda}\|_2,$$

where $\boldsymbol{\lambda}^*(\boldsymbol{v}) = \arg\max_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})$.

- $\mathcal{L}_\omega(\boldsymbol{v}, \cdot)$ *satisfies the* quadratic growth *(QG) condition, i.e., for every $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda$ it holds that:*

$$H_\omega(\boldsymbol{v}) - \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}) \ge \frac{\omega}{4} \|\boldsymbol{\lambda}^*(\boldsymbol{v}) - \boldsymbol{\lambda}\|_2,$$

where $\boldsymbol{\lambda}^*(\boldsymbol{v}) = \arg\max_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})$.

Proof. For the first property, it is enough to observe that $\mathcal{L}_\omega$ is twice differentiable in $\boldsymbol{\lambda}$ and that its Hessian is $\omega \mathbf{I}$. For the second property, we observe that $\mathcal{L}_\omega$ is quadratic in $\boldsymbol{\lambda}$ and, consequently it satisfies the PL condition with parameter ω:

$$\|\nabla_{\boldsymbol{\lambda}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})\|_2^2 \ge \omega \left(\max_{\boldsymbol{\lambda}' \in \mathbb{R}^U} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}') - \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})\right) \ge \omega \left(\max_{\boldsymbol{\lambda}' \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}') - \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})\right).$$

For the third and fourth properties, we refer to Lemma A.1 of Yang et al. (2020). □

Lemma A.6. *Let $\omega > 0$. For every $\boldsymbol{v} \in \mathcal{V}$, it holds that:*

$$H_\omega(\boldsymbol{v}) - H_\omega^* \ge \frac{\omega}{4} \|\boldsymbol{\lambda}^*(\boldsymbol{v}) - \boldsymbol{\lambda}_\omega^*\|_2. \tag{34}$$

Proof. Let us consider the following derivation:

$$\begin{aligned}
H_\omega(\boldsymbol{v}) - H_\omega^* &= H_\omega(\boldsymbol{v}) - \mathcal{L}_\omega(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*) &(35)\\
&\ge H_\omega(\boldsymbol{v}) - \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}_\omega^*) &(36)\\
&\ge \frac{\omega}{4} \|\boldsymbol{\lambda}^*(\boldsymbol{v}) - \boldsymbol{\lambda}_\omega^*\|_2. &(37)
\end{aligned}$$

having exploited the fact that, from the saddle point property, $\mathcal{L}_\omega(\boldsymbol{v}_\omega^*, \boldsymbol{\lambda}_\omega^*) \le \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}_\omega^*)$ and, then, Lemma A.5. □

Lemma A.7. *Let $\omega > 0$. The following statements hold:*

- H_ω *is L_H-smooth, i.e., for every $\boldsymbol{v}, \boldsymbol{v}' \in \mathcal{V}$, it holds that:*

$$\|\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}') - \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v})\|_2 \le L_H \|\boldsymbol{v}' - \boldsymbol{v}\|_2.$$

where $L_H := L_2 + \frac{L_1^2}{\omega}$.

- *For every $\boldsymbol{v}, \boldsymbol{v}' \in \mathcal{V}$ we have $\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}) = \nabla_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})|_{\boldsymbol{\lambda} = \boldsymbol{\lambda}^*(\boldsymbol{v})}$, where $\boldsymbol{\lambda}^*(\boldsymbol{v}) = \arg\max_{\boldsymbol{\lambda} \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})$.*

Proof. The first and second statements follow from Lemma A.5 of Nouiehed, Sanjabi, Huang, Lee and Razaviyayn (2019). □

Theorem 3.2 (Convergence of $\mathcal{P}_K$). *Under Assumptions 3.2, 3.3, 3.4, for $\chi < 1/5$, sufficiently small ϵ and ω, and a choice of* constant *learning rates $\zeta_{\boldsymbol{v}}, \zeta_\lambda$, we have $\mathcal{P}_K(\chi) \le \epsilon + \beta_1/\alpha_1$ whenever:*[8]

- $K = \mathcal{O}(\omega^{-1}\log(\epsilon^{-1}))$ *if $\psi = 2$ and the gradients are exact (i.e., $V_{\boldsymbol{v}} = V_\lambda = 0$);*
- $K = \mathcal{O}(\omega^{-1}\epsilon^{-\frac{2}{\psi}-1})$ *if $\psi \in [1,2)$ and the gradients are exact (i.e., $V_{\boldsymbol{v}} = V_\lambda = 0$);*
- $K = \mathcal{O}(\omega^{-3}\epsilon^{-\frac{4}{\psi}+1})$ *if $\psi \in [1,2]$ and the gradients are estimated (i.e., $V_{\boldsymbol{v}} = \mathcal{O}(\omega^{-2})$ and $V_\lambda = \mathcal{O}(1)$).*

Proof. The proof is subdivided into several parts. We will omit the ω subscript for notational easiness. Let us focus on a specific iteration $k \in \mathbb{N}$.

Part I: bounding the a_k term. Let us start with the a_k term:

$$H_\omega(\boldsymbol{v}_{k+1}) - H^* \le H_\omega(\boldsymbol{v}_k) - H^* + \left\langle \boldsymbol{v}_{k+1} - \boldsymbol{v}_k,\ \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\rangle + \frac{L_H}{2}\left\|\boldsymbol{v}_{k+1} - \boldsymbol{v}_k\right\|_2^2 \tag{38}$$

$$\le H_{\boldsymbol{v}}(\boldsymbol{v}_k) - H^* - \zeta_{\boldsymbol{v},k}\left\langle \widehat{\nabla}_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k),\ \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\rangle \tag{39}$$

$$+ \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2\left\|\widehat{\nabla}_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\right\|_2^2, \tag{40}$$

where the first line is due to the fact that the function H_ω is L_H-smooth (Lemma A.7), the last inequality is due to the update rule of $\boldsymbol{v}$. Now, we apply the expected value on both sides of the inequality and we use the fact that the gradient estimation is unbiased and has variance bounded by $V_{\boldsymbol{v}}$:

$$\mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1})|\mathcal{F}_{k-1}\right] - H^* \le H_\omega(\boldsymbol{v}_k) - H^* - \zeta_{\boldsymbol{v},k}\left\langle \nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k),\ \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\rangle \tag{41}$$

$$+ \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2\,\mathbb{E}\left[\left\|\widehat{\nabla}_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\right\|_2^2 |\mathcal{F}_{k-1}\right], \tag{42}$$

where $\mathcal{F}_{k-1}$ is the filtration associated with all events realized up to interaction $k-1$. We recall that:

$$\mathbb{E}\left[\left\|\widehat{\nabla}_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\right\|_2^2 |\mathcal{F}_{k-1}\right] = \mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)|\mathcal{F}_{k-1}\right] + \left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\right\|_2^2, \tag{43}$$

and that $\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\right] \le V_{\boldsymbol{v}}$ by Assumption 3.4. Thus, selecting $\zeta_{\boldsymbol{v},k} \le 1/L_H$, we have that:

$$\mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1})|\mathcal{F}_{k-1}\right] - H^* \tag{44}$$

$$\le H_\omega(\boldsymbol{v}_k) - H^* - \zeta_{\boldsymbol{v},k}\left\langle \nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\ \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\rangle \tag{45}$$

$$+ \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\right\|_2^2 + \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}} \tag{46}$$

$$= H_\omega(\boldsymbol{v}_k) - H^* - \zeta_{\boldsymbol{v},k}\left\langle \nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\ \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\rangle \tag{47}$$

$$+ \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) \pm \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\|_2^2 + \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}}. \tag{48}$$

Consider that:

$$\frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k) + \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\|_2^2 \tag{49}$$

$$= \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\|_2^2 - \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\|_2^2 \tag{50}$$

$$+ \zeta_{\boldsymbol{v},k}\left\langle \nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k),\ \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\right\rangle. \tag{51}$$

[8] In the context of this statement, the $\mathcal{O}(\cdot)$ notation preserves dependences on ϵ and ω only.

Thus, the following holds:

$$\mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1})|\mathcal{F}_{k-1}\right] - H^* \tag{52}$$

$$\leq H_\omega(\boldsymbol{v}_k) - H^* - \zeta_{\boldsymbol{v},k}\left\langle \nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k),\ \nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\rangle + \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}} \tag{53}$$

$$+ \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 - \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 \tag{54}$$

$$+ \zeta_{\boldsymbol{v},k}\left\langle \nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k),\ \nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\rangle \tag{55}$$

$$= H_\omega(\boldsymbol{v}_k) - H^* - \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 + \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 \tag{56}$$

$$+ \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}}. \tag{57}$$

Thus, we have obtained:

$$\mathsf{A} := \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1})|\mathcal{F}_{k-1}\right] - H^* \tag{58}$$

$$\leq H_\omega(\boldsymbol{v}_k) - H^* - \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 + \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 \tag{59}$$

$$+ \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}}, \tag{60}$$

holding via the selection of $\zeta_{\boldsymbol{v},k} \leq 1/L_H$. Notice that, from A the following directly follows:

$$\mathsf{D} := \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1})|\mathcal{F}_{k-1}\right] - H_\omega(\boldsymbol{v}_k) \tag{61}$$

$$\leq -\frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 + \frac{\zeta_{\boldsymbol{v},k}}{2}\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 + \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}}. \tag{62}$$

Part II: bounding the b_k term. We are ready to analyze the b_k term. Recall that for ridge regularization of the Lagrangian function presented in the main paper, we have that $\mathcal{L}_\omega$ is ω-smooth and fulfills the PL condition with constant ω, as shown in Lemma A.5. Since $\mathcal{L}$ is a quadratic function of λ and $\lambda^*(\boldsymbol{v}_{k+1}) \in \Lambda$, we have that considering the non-projected λ_{k+1} can only increase the distance. Thus, we will ignore projection for the rest of the proof. We have:

$$H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_{k+1}) \tag{63}$$

$$\leq H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k) - \left\langle \lambda_{k+1} - \lambda_k,\ \nabla_\lambda\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)\right\rangle + \frac{\omega}{2}\left\|\lambda_{k+1} - \lambda_k\right\|_2^2 \tag{64}$$

$$= H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k) - \zeta_{\lambda,k}\left\langle \widehat{\nabla}_\lambda\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k),\ \nabla_\lambda\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)\right\rangle \tag{65}$$

$$+ \frac{\omega}{2}\zeta_{\lambda,k}^2\left\|\widehat{\nabla}_\lambda\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)\right\|_2^2, \tag{66}$$

that is possible under Assumption 3.3 (i.e., $\mathcal{L}_\omega$ is L_2-smooth) and due to the update rules we are considering. Now, by applying the expectation on both sides, we obtain the following:

$$\mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_{k+1})|\mathcal{F}_{k-1}\right] \tag{67}$$

$$\leq \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)|\mathcal{F}_{k-1}\right] - \zeta_{\lambda,k}\left\|\nabla_\lambda\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)\right\|_2^2 \tag{68}$$

$$+ \frac{\omega}{2}\zeta_{\lambda,k}^2\,\mathbb{E}\left[\left\|\widehat{\nabla}_\lambda\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)\right\|_2^2|\mathcal{F}_{k-1}\right] \tag{69}$$

$$\leq \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)|\mathcal{F}_{k-1}\right] - \zeta_{\lambda,k}\left\|\nabla_\lambda\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)\right\|_2^2 + \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda \tag{70}$$

$$+ \frac{\omega}{2}\zeta_{\lambda,k}^2\left\|\widehat{\nabla}_\lambda\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda)\right\|_2^2 \tag{71}$$

$$\leq \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)|\mathcal{F}_{k-1}\right] - \frac{\zeta_{\lambda,k}}{2}\left\|\nabla_\lambda\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_k)\right\|_2^2 + \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda, \tag{72}$$

where the last line follows by selecting $\zeta_{\lambda,k} \leq 1/\omega$. Since $\mathcal{L}_\omega$ enjoys the PL condition w.r.t λ with constant ω, for every pair $(\boldsymbol{v}, \boldsymbol{\lambda})$ we have:

$$\left\|\nabla_\lambda \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda})\right\|_2^2 \geq \omega \left(\max_{\overline{\lambda} \in \mathbb{R}^U} \mathcal{L}_\omega(\boldsymbol{v}, \overline{\boldsymbol{\lambda}}) - \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}) \right) \geq \omega \left(\max_{\overline{\lambda} \in \Lambda} \mathcal{L}_\omega(\boldsymbol{v}, \overline{\boldsymbol{\lambda}}) - \mathcal{L}_\omega(\boldsymbol{v}, \boldsymbol{\lambda}) \right). \tag{73}$$

By applying the PL condition:

$$\begin{aligned}
&\mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_{k+1}) | \mathcal{F}_{k-1}\right] && (74)\\
&\leq \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_{k}) | \mathcal{F}_{k-1}\right] - \frac{\zeta_{\lambda,k}}{2} \left\|\nabla_\lambda \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_k)\right\|_2^2 + \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda && (75)\\
&\leq \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_{k}) | \mathcal{F}_{k-1}\right] - \frac{\zeta_{\lambda,k}}{2} \omega\, \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_{k}) | \mathcal{F}_{k-1}\right] && (76)\\
&\quad + \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda && (77)\\
&= \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_{k}) | \mathcal{F}_{k-1}\right] + \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda, && (78)
\end{aligned}$$

where we enforce $1 - \frac{\zeta_{\lambda,k}}{2}\omega \geq 0$, i.e., $\zeta_{\lambda,k} \leq 2/\omega$. However, we do not have a proper recursive term, thus consider the following:

$$\begin{aligned}
H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_k) &= \underbrace{H_\omega(\boldsymbol{v}_k) - \mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k)}_{\text{Recursive Term}} + \underbrace{\mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_k)}_{\mathsf{C}} && (79)\\
&\quad + \underbrace{H_\omega(\boldsymbol{v}_{k+1}) - H_\omega(\boldsymbol{v}_k)}_{\mathsf{D}}. && (80)
\end{aligned}$$

A bound on D has already been derived, so let us bound the term C:

$$\begin{aligned}
\mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_k) &\leq - \left\langle \boldsymbol{v}_{k+1} - \boldsymbol{v}_k,\ \nabla_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k)\right\rangle + \frac{L_2}{2}\left\|\boldsymbol{v}_{k+1} - \boldsymbol{v}_k\right\|_2^2 && (81)\\
&\leq \zeta_{\boldsymbol{v},k} \left\langle \widehat{\nabla}_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k),\ \nabla_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k)\right\rangle + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}^2 \left\|\widehat{\nabla}_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k)\right\|_2^2, && (82)
\end{aligned}$$

again because of Assumption 3.3 and the update rule. Now, as usual, we consider the expectation conditioned to the filtration $\mathcal{F}_{k-1}$ and the properties of the variance, to obtain:

$$\begin{aligned}
\mathsf{C} &:= \mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k) - \mathbb{E}\left[\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_k) | \mathcal{F}_{k-1}\right] && (83)\\
&\leq \zeta_{\boldsymbol{v},k}\left(1 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right) \left\|\nabla_{\boldsymbol{v}} \mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k)\right\|_2^2 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}}, && (84)
\end{aligned}$$

having set $\zeta_{\boldsymbol{v},k} \leq 1/L_2$. We are finally able to conclude the bound of the term B:

$$\begin{aligned}
\mathsf{B} &:= \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_{k+1}) | \mathcal{F}_{k-1}\right] && (85)\\
&\leq \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_{k}) | \mathcal{F}_{k-1}\right] + \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda && (86)\\
&= \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) \left(H_\omega(\boldsymbol{v}_k) - \mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k)\right) && (87)\\
&\quad + \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) \left(\mathcal{L}_\omega(\boldsymbol{v}_k, \boldsymbol{\lambda}_k) - \mathbb{E}\left[\mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \boldsymbol{\lambda}_k) | \mathcal{F}_{k-1}\right]\right) && (88)\\
&\quad + \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) \left(\mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) | \mathcal{F}_{k-1}\right] - H_\omega(\boldsymbol{v}_k)\right) + \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda. && (89)
\end{aligned}$$

Now we apply the bounds on C and D (the latter is from Eq. 62), obtaining:

$$
\begin{aligned}
&\mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_{k+1}) | \mathcal{F}_{k-1}\right] && (90)\\
&\leq \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\left(H_\omega(\boldsymbol{v}_k) - \mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\right) && (91)\\
&\quad + \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\left(\zeta_{\boldsymbol{v},k}\left(1 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right)\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\|_2^2 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}}\right) && (92)\\
&\quad + \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\left(-\frac{\zeta_{\boldsymbol{v},k}}{2}\|\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\|_2^2 + \frac{\zeta_{\boldsymbol{v},k}}{2}\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\|_2^2\right. && (93)\\
&\qquad \left. + \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}}\right) + \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda, && (94)
\end{aligned}
$$

that is the second fundamental term.

Part III: bounding the potential function $P_k(\chi)$. Before going on, we recall that so far we enforced: $\zeta_{\boldsymbol{v},k} \leq 1/L_H$ (since $L_H \geq L_2$) and $\zeta_{\lambda,k} \leq 1/\omega$, for every $t \in [\![K]\!]$. What we want to bound here is the potential function $P_{k+1}(\chi) = a_{k+1} + \chi b_{k+1}$. Using the final results of Part I and Part II:

$$
\begin{aligned}
a_{k+1} + \chi b_{k+1} &= \mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - H^*\right] + \chi\,\mathbb{E}\left[H_\omega(\boldsymbol{v}_{k+1}) - \mathcal{L}_\omega(\boldsymbol{v}_{k+1}, \lambda_{k+1})\right] && (95)\\
&\leq \mathbb{E}\left[H_\omega(\boldsymbol{v}_k) - H^*\right] - \frac{\zeta_{\boldsymbol{v},k}}{2}\mathbb{E}\left[\|\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\|_2^2\right] && (96)\\
&\quad + \frac{\zeta_{\boldsymbol{v},k}}{2}\mathbb{E}\left[\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\|_2^2\right] + \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}} && (97)\\
&\quad + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\mathbb{E}\left[H_\omega(\boldsymbol{v}_k) - \mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\right] && (98)\\
&\quad + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\left(\zeta_{\boldsymbol{v},k}\left(1 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right)\mathbb{E}\left[\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\|_2^2\right] + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}}\right) && (99)\\
&\quad + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\left(-\frac{\zeta_{\boldsymbol{v},k}}{2}\mathbb{E}\left[\|\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\|_2^2\right]\right. && (100)\\
&\qquad \left. + \frac{\zeta_{\boldsymbol{v},k}}{2}\mathbb{E}\left[\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\|_2^2\right] + \frac{L_H}{2}\zeta_{\boldsymbol{v},k}^2 V_{\boldsymbol{v}}\right) && (101)\\
&\quad + \chi\frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda && (102)\\
&= a_k + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) b_k && (103)\\
&\quad - \frac{\zeta_{\boldsymbol{v},k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\mathbb{E}\left[\|\nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\|_2^2\right] && (104)\\
&\quad + \frac{\zeta_{\boldsymbol{v},k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\mathbb{E}\left[\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k) - \nabla_{\boldsymbol{v}} H_\omega(\boldsymbol{v}_k)\|_2^2\right] && (105)\\
&\quad + \zeta_{\boldsymbol{v},k}\left(1 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right)\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\mathbb{E}\left[\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k, \lambda_k)\|_2^2\right] && (106)\\
&\quad + \frac{\zeta_{\boldsymbol{v},k}^2}{2}\left(L_H + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)(L_H + L_2)\right) V_{\boldsymbol{v}} + \chi\frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda. && (107)
\end{aligned}
$$

Now we can re-arrange the terms by noticing that:

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)\right\|_2^2 &= \left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)+\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 && (108)\\
&= \left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 && (109)\\
&\quad + \left\|\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 + 2\left\langle \nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k),\ \nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\rangle && (110)\\
&\le 2\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 + 2\left\|\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2, && (111)
\end{aligned}$$

where the last inequality holds by Young's inequality. Then we can write what follows:

$$\begin{aligned}
& a_{k+1}+\chi b_{k+1} && (112)\\
&\le a_k + \chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)b_k && (113)\\
&\quad + \Bigg(2\zeta_{\boldsymbol{v},k}\left(1+\frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right)\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right) && (114)\\
&\quad\quad -\frac{\zeta_{\boldsymbol{v},k}}{2}\left(1+\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\Bigg)\mathbb{E}\left[\left\|\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2\right] && (115)\\
&\quad + \left(2\zeta_{\boldsymbol{v},k}\left(1+\frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right)\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)+\frac{\zeta_{\boldsymbol{v},k}}{2}\left(1+\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right) && (116)\\
&\quad\quad \cdot\mathbb{E}\left[\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2\right] && (117)\\
&\quad + \frac{\zeta_{\boldsymbol{v},k}^2}{2}\left(L_H+\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)(L_H+L_2)\right)V_{\boldsymbol{v}} + \chi\frac{\omega}{2}\zeta_{\lambda,k}^2 V_{\lambda}. && (118)
\end{aligned}$$

Let us now proceed to bound $\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2$. By Lemma A.7, we have that $\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}) = \nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v},\boldsymbol{\lambda}^*(\boldsymbol{v}))$ for every $\boldsymbol{\lambda}^*(\boldsymbol{v})\in\arg\max_{\overline{\boldsymbol{\lambda}}\in\Lambda}\mathcal{L}_\omega(\boldsymbol{v},\overline{\boldsymbol{\lambda}})$, thus we can write:

$$\begin{aligned}
\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 &= \left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}^*(\boldsymbol{v}_k))\right\|_2^2 && (119)\\
&\le L_3^2\left\|\boldsymbol{\lambda}^*(\boldsymbol{v}_k)-\boldsymbol{\lambda}_k\right\|_2^2, && (120)
\end{aligned}$$

since Assumption 3.3 holds.

For a fixed value of $\boldsymbol{v}$, by Lemma A.5 it follows that $\mathcal{L}_\omega(\boldsymbol{v},\cdot)$ satisfies the quadratic growth condition (since it satisfies the PL condition), for which the following holds:

$$\left\|\boldsymbol{\lambda}^*(\boldsymbol{v}_k)-\boldsymbol{\lambda}_k\right\|_2^2 \le \frac{4}{\omega}\left(H_\omega(\boldsymbol{v}_k)-\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)\right), \tag{121}$$

and thus we have:

$$\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2 \le \frac{4L_3^2}{\omega}\left(H_\omega(\boldsymbol{v}_k)-\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)\right). \tag{122}$$

By applying the total expectation, it trivially follows:

$$\mathbb{E}\left[\left\|\nabla_{\boldsymbol{v}}\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)-\nabla_{\boldsymbol{v}}H_\omega(\boldsymbol{v}_k)\right\|_2^2\right] \le \frac{4L_3^2}{\omega}\mathbb{E}\left[H_\omega(\boldsymbol{v}_k)-\mathcal{L}_\omega(\boldsymbol{v}_k,\boldsymbol{\lambda}_k)\right] = \frac{4L_3^2}{\omega}b_k. \tag{123}$$

Thus, we have:

$$a_{k+1} + \chi b_{k+1} \tag{124}$$

$$\leq a_k + \chi \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) b_k \tag{125}$$

$$+ \left(2\zeta_{\upsilon,k}\left(1 + \frac{L_2}{2}\zeta_{\upsilon,k}\right)\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right. \tag{126}$$

$$\left. - \frac{\zeta_{\upsilon,k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right)\mathbb{E}\left[\|\nabla_{\upsilon} H_\omega(\upsilon_k)\|_2^2\right] \tag{127}$$

$$+ \left(2\zeta_{\upsilon,k}\left(1 + \frac{L_2}{2}\zeta_{\upsilon,k}\right)\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) + \frac{\zeta_{\upsilon,k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right)\frac{4L_3^2}{\omega} b_k \tag{128}$$

$$+ \frac{\zeta_{\upsilon,k}^2}{2}\left(L_H + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)(L_H + L_2)\right)V_{\upsilon} + \chi\frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda. \tag{129}$$

Part IV: apply the ψ-gradient domination. Now we need to bound the term $\|\nabla H_\omega(\upsilon_k)\|_2^2$. We consider Assumption 3.2 and we get: $\|\nabla_{\upsilon} H_\omega(\upsilon_k)\|_2^{\psi} \geq \alpha_1\left(H_\omega(\upsilon_k) - H^*\right) - \beta_1$. By defining $\widetilde{H}^* := H^* + \beta_1/\alpha_1$, we also have:

$$\|\nabla_{\upsilon} H_\omega(\upsilon)\|_2^{\psi} \geq \alpha_1 \max\left\{0,\ H_\omega(\upsilon) - \widetilde{H}^*\right\} \tag{130}$$

$$\Longrightarrow$$

$$\|\nabla_{\upsilon} H_\omega(\upsilon)\|_2^2 \geq \alpha_1^{\frac{2}{\psi}} \max\left\{0,\ H_\omega(\upsilon) - \widetilde{H}^*\right\}^{\frac{2}{\psi}}.$$

If we apply the total expectation on both sides of the inequality, we get:

$$\mathbb{E}\left[\|\nabla_{\upsilon} H_\omega(\upsilon)\|_2^2\right] \geq \alpha_1^{\frac{2}{\psi}}\,\mathbb{E}\left[\max\left\{0,\ H_\omega(\upsilon) - \widetilde{H}^*\right\}^{\frac{2}{\psi}}\right] \tag{131}$$

$$\geq \alpha_1^{\frac{2}{\psi}}\,\mathbb{E}\left[\max\left\{0,\ H_\omega(\upsilon) - \widetilde{H}^*\right\}\right]^{\frac{2}{\psi}} \tag{132}$$

$$\geq \alpha_1^{\frac{2}{\psi}} \max\left\{0,\ \mathbb{E}\left[H_\omega(\upsilon) - \widetilde{H}^*\right]\right\}^{\frac{2}{\psi}}, \tag{133}$$

which is achieved by a double application of Jensen's inequality, since $z^{2/\psi}$ is convex for $\psi \in [1,2]$ and $z \geq 0$, and the maximum is convex. Let us start from Equation (124):

$$a_{k+1} + \chi b_{k+1} \tag{134}$$

$$\leq a_k + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) b_k \tag{135}$$

$$+ \underbrace{\left(2\zeta_{\upsilon,k}\left(1 + \frac{L_2}{2}\zeta_{\upsilon,k}\right)\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) - \frac{\zeta_{\upsilon,k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right)}_{=:-C} \tag{136}$$

$$\cdot\,\mathbb{E}\left[\|\nabla H_\omega(\upsilon_k)\|_2^2\right] \tag{137}$$

$$+ \left(2\zeta_{\upsilon,k}\left(1 + \frac{L_2}{2}\zeta_{\upsilon,k}\right)\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) + \frac{\zeta_{\upsilon,k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right)\frac{4L_3^2}{\omega} b_k \tag{138}$$

$$\underbrace{+ \frac{\zeta_{\upsilon,k}^2}{2}\left(L_H + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)(L_H + L_2)\right)V_{\upsilon} + \chi\frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda}_{=:V}. \tag{139}$$

We first enforce the negativity of $-C$. To this end:

$$-C = \left(2\zeta_{\boldsymbol{v},k} \underbrace{\left(1 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right)}_{\leq 3/2} \chi \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) - \frac{\zeta_{\boldsymbol{v},k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right) \tag{140}$$

$$\leq \zeta_{\boldsymbol{v},k}\left(3\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) - \frac{1}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right) \tag{141}$$

$$\leq \frac{\zeta_{\boldsymbol{v},k}}{2}\left(5\chi\underbrace{\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)}_{\leq 1} - 1\right) \leq \frac{\zeta_{\boldsymbol{v},k}}{2}(5\chi - 1) \leq 0. \tag{142}$$

Thus, it is enough to enforce $5\chi - 1 \leq 0 \implies \chi \leq 1/5$. We now plug in the gradient domination inequalities:

$$a_{k+1} + \chi b_{k+1} \tag{143}$$

$$\leq a_k + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) b_k - C\alpha_1^{\frac{2}{\psi}} \max\left\{0;\ \mathbb{E}\left[H_\omega(\boldsymbol{v}) - \widetilde{H}^*\right]\right\}^{\frac{2}{\psi}} + V \tag{144}$$

$$+ \left(2\zeta_{\boldsymbol{v},k}\left(1 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right)\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) + \frac{\zeta_{\boldsymbol{v},k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right)\frac{4L_3^2}{\omega} b_k. \tag{145}$$

Now we introduce the symbol $\widetilde{a}_k := \mathbb{E}\left[H_\omega(\boldsymbol{v}_k) - \widetilde{H}^*\right] = a_k - \beta_1/\alpha_1$, to get:

$$\widetilde{a}_{k+1} + \chi b_{k+1} \tag{146}$$

$$\leq \widetilde{a}_k - C\alpha_1^{\frac{2}{\psi}} \max\left\{0, \widetilde{a}_k\right\}^{\frac{2}{\psi}} + V \tag{147}$$

$$+ \left(\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) + \left(2\zeta_{\boldsymbol{v},k}\left(1 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right)\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right.\right. \tag{148}$$

$$\left.\left. + \frac{\zeta_{\boldsymbol{v},k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right)\frac{4L_3^2}{\omega}\right) b_k. \tag{149}$$

For what follows, we call B the term that is multiplying b_k:

$$B := \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) + \left(2\zeta_{\boldsymbol{v},k}\left(1 + \frac{L_2}{2}\zeta_{\boldsymbol{v},k}\right)\chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right. \tag{150}$$

$$\left. + \frac{\zeta_{\boldsymbol{v},k}}{2}\left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right)\right)\frac{4L_3^2}{\omega} \tag{151}$$

Let refer to $\widetilde{a}_k + \chi b_k$ as $\widetilde{P}_t(\chi)$ with $\chi \in (0, 1)$. For the sake of clarity, we re-write our main inequality as:

$$\widetilde{P}_t(\chi) = \widetilde{a}_{k+1} + \chi b_{k+1} \leq \widetilde{a}_k + Bb_k - C\max\left\{0;\ \widetilde{a}_k\right\}^{\frac{2}{\psi}} + V. \tag{152}$$

Then, from Lemma B.1, having set $a \leftarrow \widetilde{a}_k$ and $b \leftarrow \chi b_k$, we have:

$$\widetilde{P}_{t+1}(\chi) = \widetilde{a}_{k+1} + \chi b_{k+1} \leq \widetilde{a}_k + Bb_k + C(\chi b_k)^{\frac{2}{\psi}} - 2^{1-\frac{2}{\psi}} C\max\left\{0,\ \widetilde{a}_k + \chi b_k\right\}^{\frac{2}{\psi}} + V. \tag{153}$$

By choosing χ so that $\chi b_k \leq 1$, i.e., $\chi \leq 1/\max_{k\in[K]} b_k$, we have:

$$\widetilde{P}_{t+1}(\chi) = \widetilde{a}_{k+1} + \chi b_{k+1} \leq \widetilde{a}_k + Bb_k + C(\chi b_k)^{\frac{2}{\psi}} - 2^{1-\frac{2}{\psi}} C \max\left\{0,\ \widetilde{a}_k + \chi b_k\right\}^{\frac{2}{\psi}} + V \tag{154}$$

$$\leq \widetilde{a}_k + (B+\chi C)b_k - 2^{1-\frac{2}{\psi}} C \max\left\{0,\ \widetilde{a}_k + \chi b_k\right\}^{\frac{2}{\psi}} + V \tag{155}$$

$$= \widetilde{P}_t(B+\chi C) - 2^{1-\frac{2}{\psi}} C \max\left\{0,\ \widetilde{P}_t(\chi)\right\}^{\frac{2}{\psi}} + V. \tag{156}$$

To unfold the recursion, we need to ensure that $B + \chi C \leq \chi$, which leads to a condition relating the two learning rates:

$$B + \chi C \tag{157}$$

$$= \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) + \left(2\zeta_{\upsilon,k}\underbrace{\left(1+\frac{L_2}{2}\zeta_{\upsilon,k}\right)}_{\leq 3/2}\chi\underbrace{\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)}_{\leq 1}\right. \tag{158}$$

$$\left.+\frac{\zeta_{\upsilon,k}}{2}\left(1+\chi\underbrace{\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)}_{\leq 1}\right)\right)\frac{4L_3^2}{\omega} \tag{159}$$

$$+\chi\left(\underbrace{-2\zeta_{\upsilon,k}\left(1+\frac{L_2}{2}\zeta_{\upsilon,k}\right)\chi\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)}_{\leq 0}+\frac{\zeta_{\upsilon,k}}{2}\left(1+\chi\underbrace{\left(1-\frac{\zeta_{\lambda,k}}{2}\omega\right)}_{\leq 1}\right)\right)\alpha_1^{\frac{2}{\psi}} \tag{160}$$

$$\leq \chi - \chi\frac{\zeta_{\lambda,k}}{2}\omega + \zeta_{\upsilon,k}\left(\frac{2L_3^2}{\omega}(1+7\chi) + \frac{1+\chi}{2}\alpha_2^{\frac{2}{\psi}}\right) \leq \chi \tag{161}$$

$$\implies \zeta_{\upsilon,k} \leq \frac{\omega^2\chi\zeta_{\lambda,k}}{(1+\chi)\omega\alpha_1^{\frac{2}{\psi}} + 4L_3^2(1+7\chi)}, \tag{162}$$

where we exploited $\zeta_{\upsilon,k} \leq 1/L_2$ and $\zeta_{\lambda,k} \leq 2/\omega$. Thus, we have:

$$\widetilde{P}_{k+1}(\chi) \leq \widetilde{P}_k(\chi) - 2^{1-\frac{2}{\psi}} C \max\left\{0,\ \widetilde{P}_k(\chi)\right\}^{\frac{2}{\psi}} + V. \tag{163}$$

Collecting all conditions on the learning rates, we have:

$$\zeta_{\upsilon,k} \leq \min\left\{\frac{1}{L_H}, \frac{1}{L_2}, \frac{\omega^2\chi\zeta_{\lambda,k}}{(1+\chi)\omega\alpha_1^{\frac{2}{\psi}} + 4L_3^2(1+7\chi)}\right\}, \tag{164}$$

$$\zeta_{\lambda,k} \leq \min\left\{\frac{1}{\omega}, \frac{2}{\omega}\right\} = \frac{1}{\omega}. \tag{165}$$

As a further simplification, let us observe that:

$$C = \left(-2\zeta_{\upsilon,k} \underbrace{\left(1 + \frac{L_2}{2}\zeta_{\upsilon,k}\right)}_{\leq 3/2} \chi \left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right) + \frac{\zeta_{\upsilon,k}}{2} \left(1 + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\right) \right) \alpha_1^{\frac{2}{\psi}} \tag{166}$$

$$\geq \frac{\zeta_{\upsilon,k}}{2} \left(1 + 5\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)\chi\right) \alpha_1^{\frac{2}{\psi}} \geq \frac{\zeta_{\upsilon,k}\alpha_1^{\frac{2}{\psi}}}{2}. \tag{167}$$

$$V = \frac{\zeta_{\upsilon,k}^2}{2} \left(L_H + \chi\left(1 - \frac{\zeta_{\lambda,k}}{2}\omega\right)(L_H + L_2)\right) V_\upsilon + \chi \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda \tag{168}$$

$$\leq \frac{\zeta_{\upsilon,k}^2}{2} \left((1 + 2\chi)L_2 + (1+\chi)\frac{L_1^2}{\omega}\right) V_\upsilon + \chi \frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda =: \widetilde{V}. \tag{169}$$

Denoting with $\widetilde{C} =: 2^{1-\frac{1}{\psi}} \frac{\zeta_{\upsilon,k}\alpha_1^{\frac{2}{\psi}}}{2}$, we are going to study the recurrence:

$$\widetilde{P}_{k+1}(\chi) \leq \widetilde{P}_k(\chi) - \widetilde{C} \max\left\{0,\ \widetilde{P}_k(\chi)\right\}^{\frac{2}{\psi}} + \widetilde{V}. \tag{170}$$

Part V: Rates Computation

Part V(a): Exact gradients We consider the case $\widetilde{V} = 0$. Let us start with $\psi = 2$. From Lemma C.3, we have:

$$\widetilde{P}_K(\xi) \leq \left(1 - \widetilde{C}\right)^K \widetilde{P}_0(\xi) \leq \epsilon \tag{171}$$

$$\implies K \leq \frac{\log \frac{\widetilde{P}_0(\xi)}{\epsilon}}{\log \frac{1}{1-\widetilde{C}}} \leq \widetilde{C}^{-1} \log \frac{\widetilde{P}_0(\xi)}{\epsilon} = \frac{2 \log \frac{\widetilde{P}_0(\xi)}{\epsilon}}{2^{1-\frac{1}{\psi}} \zeta_{\rho,t} \alpha_1^{\frac{2}{\psi}}} \tag{172}$$

The inequality on K holds under the conditions:

$$\widetilde{C} \leq \frac{2}{\psi \widetilde{P}_0(\chi)^{\frac{2}{\psi}-1}} \implies \zeta_{\upsilon,k} \leq \frac{2^{1+\frac{2}{\psi}}}{\psi \alpha_1^{\frac{2}{\psi}} \widetilde{P}_0(\chi)^{\frac{2}{\psi}-1}}, \tag{173}$$

$$\zeta_{\upsilon,k} \leq \min\left\{ \frac{1}{L_H}, \frac{1}{L_2}, \frac{\omega^2 \chi \zeta_{\lambda,k}}{(1+\chi)\omega\alpha_1^{\frac{2}{\psi}} + 4L_3^2(1+7\chi)} \right\} \tag{174}$$

$$= \min\left\{ \frac{1}{L_2 + \frac{L_1^2}{\omega}}, \frac{\omega^2 \chi \zeta_{\lambda,k}}{(1+\chi)\omega\alpha_1^{\frac{2}{\psi}} + 4L_3^2(1+7\chi)} \right\}, \tag{175}$$

$$\zeta_{\lambda,k} \leq \frac{1}{\omega}, \tag{176}$$

where the first one derives from the hypothesis of Lemma C.3 and the other two from the conditions on the learning rates derived in the previous parts. We set:

$$\zeta_{\lambda,k} = \omega^{-1},$$
$$\zeta_{\upsilon,k} = \min\left\{ \frac{2^{1+\frac{2}{\psi}}}{\psi \alpha_1^{\frac{2}{\psi}} \widetilde{P}_0(\chi)^{\frac{2}{\psi}-1}}, \frac{1}{L_2 + \frac{L_1^2}{\omega}}, \frac{\omega\chi}{(1+\chi)\omega\alpha_1^{\frac{2}{\psi}} + 4L_3^2(1+7\chi)} \right\} = \mathcal{O}(\omega).$$

Thus, the sample complexity becomes $K = \mathcal{O}\left(\omega^{-1} \log \frac{1}{\epsilon}\right)$.

Consider now $\psi \in [1, 2)$. We have from Lemma C.3:

$$\widetilde{P}_K(\chi) \leq \left(\left(\frac{2}{\psi} - 1 \right) \widetilde{C} K \right)^{-\frac{\psi}{2-\psi}} \leq \epsilon \tag{177}$$
$$\implies K \leq \frac{\psi}{2-\psi} \widetilde{C}^{-1} \epsilon^{-\frac{2}{\psi}+1} = \frac{2\psi}{(2-\psi) 2^{1-\frac{1}{\psi}} \zeta_{\upsilon,k} \alpha_1^{\frac{2}{\psi}}} \epsilon^{-\frac{2}{\psi}+1}, \tag{178}$$

holding under the same conditions as before. With the same choices of learning rates, we obtain the sample complexity $K = \mathcal{O}\left(\omega^{-1}\epsilon^{-\frac{2}{\psi}+1}\right)$ as sample complexity.

Part V(b): Estimated gradients We consider $\widetilde{V} > 0$. In this case, from Lemma C.5, we have:

$$\widetilde{P}_K(\chi) \leq \left(1 - \widetilde{C}^{1-\frac{\psi}{2}} \widetilde{V}^{\frac{\psi}{2}}\right)^K \widetilde{P}_0(\chi) + \left(\frac{\widetilde{V}}{\widetilde{C}}\right)^{\frac{\psi}{2}}. \tag{179}$$

We enforce both terms to be smaller or equal to $\epsilon/2$. With the first one, we can evaluate the sample complexity:

$$\left(1 - \widetilde{V}^{1-\frac{\psi}{2}} \widetilde{C}^{\frac{\psi}{2}}\right)^K \widetilde{P}_0(\chi) \leq \frac{\epsilon}{2} \tag{180}$$
$$\implies K \leq \frac{\log \frac{2\widetilde{P}_0(\chi)}{\epsilon}}{\widetilde{V}^{1-\frac{\psi}{2}} \widetilde{C}^{\frac{\psi}{2}}} \tag{181}$$
$$= \frac{\log \frac{2\widetilde{P}_0(\chi)}{\epsilon}}{\left(\frac{\zeta_{\upsilon,k}^2}{2}\left((1+2\chi)L_2 + (1+\chi)\frac{L_1^2}{\omega}\right)V_\upsilon + \chi\frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda\right)^{1-\frac{\psi}{2}} \left(2^{1-\frac{1}{\psi}} \frac{\zeta_{\upsilon,k}\alpha_1^{\frac{2}{\psi}}}{2}\right)^{\frac{\psi}{2}}} \tag{182}$$

Regarding the second one, we have:

$$\left(\frac{\widetilde{V}}{\widetilde{C}}\right)^{\frac{\psi}{2}} \leq \frac{\epsilon}{2} \implies \left(\frac{\frac{\zeta_{\upsilon,k}^2}{2}\left((1+2\chi)L_2 + (1+\chi)\frac{L_1^2}{\omega}\right)V_\upsilon + \chi\frac{\omega}{2}\zeta_{\lambda,k}^2 V_\lambda}{2^{1-\frac{1}{\psi}} \frac{\zeta_{\upsilon,k}\alpha_1^{\frac{2}{\psi}}}{2}} \right)^{\frac{\psi}{2}} \leq \frac{\epsilon}{2} \tag{183}$$

By enforcing the relation between the two learning rates, we set $\zeta_{\upsilon,k} = \mathcal{O}(\omega^2\zeta_{\lambda,k})$. By enforcing the previous inequality, recalling that $L_2 \leq \mathcal{O}(\omega^{-1})$ and $V_\upsilon \leq \mathcal{O}(\omega^{-2})$, we obtain $\zeta_\lambda = \mathcal{O}(\omega\epsilon^{2/\psi})$, from which $\zeta_\upsilon = \mathcal{O}(\omega^3\epsilon^{2/\psi})$. Substituting these values into the sample complexity upper bound, we get (highlighting the terms possibly depending on ω):

$$K \leq \mathcal{O}\left(\frac{\log \frac{1}{\epsilon}}{((L_2 + \omega^{-1})V_{\upsilon}\zeta_{\upsilon}^2 + \omega\zeta_{\lambda}^2)^{1-\psi/2}\zeta_{\upsilon}^{\psi/2}}\right) \tag{184}$$

$$= \mathcal{O}\left(\frac{\log \frac{1}{\epsilon}}{((L_2 + \omega^{-1})V_{\upsilon}(\omega^3\epsilon^{2/\psi})^2 + \omega(\omega\epsilon^{2/\psi})^2)^{1-\psi/2}(\omega^3\epsilon^{2/\psi})^{\psi/2}}\right) \tag{185}$$

$$\leq \mathcal{O}\left(\frac{\log \frac{1}{\epsilon}}{\omega^3\epsilon^{4/\psi-1}}\right), \tag{186}$$

having bounded the sum at the denominator with the second addendum. □

Theorem 4.6 (Sample Complexity for Deterministic Deployment)**.** *Suppose to run* C-PG *for K iterations employing a (hyper)policy complying with Definitions 4.2 (AB) or 4.3 (PB). Suppose to be under Assumptions 4.1 (PB) or 4.2 (AB), 4.3, 4.4 (AB), and 4.5. For $\psi \in [1,2]$, $\chi < 1/5$, sufficiently small ϵ and ω, and a choice of* constant *learning rates $\zeta_\lambda = \mathcal{O}(\omega\sigma^2\epsilon^{2/\psi})$ and $\zeta_\theta = \omega\zeta_\lambda$, whenever $K = \mathcal{O}(\omega^{-3}\sigma^{-2}\epsilon^{-\frac{4}{\psi}+1})$ and the gradients are estimated, we have that:*

$$\mathcal{P}_{D,K}(\chi) \leq \epsilon + \frac{\beta_\dagger(\sigma,\psi)}{\alpha_D} + 4(1+\Lambda_{\max})L_{1\dagger}\sigma\sqrt{d_\dagger},$$

where $\beta_\dagger(\sigma,\psi)$ is quantified in Theorem 4.2, $\Lambda_{\max} := \omega^{-1}UJ_{\max}$, $L_{1P} := L_{1D,\max}$, $L_{1A} := L_{1\underline{\mu},\max}$, $d_P := d_\Theta$, and $d_A := d_{\mathcal{A}}$.

Proof. Under the considered set of assumptions, we recover the results of Theorems 4.2, 4.3, and 4.5 and of Lemma 4.4, matching the conditions needed to establish the sample complexity exhibited in Theorem 3.2 for ensuring that $\mathcal{P}_{\dagger,K}(\chi) \leq \epsilon + \frac{\beta_\dagger(\sigma.\psi)}{\alpha_D}$, where we employed the coefficients of the inherited weak ψ-GD (Theorem 4.2).

In particular, recovering "Part V: Rates Computation" of the proof of Theorem 3.2 with $\psi \in [1,2]$ and inexact gradients, to compute the sample complexity needed to ensure last-iterate global convergence of $\mathcal{P}_{\dagger,K}(\chi)$, we have to ensure the following conditions:

$$(i).\;\; K \leq \frac{\log \frac{2\widetilde{P}_{\dagger,0}(\chi)}{\epsilon}}{\left(\frac{\zeta_\theta^2}{2}\left((1+2\chi)L_2+(1+\chi)\frac{L_1^2}{\omega}\right)V_{\dagger,\theta}+\chi\frac{\omega}{2}\zeta_\lambda^2V_\lambda\right)^{1-\frac{\psi}{2}}\left(2^{1-\frac{1}{\psi}}\frac{\zeta_\theta\alpha_{\mathrm{D}}^{\frac{2}{\psi}}}{2}\right)^{\frac{\psi}{2}}},$$

$$(ii).\;\; \left(\frac{\frac{\zeta_\theta^2}{2}\left((1+2\chi)L_2+(1+\chi)\frac{L_1^2}{\omega}\right)V_{\dagger,\theta}+\chi\frac{\omega}{2}\zeta_\lambda^2V_\lambda}{2^{1-\frac{1}{\psi}}\frac{\zeta_\theta\alpha_{\mathrm{D}}^{\frac{2}{\psi}}}{2}}\right)^{\frac{\psi}{2}} \leq \frac{\epsilon}{2},$$

where $L_1 = \mathcal{O}(1)$ and $L_2 = \mathcal{O}(\omega^{-1})$ are quantified in Theorem 4.3, while $V_{\dagger,\theta} = \mathcal{O}(\omega^{-2}\sigma^{-2})$ and $V_\lambda = \mathcal{O}(1)$ are quantified in Lemma 4.4.

Now, considering (ii), by enforcing the relation between the two learning rates specified in Theorem 3.2, we can set $\zeta_\theta = \mathcal{O}(\omega^2\zeta_\lambda)$. Exploiting the characterization of L_1 and L_2 (Theorem 4.3) and the one of $V_{\dagger,\theta}$ and V_λ (Lemma 4.4), we have the following:

$$\left(\frac{\frac{\zeta_\theta^2}{2}\left((1+2\chi)L_2+(1+\chi)\frac{L_1^2}{\omega}\right)V_{\dagger,\theta}+\chi\frac{\omega}{2}\zeta_\lambda^2 V_\lambda}{2^{1-\frac{1}{\psi}}\frac{\zeta_\theta\alpha_{\mathrm{D}}^{\frac{2}{\psi}}}{2}}\right)^{\frac{\psi}{2}} = \mathcal{O}\left(\frac{\omega^4\zeta_\lambda^2\left(\omega^{-1}+\omega^{-1}\right)\omega^{-2}\sigma^{-2}+\omega\zeta_\lambda^2}{\omega^2\zeta_\lambda}\right)^{\frac{\psi}{2}}$$

$$= \mathcal{O}\left(\frac{\omega\zeta_\lambda^2\sigma^{-2}+\omega\zeta_\lambda^2}{\omega^2\zeta_\lambda}\right)^{\frac{\psi}{2}}$$

$$= \mathcal{O}\left(\zeta_\lambda\omega^{-1}\sigma^{-2}\right)^{\frac{\psi}{2}}.$$

Thus, enforcing $\mathcal{O}\left(\zeta_\lambda\omega^{-1}\sigma^{-2}\right)^{\frac{\psi}{2}} \le \frac{\epsilon}{2}$, we have that $\zeta_\lambda = \mathcal{O}(\omega\sigma^2\epsilon^{2/\psi})$, thus implying $\zeta_\theta = \mathcal{O}(\omega^3\sigma^2\epsilon^{2/\psi})$.

If we now consider the term (i), by exploiting the form of L_1, L_2, $V_{\dagger,\theta}$, and V_λ and the choice of ζ_λ and ζ_θ, we have the following:

$$\begin{aligned}
K &\le \frac{\log\frac{2\widetilde{P}_{\dagger,0}(\chi)}{\epsilon}}{\left(\frac{\zeta_\theta^2}{2}\left((1+2\chi)L_2+(1+\chi)\frac{L_1^2}{\omega}\right)V_{\dagger,\theta}+\chi\frac{\omega}{2}\zeta_\lambda^2 V_\lambda\right)^{1-\frac{\psi}{2}}\left(2^{1-\frac{1}{\psi}}\frac{\zeta_\theta\alpha_{\mathrm{D}}^{\frac{2}{\psi}}}{2}\right)^{\frac{\psi}{2}}} \\
&= \mathcal{O}\left(\frac{\log\frac{1}{\epsilon}}{\left(\omega\zeta_\lambda^2\sigma^{-2}\right)^{1-\psi/2}\left(\omega^2\zeta_\lambda\right)^{\psi/2}}\right) \\
&= \mathcal{O}\left(\frac{\log\frac{1}{\epsilon}}{\omega^{1+\psi/2}\zeta_\lambda^{2-\psi/2}\sigma^{\psi-2}}\right) \\
&= \mathcal{O}\left(\frac{\log\frac{1}{\epsilon}}{\omega^3\sigma^2\epsilon^{-1+4/\psi}}\right),
\end{aligned}$$

where in the last line, we exploited the fact that $\zeta_\lambda = \mathcal{O}(\omega\sigma^2\epsilon^{2/\psi})$.

This iteration complexity, which naturally translates into a sample complexity given that we can employ a constant batch size N, ensures that $\mathcal{P}_{\dagger,K}(\chi) \le \epsilon + \frac{\beta_\dagger(\sigma,\psi)}{\alpha_{\mathrm{D}}}$. By leveraging the result of Theorem 4.5, we have that the same sample complexity ensures that:

$$\mathcal{P}_{\mathrm{D},K}(\chi) \le \epsilon + \frac{\beta_\dagger(\sigma,\psi)}{\alpha_{\mathrm{D}}} + 4(1+\Lambda_{\max})L_{1\dagger}\sigma\sqrt{d_\dagger}.$$

□

B. Technical Lemmas

Lemma B.1. *Let $a \in \mathbb{R}$, $b \ge 0$, and $\psi \in [1,2]$. It holds that:*

$$\max\{0,a\}^{\frac{2}{\psi}} \ge 2^{1-\frac{2}{\psi}}\max\{0,a+b\}^{\frac{2}{\psi}} - b^{\frac{2}{\psi}}. \tag{187}$$

Proof. Let us consider the following derivation:

$$\max\{0,a\}^{\frac{2}{\psi}} = \begin{cases} a^{\frac{2}{\psi}} & \text{if } a>0 \\ 0 & \text{otherwise} \end{cases} \tag{188}$$

$$\geq \begin{cases} 2^{1-\frac{2}{\psi}}(a+b)^{\frac{2}{\psi}} - b^{\frac{2}{\psi}} & \text{if } a>0 \\ 0 & \text{otherwise} \end{cases} \tag{189}$$

$$= \begin{cases} 2^{1-\frac{2}{\psi}}(a+b)^{\frac{2}{\psi}} - b^{\frac{2}{\psi}} & \text{if } a>0 \\ 0 & \text{if } -b<a\leq 0 \\ 0 & \text{otherwise} \end{cases} \tag{190}$$

$$\geq \begin{cases} 2^{1-\frac{2}{\psi}}(a+b)^{\frac{2}{\psi}} - b^{\frac{2}{\psi}} & \text{if } a>0 \\ 2^{1-\frac{2}{\psi}}(a+b)^{\frac{2}{\psi}} - b^{\frac{2}{\psi}} & \text{if } -b<a\leq 0 \\ -b^{\frac{2}{\psi}} & \text{otherwise} \end{cases} \tag{191}$$

$$= \begin{cases} 2^{1-\frac{2}{\psi}}(a+b)^{\frac{2}{\psi}} - b^{\frac{2}{\psi}} & \text{if } a+b>0 \\ -b^{\frac{2}{\psi}} & \text{otherwise} \end{cases} \tag{192}$$

$$= 2^{1-\frac{2}{\psi}}\max\{0,a+b\}^{\frac{2}{\psi}} - b^{\frac{2}{\psi}}, \tag{193}$$

$$\tag{194}$$

where the first inequality follows from $(x+y)^{\frac{2}{\psi}} \leq 2^{\frac{2}{\psi}-1}(x^{\frac{2}{\psi}} + y^{\frac{2}{\psi}})$ for $x,y\geq 0$, from Holder's inequality; the second inequality from observing that $2^{1-\frac{2}{\psi}}(a+b)^{\frac{2}{\psi}} - b^{\frac{2}{\psi}} \leq (2^{1-\frac{2}{\psi}}-1)b^{\frac{2}{\psi}} \leq 0$ for $-b<a\leq 0$. □

C. Recurrences

In this section, we provide auxiliary results about convergence rate of a certain class of recurrences that will be employed for the convergence analysis of the proposed algorithms. Specifically, we study the recurrence:

$$r_{k+1} \leq r_k - a\max\{0,r_k\}^{\phi} + b \tag{195}$$

for $a>0$, $b\geq 0$, and $\phi\in[1,2]$. To this end, we consider the helper sequence:

$$\begin{cases} \rho_0 = r_0 \\ \rho_{k+1} = \rho_k - a\max\{0,\rho_k\}^{\phi} + b \end{cases} \tag{196}$$

The line of the proof follows that of Montenegro et al. (2024a). Let us start showing that for sufficiently small a, the sequence ρ_k upper bounds r_k.

Lemma C.1. *If $a \leq \frac{1}{\phi\rho_k^{\phi-1}}$ for every $k\geq 0$, then, $r_k \leq \rho_k$ for every $k \geq 0$.*

Proof. By induction on k. For $k = 0$, the statement holds since $\rho_0 = r_0$. Suppose the statement holds for every $j \le k$, we prove that it holds for $k + 1$:

$$\rho_{k+1} = \rho_k - a \max\{0, \rho_k\}^\phi + b \tag{197}$$
$$\ge r_k - a \max\{0, r_k\}^\phi + b \tag{198}$$
$$\ge r_{k+1}, \tag{199}$$

where the first inequality holds by the inductive hypothesis and by observing that the function $f(x) = x - a \max\{0, x\}^\phi$ is non-decreasing in x when $a \le \frac{1}{\phi \rho_k^{\phi-1}}$. Indeed, if $x < 0$, then $f(x) = x$, which is non-decreasing; if $x \ge 0$, we have $f(x) = x - ax^\phi$, that can be proved to be non-decreasing in the interval $\left[0, (a\phi)^{-\frac{1}{\phi-1}}\right]$ simply by studying the sign of the derivative. Thus, we enforce the following requirement to ensure that ρ_k falls in the non-decreasing region:

$$\rho_k \le (a\phi)^{-\frac{1}{\phi-1}} \implies a \le \frac{1}{\phi \rho_k^{\phi-1}}. \tag{200}$$

So does r_k by the inductive hypothesis. □

Thus, from now on, we study the properties of the sequence ρ_k. Let us note that, if ρ_k is convergent, then it converges to the fixed-point $\overline{\rho}$ computed as follows:

$$\overline{\rho} = \overline{\rho} - a \max\{0, \overline{\rho}\}^\phi + b \implies \overline{\rho} = \left(\frac{b}{a}\right)^{\frac{1}{\phi}}, \tag{201}$$

having retained the positive solution of the equation only, since the negative one never attains the maximum $\max\{0, \overline{\rho}\}$. Let us now study the monotonicity properties of the sequence ρ_k.

Lemma C.2. *The following statements hold:*

- *If* $r_0 > \overline{\rho}$ *and* $a \le \frac{1}{\phi r_0^{\phi-1}}$*, then for every* $k \ge 0$ *it holds that:* $\overline{\rho} \le \rho_{k+1} \le \rho_k$*.*
- *If* $r_0 < \overline{\rho}$ *and* $a \le \frac{1}{\phi \overline{\rho}^{\phi-1}}$*, then for every* $k \ge 0$ *it holds that:* $\overline{\rho} \ge \rho_{k+1} \ge \rho_k$*.*

Proof. The proof is analogous to that of (Montenegro et al., 2024a, Lemma F.3). □

From now on, we focus on the case in which $r_0 \ge \overline{\rho}$, since, as we shall see later, the opposite case is irrelevant for the convergence guarantees. We now consider two cases: $b = 0$ and $b > 0$.

C.1. Analysis when $b = 0$

From the policy optimization perspective, this case corresponds to the one in which the gradients are exact (no variance). Recall that here $\overline{\rho} = 0$. We have the following convergence result.

Lemma C.3. *If* $a \le \frac{1}{\phi r_0^{\phi-1}}$, $r_0 \ge 0$*, and* $b = 0$ *it holds that:*

$$\rho_{k+1} \le \begin{cases} (1-a)^{k+1} r_0 & \text{if } \phi = 1 \\ \min\left\{ r_0, ((\phi-1)a(k+1))^{-\frac{1}{\phi-1}} \right\} & \text{if } \phi \in (1, 2] \end{cases}. \tag{202}$$

Proof. Since $r_0 \ge 0 = \overline{\rho}$, from Lemma C.2, we know that $\rho_k \ge 0$ and, thus, $\max\{0, \rho_k\} = \rho_k$. For $\phi = 1$, we have:

$$\rho_{k+1} = \rho_k - a\rho_k = (1-a)\rho_k = (1-a)^{k+1}\rho_0 = (1-a)^{k+1} r_0. \tag{203}$$

For $\phi \in (1, 2]$, we have:

$$\rho_{k+1} = \rho_k - a\rho_k^{\phi}. \tag{204}$$

We proceed by induction. For $k = 0$, the statement hold since $\rho_0 = r_0$ and $r_0 \leq (\phi a)^{-\frac{1}{\psi-1}} \leq ((\phi-1)a)^{-\frac{1}{\psi-1}}$ from the condition on the learning rate. Suppose the thesis holds for $j \leq k$, we prove it for $k+1$. $\rho_{k+1} \leq r_0$ by monotonicity, and, from the inductive hypothesis:

$$\rho_{k+1} = \rho_k - a\rho_k^{\phi} \leq (\phi a k)^{-\frac{1}{\phi-1}} - a(\phi a k)^{-\frac{\phi}{\phi-1}} \tag{205}$$

$$= \underbrace{(\phi a k)^{-\frac{1}{\phi-1}} - (\phi a(k+1))^{-\frac{1}{\phi-1}} - a(\phi a k)^{-\frac{\phi}{\phi-1}}}_{(*)} + (\phi a(k+1))^{-\frac{1}{\phi-1}}. \tag{206}$$

We now prove that $(*)$ is non-positive:

$$(*) = ((\phi-1)ak)^{-\frac{1}{\phi-1}} - ((\phi-1)a(k+1))^{-\frac{1}{\phi-1}} - a((\phi-1)ak)^{-\frac{\phi}{\phi-1}} \tag{207}$$

$$= ((\phi-1)a)^{-\frac{1}{\phi-1}} k^{-\frac{\phi}{\phi-1}} \underbrace{\left(k - (k+1)\left(\frac{k}{k+1}\right)^{\frac{\phi}{\phi-1}}\right)}_{\leq \frac{1}{\phi-1}} - a^{-\frac{1}{\phi-1}}((\phi-1)k)^{-\frac{\phi}{\phi-1}} \tag{208}$$

$$\leq a^{-\frac{1}{\phi-1}} k^{-\frac{\phi}{\phi-1}} (\phi-1)^{-\frac{1}{\phi-1}} \left(\frac{1}{\phi-1} - \frac{1}{\phi-1}\right) \leq 0, \tag{209}$$

having observed that:

$$\sup_{k\geq 1}\left(k - (k+1)\left(\frac{k}{k+1}\right)^{\frac{\phi}{\phi-1}}\right) = \lim_{k\to+\infty}\left(k - (k+1)\left(\frac{k}{k+1}\right)^{\frac{\phi}{\phi-1}}\right) = \frac{1}{\phi-1}. \tag{210}$$

□

C.2. Analysis for $b > 0$

From the policy optimization perspective, this corresponds to the case in which the gradients are estimated, i.e., the variance is positive. In this case, we proceed considering the helper sequence:

$$\begin{cases} \eta_0 = \rho_0 \\ \eta_{k+1} = \left(1 - a\overline{\rho}^{\phi-1}\right)\eta_k + b & \text{if } k \geq 0 \end{cases}. \tag{211}$$

We show that the sequence η_k upper bounds ρ_k when $\rho_0 = r_0 \geq \overline{\rho}$.

Lemma C.4. *If $r_0 > \overline{\rho}$ and $a \leq \frac{1}{\phi r_0^{\phi-1}}$, then, for every $k \geq 0$, it holds that $\eta_k \geq \rho_k$.*

Proof. The proof is analogous to that of (Montenegro et al., 2024a, Lemma F.4). □

Thus, we can provide the convergence guarantee.

Lemma C.5. *If $a \leq \frac{1}{\phi r_0^{\phi-1}}$, $r_0 \geq 0$, and $b > 0$ it holds that:*

$$\eta_{k+1} \leq \left(1 - b^{1-\frac{1}{\phi}} a^{\frac{1}{\phi}}\right)^{k+1} + \left(\frac{b}{a}\right)^{\frac{1}{\phi}}. \tag{212}$$

Proof. By unrolling the recursion:

$$\eta_{k+1} = \left(1 - a\overline{\rho}^{\phi-1}\right)\eta_k + b \tag{213}$$

$$= \left(1 - a\overline{\rho}^{\phi-1}\right)^{k+1} r_0 + b\sum_{j=0}^{k}\left(1 - a\overline{\rho}^{\phi-1}\right)^j \tag{214}$$

$$\leq \left(1 - a\overline{\rho}^{\phi-1}\right)^{k+1} r_0 + b\sum_{j=0}^{+\infty}\left(1 - a\overline{\rho}^{\phi-1}\right)^j \tag{215}$$

$$= \left(1 - b^{1-\frac{1}{\phi}}a^{\frac{1}{\phi}}\right)^{k+1} + \frac{b}{a\overline{\rho}^{\phi-1}} \tag{216}$$

$$= \left(1 - b^{1-\frac{1}{\phi}}a^{\frac{1}{\phi}}\right)^{k+1} + \left(\frac{b}{a}\right)^{\frac{1}{\phi}}. \tag{217}$$

□

D. Experimental Details

D.1. Employed Policies and Hyperpolicies

Linear Gaussian Policy. A *linear parametric gaussian* policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{A})$ with variance σ^2 samples the actions as $a_t \sim \mathcal{N}(\boldsymbol{\theta}^\top \boldsymbol{s}_t, \sigma^2 \boldsymbol{I}_{d_S})$, where $\boldsymbol{s}_t$ is the observed state at time t and $\boldsymbol{\theta}$ is the parameter vector.

Tabular Softmax Policy. A *tabular softmax* policy $\pi_{\boldsymbol{\theta}} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{A})$ with a constant temperature τ is such that:

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_j|\boldsymbol{s}_i) = \frac{\exp\left(\frac{\theta_{i,j}}{\tau}\right)}{\sum_{z=1}^{|\mathcal{A}|}\exp\left(\frac{\theta_{i,z}}{\tau}\right)},$$

where $\boldsymbol{\theta}_{i,j}$ is the parameter associated with the i-th state and the j-th action. Notice that the total number of parameters for this kind of policy is $|\mathcal{S}||\mathcal{A}|$.

Linear Deterministic Policy. A *linear parametric deterministic* policy $\mu_{\boldsymbol{\theta}} : \mathcal{S} \to \mathcal{A}$ samples the actions as $\boldsymbol{a}_t = \boldsymbol{\theta}^\top \boldsymbol{s}_t$, where $\boldsymbol{s}_t$ is the observed state at time t and $\boldsymbol{\theta}$ is the parameter vector.

Gaussian Hyperpolicy. A *parametric gaussian* hyperpolicy $\nu_{\boldsymbol{\rho}} \in \Delta(\Theta)$ with variance σ^2 samples the parameters $\boldsymbol{\theta}$ for the underlying generic parametric policy $\pi_{\boldsymbol{\theta}}$ as $\boldsymbol{\theta}_t \sim \mathcal{N}(\boldsymbol{\rho}, \sigma^2 \boldsymbol{I}_{d_\Theta})$, where $\boldsymbol{\rho}$ is the parameter vector for the hyperpolicy.

D.2. Environments

Discrete Grid World with Walls. Discrete Grid World with Walls (DGWW) is a simple discrete environment we employed to compare C-PGAE against the sample-based versions of NPG-PD (Ding et al., 2020, Appendix H) and RPG-PD (Ding et al., 2024, Appendix C.9). DGWW is a grid-like bidimensional environment in which an agent can assume only integer coordinate positions and in which an agent can play four actions stating whether to go up, right, left, or down. The goal is to reach the center of the grid performing the minimum amount of steps, begin the initial state uniformly sampled among the four vertices of the grid. The agent is rewarded negatively and proportionally to its distance from the center, where the reward is 0. Around the goal state there is a "U-shaped" obstacle with an opening on the top side. In particular, when the agent lands in a state in which the wall is present, it receives a cost of 1, otherwise the cost signal is always equal to 0. In our experiments, we employed a DGWW environment of such a kind, with $|\mathcal{S}| = 49$, i.e., with each dimension with length equal to 7.

Linear Quadratic Regulator with Costs. The Linear Quadratic Regulator (LQR, Anderson and Moore, 2007) is a continuous environment we employed in the regularization sensitivity study of C-PGAE and C-PGPE, and in the comparison among the same algorithms against the sample-based version of NPG-PD2 (Ding et al., 2022, Algorithm 1) and its ridge-regularized version RPG-PD2 (not provided by the authors, but designed by us). LQR is a dynamical system governed by the following state evolution:

$$\boldsymbol{s}_{t+1} = A\boldsymbol{s}_t + B\boldsymbol{a}_t,$$

where $A \in \mathbb{R}^{d_S \times d_S}$ and $B \in \mathbb{R}^{d_S \times d_\mathcal{A}}$.

In the standard version of the environment, the reward is computed at each step as:

$$r_t = -\boldsymbol{s}_t^\top R\boldsymbol{s}_t - \boldsymbol{a}_t^\top Q\boldsymbol{a}_t,$$

where $R \in \mathbb{R}^{d_S \times d_S}$ and $Q \in \mathbb{R}^{d_\mathcal{A} \times d_\mathcal{A}}$.

We modified this version of the LQR environment introducing costs. In particular, in our *CostLQR*, the state evolution is treated as in the original case, while the reward at step t is computed as:

$$r_t = -\boldsymbol{s}_t^\top R\boldsymbol{s}_t,$$

where $R \in \mathbb{R}^{d_S \times d_S}$. Moreover, we added a cost signal c which is computed as follows at every time step t:

$$c_t = \boldsymbol{a}_t^\top Q\boldsymbol{a}_t,$$

where $Q \in \mathbb{R}^{d_\mathcal{A} \times d_\mathcal{A}}$.

In our experiments, we consider a *CostLQR* environment whose main characteristics are reported in Table 4.

Additionally, we considered a uniform initial state distribution in $[-3, 3]$ and the following matrices:

$$A = B = 0.9\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}, \qquad Q = \begin{bmatrix}0.9 & 0\\ 0 & 0.1\end{bmatrix}, \qquad R = \begin{bmatrix}0.1 & 0\\ 0 & 0.9\end{bmatrix}.$$

MuJoCo with Costs. For our experiments on risk minimization, we utilized environments from the MuJoCo control suite (Todorov et al., 2012), which offers a variety of continuous control environments. To tailor these environments to our specific requirements, we introduced a cost function that represents the energy associated with the control actions. In standard MuJoCo environments, a portion of the reward is typically calculated as the cost of the control action, which is proportional to the deviation of the chosen action from predefined action bounds. In our MuJoCo modification, at each time step, we make the environment return a cost computed as:

$$\left\|\boldsymbol{a}_t - \min\left\{\max\left\{\boldsymbol{a}_t, a_{\min}\right\}, a_{\max}\right\}\right\|_2,$$

where $a_{\min}$ and $a_{\max}$ are, respectively, the bounds for the minimum and maximum value for each component of the action vector. Then, the action $\min\left\{\max\left\{\boldsymbol{a}_t, a_{\min}\right\}, a_{\max}\right\}$ is passed to the environment. In our experiment, we consider *Swimmer-v4* and *Hopper-v4* MuJoCo environments, whose main features are summarized in Table 4.

RobotWorld. For the comparison among C-PG, AD-PGPD (Rozada et al., 2025), and PGDual (Zhao and You, 2021; Brunke et al., 2022), we employed the *Robot World* environment (Rozada et al., 2025). This environment is a modification of the *CostLQR* one in which both reward and cost functions are given by the following quadratic functions:

$$r(\boldsymbol{s}, \boldsymbol{a}) = \langle G_1; |\boldsymbol{s}|\rangle + \langle R_1; |\boldsymbol{a}|\rangle - \frac{1}{2}\|\boldsymbol{a}\|_2^2,$$
$$c(\boldsymbol{s}, \boldsymbol{a}) = \left\langle G_2; \boldsymbol{s}^2\right\rangle + \left\langle R_2; \boldsymbol{a}^2\right\rangle,$$

where for every $\boldsymbol{x} \in \mathbb{R}^n$ we define $|\boldsymbol{x}| := (|x_1|, \dots, |x_n|)^\top$ and $\boldsymbol{x}^2 := (x_1^2, \dots, x_n^2)^\top$. In the setting we considered for our experiments, we employed the following values for $\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{R}_1$, and $\boldsymbol{R}_2$:

$$G_1 = -(1, 1, 0.001, 0.001), \quad G_2 = -(0.001, 0.001, 1, 1), \quad \text{and} \quad R_1 = R_2 = -(0.01, 0.01).$$

Furthermore, differently from the usual LQR environment, in *RobotWorld* the agent is allowed to control both velocity and acceleration of the agent.

D.3. Experimental Details

Environment	**State Dim.** d_S	**Action Dim.** $d_{\mathcal{A}}$	**Action Range** $[a_{\min}, a_{\max}]$	**State Range** $[s_{\min}, s_{\max}]$
CostLQR	2	2	$(-\infty, +\infty)$	$(-\infty, +\infty)$
Swimmer-v4	8	2	$[-1, 1]$	$(-\infty, +\infty)$
Hopper-v4	11	3	$[-1, 1]$	$(-\infty, +\infty)$
RobotWorld	4	2	$(-\infty, +\infty)$	$(-\infty, +\infty)$

Table 4
Main features of *CostLQR*, *Swimmer-v4*, *Hopper-v4*, and *RobotWorld*.

Details for the Comparison in *DGWW* Experiment	
Environment	*DGWW*
Horizon	$T = 100$
Policy	Tabular Softmax
Constraint Threshold	$b = 0.2$
Iterations	$K = 3000$
Batch Size	$N = 10$
Learning Rates C-PGAE	$\zeta_\theta = 0.01$ and $\zeta_\lambda = 0.1$
Learning Rates NPG-PD	$\zeta_\theta = 0.01$ and $\zeta_\lambda = 0.1$
Learning Rates RPG-PD	$\zeta_\theta = 0.01$ and $\zeta_\lambda = 0.01$
Regularization C-PGAE	$\omega = 10^{-4}$
Regularization RPG-PD	$\omega = 10^{-4}$

Table 5
Details for the comparison of C-PGAE against NPG-PD and RPG-PD in DGWW (Section 6.1).

Details for the Comparison in *CostLQR* Experiment	
Environment	Bidimensional *CostLQR*
Horizon	$T = 50$
(Hyper)Policy	Linear Gaussian with $\sigma^2 = 10^{-3}$
Constraint Threshold	$b = 0.9$
Iterations (C-PGPE and C-PGAE)	$K = 6000$
Iterations (NPG-PD2 and RPG-PD2)	$K = 1000$
Batch Size (C-PGPE and C-PGAE)	$N = 100$
Batch Size (NPG-PD2 and RPG-PD2)	$N = 600$
Learning Rate (Adam) C-PGPE	$\zeta_{\rho,0} = 10^{-3}$ and $\zeta_{\lambda,0} = 10^{-2}$
Learning Rates (Adam) C-PGAE	$\zeta_{\theta,0} = 10^{-3}$ and $\zeta_{\lambda,0} = 10^{-2}$
Learning Rates (Adam) NPG-PD2	$\zeta_{\theta,0} = 3 \cdot 10^{-3}$ and $\zeta_{\lambda,0} = 10^{-2}$
Learning Rates (Adam) RPG-PD2	$\zeta_{\theta,0} = 3 \cdot 10^{-3}$ and $\zeta_{\lambda,0} = 10^{-2}$
Regularization C-PGPE	$\omega = 10^{-4}$
Regularization C-PGAE	$\omega = 10^{-4}$
Regularization RPG-PD2	$\omega = 10^{-4}$

Table 6
Details for the comparison of C-PGPE and C-PGAE against NPG-PD2 and RPG-PD2 in a bidimensional *CostLQR* (Section 6.1).

Details for the Comparison in *RobotWorld* Experiment	
Environment	*RobotWorld*
Horizon	$T = 100$
Hyperpolicy (C-PGPE)	Linear Gaussian with $\sigma^2 = 10^{-6}$
Policy (C-PGAE)	Linear Gaussian with $\sigma^2 = 5 \cdot 10^{-2}$
Policy (AD-PGPD and PGDual)	Linear Deterministic
Constraint Threshold	$b = 1000$
Iterations (C-PGPE and C-PGAE)	$K = 10^3$
Iterations (AD-PGPD and PGDual)	$K = 4 \cdot 10^4$
Batch Size (C-PGPE and C-PGAE)	$N = 100$
Batch Size (AD-PGPD and PGDual)	$N = 400$
Learning Rates (Adam) C-PGPE	$\zeta_{\rho,0} = 5 \cdot 10^{-6}$ and $\zeta_{\lambda,0} = 5 \cdot 10^{-3}$
Learning Rates (Adam) C-PGAE	$\zeta_{\theta,0} = 5 \cdot 10^{-6}$ and $\zeta_{\lambda,0} = 10^{-4}$
Learning Rates (Adam) AD-PGPD	$\zeta_{\theta,0} = 10^{-5}$ and $\zeta_{\lambda,0} = 10^{-5}$
Learning Rates (Adam) PGDual	$\zeta_{\theta,0} = 10^{-4}$ and $\zeta_{\lambda,0} = 10^{-5}$
Regularization (C-PGPE and C-PGAE)	$\omega = 10^{-4}$
Regularization AD-PGPD	$\omega = 2 \cdot 10^{-1}$

Table 7
Details for the comparison of C-PGPE and C-PGAE against AD-PGPD and PGDual in the RobotWorld Environment (Section 6.1).

Details for the Deterministic Deployment in *CostSwimmer-v4*	
Environment	*CostSwimmer-v4*
Horizon	$T = 100$
(Hyper)policy (C-PGPE and C-PGAE)	Linear Gaussian with $\sigma^2 \in \{10^{-2}, 5 \cdot 10^{-2}, 10^{-1}, 5 \cdot 10^{-1}, 1\}$
Constraint Threshold	$b = 50$
Iterations (C-PGPE and C-PGAE)	$K = 3 \cdot 10^3$
Batch Size (C-PGPE and C-PGAE)	$N = 100$
Learning Rates (Adam) C-PGPE	$\zeta_{\rho,0} = 10^{-3}$ and $\zeta_{\lambda,0} = 10^{-2}$
Learning Rates (Adam) C-PGAE	$\zeta_{\theta,0} = 10^{-3}$ and $\zeta_{\lambda,0} = 10^{-2}$
Regularization (C-PGPE and C-PGAE)	$\omega = 10^{-4}$

Table 8
Details for the comparison of C-PGPE and C-PGAE in *CostSwimmer-v4* (Section 6.2).

Details for the Deterministic Deployment in *CostHopper-v4*	
Environment	*CostHopper-v4*
Horizon	$T = 100$
(Hyper)policy (C-PGPE and C-PGAE)	Linear Gaussian with $\sigma^2 \in \{10^{-2}, 5 \cdot 10^{-2}, 10^{-1}, 5 \cdot 10^{-1}, 1\}$
Constraint Threshold	$b = 50$
Iterations (C-PGPE and C-PGAE)	$K = 3 \cdot 10^3$
Batch Size (C-PGPE and C-PGAE)	$N = 100$
Learning Rates (Adam) C-PGPE	$\zeta_{\rho,0} = 10^{-2}$ and $\zeta_{\lambda,0} = 10^{-1}$
Learning Rates (Adam) C-PGAE	$\zeta_{\theta,0} = 10^{-2}$ and $\zeta_{\lambda,0} = 10^{-1}$
Regularization (C-PGPE and C-PGAE)	$\omega = 10^{-4}$

Table 9
Details for the comparison of C-PGPE and C-PGAE in *CostHopper-v4* (Section 6.2).

Details for the Regularization Sensitivity Study in *CostLQR*	
Environment	Bidimensional *CostLQR*
Horizon	$T = 50$
(Hyper)Policy	Linear Gaussian with $\sigma^2 = 10^{-3}$
Constraint Threshold	$b = 0.2$
Iterations	$K = 10^4$
Batch Size	$N = 100$
Learning Rate (Adam) C-PGPE	$\zeta_{\rho,0} = 10^{-3}$ and $\zeta_{\lambda,0} = 10^{-2}$
Learning Rates (Adam) C-PGAE	$\zeta_{\theta,0} = 10^{-3}$ and $\zeta_{\lambda,0} = 10^{-2}$
Regularization	$\omega \in \left\{0, 10^{-4}, 10^{-2}\right\} 10^{-4}$

Table 10
Details for the regularization sensitivity study of C-PGPE and C-PGAE in a bidimensional *CostLQR* (Section 6.3).

References

Achiam, J., Held, D., Tamar, A., Abbeel, P., 2017. Constrained policy optimization, in: Proceedings of the International Conference on Machine Learning (ICML), PMLR. pp. 22–31.

Altman, E., 1999. Constrained Markov Decision Processes. CRC Press.

Anderson, B.D., Moore, J.B., 2007. Optimal control: linear quadratic methods. Courier Corporation.

Azizzadenesheli, K., Yue, Y., Anandkumar, A., 2018. Policy gradient in partially observable environments: Approximation and convergence. arXiv preprint arXiv:1810.07900 .

Bai, Q., Bedi, A.S., Agarwal, M., Koppel, A., Aggarwal, V., 2022. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach, in: AAAI Conference on Artificial Intelligence, AAAI Press. pp. 3682–3689.

Bai, Q., Bedi, A.S., Aggarwal, V., 2023. Achieving zero constraint violation for constrained reinforcement learning via conservative natural policy gradient primal-dual algorithm, in: AAAI Conference on Artificial Intelligence, AAAI Press. pp. 6737–6744.

Baxter, J., Bartlett, P.L., 2001. Infinite-horizon policy-gradient estimation. Journal of Artificial Intelligence Research 15, 319–350.

Bertsekas, D.P., 2014. Constrained optimization and Lagrange multiplier methods. Academic press.

Bhandari, J., Russo, D., 2024. Global optimality guarantees for policy gradient methods. Operations Research .

Bhatnagar, S., Lakshmanan, K., 2012. An online actor-critic algorithm with function approximation for constrained markov decision processes. Journal of Optimization Theory and Applications 153, 688–708.

Brunke, L., Greeff, M., Hall, A.W., Yuan, Z., Zhou, S., Panerati, J., Schoellig, A.P., 2022. Safe learning in robotics: From learning-based control to safe reinforcement learning. Annual Review of Control, Robotics, and Autonomous Systems 5, 411–444.

Chow, Y., Ghavamzadeh, M., Janson, L., Pavone, M., 2017. Risk-constrained reinforcement learning with percentile risk criteria. Journal of Machine Learning Research 18, 6070–6120.

Chow, Y., Nachum, O., Duéñez-Guzmán, E.A., Ghavamzadeh, M., 2018. A lyapunov-based approach to safe reinforcement learning, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 8103–8112.

Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., Tassa, Y., 2018. Safe exploration in continuous action spaces. arXiv preprint arXiv:1801.08757 .

Deisenroth, M.P., Neumann, G., Peters, J., 2013. A survey on policy search for robotics. Foundations and Trends in Robotics 2, 1–142.

Ding, D., Wei, C.Y., Zhang, K., Ribeiro, A., 2024. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. Advances in Neural Information Processing Systems (NeurIPS) 36.

Ding, D., Wei, X., Yang, Z., Wang, Z., Jovanovic, M.R., 2021. Provably efficient safe exploration via primal-dual policy optimization, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR. pp. 3304–3312.

Ding, D., Zhang, K., Basar, T., Jovanovic, M., 2020. Natural policy gradient primal-dual method for constrained markov decision processes. Advances in Neural Information Processing Systems (NeurIPS) 33, 8378–8390.

Ding, D., Zhang, K., Duan, J., Başar, T., Jovanović, M.R., 2022. Convergence and sample complexity of natural policy gradient primal-dual methods for constrained mdps. arXiv preprint arXiv:2206.02346 .

Fatkhullin, I., Barakat, A., Kireeva, A., He, N., 2023. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies, in: Proceedings of the International Conference on Machine Learning (ICML), PMLR. pp. 9827–9869.

Fujimoto, S., van Hoof, H., Meger, D., 2018. Addressing function approximation error in actor-critic methods, in: Proceedings of the International Conference on Machine Learning (ICML), PMLR. pp. 1587–1596.

Ghavamzadeh, M., Engel, Y., 2006. Bayesian policy gradient algorithms. Advances in Neural Information Processing Systems (NIPS) 19.

Gladin, E., Lavrik-Karmazin, M., Zainullina, K., Rudenko, V., Gasnikov, A.V., Takác, M., 2023. Algorithm for constrained markov decision process with linear convergence, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR. pp. 11506–11533.

Gravell, B., Esfahani, P.M., Summers, T., 2020. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. IEEE Transactions on Automatic Control 66, 5283–5298.

Hsieh, Y., Iutzeler, F., Malick, J., Mertikopoulos, P., 2019. On the convergence of single-call stochastic extra-gradient methods, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 6936–6946.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: International Conference on Learning Representations (ICLR).

Likmeta, A., Metelli, A.M., Tirinzoni, A., Giol, R., Restelli, M., Romano, D., 2020. Combining reinforcement learning with rule-based controllers for transparent and general decision-making in autonomous driving. Robotics and Autonomous Systems 131, 103568.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N.M.O., Erez, T., Tassa, Y., Silver, D., Wierstra, D., 2015. Continuous control with deep reinforcement learning. CoRR abs/1509.02971.

Liu, T., Zhou, R., Kalathil, D., Kumar, P., Tian, C., 2021. Policy optimization for constrained mdps with provable fast global convergence. arXiv preprint arXiv:2111.00552 .

Liu, Y., Ding, J., Liu, X., 2020. IPO: interior-point policy optimization under constraints, in: AAAI Conference on Artificial Intelligence, AAAI Press. pp. 4940–4947.

Masiha, S., Salehkaleybar, S., He, N., Kiyavash, N., Thiran, P., 2022. Stochastic second-order methods improve best-known sample complexity of sgd for gradient-dominated functions. Advances in Neural Information Processing Systems (NeurIPS) 35, 10862–10875.

Mei, J., Xiao, C., Szepesvari, C., Schuurmans, D., 2020. On the global convergence rates of softmax policy gradient methods, in: Proceedings of the International Conference on Machine Learning (ICML), PMLR. pp. 6820–6829.

Metelli, A.M., Papini, M., Faccio, F., Restelli, M., 2018. Policy optimization via importance sampling, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 5447–5459.

Mondal, W.U., Aggarwal, V., 2024. Last-iterate convergence of general parameterized policies in constrained mdps. arXiv preprint arXiv:2408.11513 .

Montenegro, A., Mussi, M., Metelli, A.M., Papini, M., 2024a. Learning optimal deterministic policies with stochastic policy gradients, in: Proceedings of the International Conference on Machine Learning (ICML).

Montenegro, A., Mussi, M., Metelli, A.M., Papini, M., 2025. Convergence analysis of policy gradient methods with dynamic stochasticity, in: Proceedings of the International Conference on Machine Learning (ICML).

Montenegro, A., Mussi, M., Papini, M., Metelli, A.M., 2024b. Last-iterate global convergence of policy gradients for constrained reinforcement learning, in: Advances in Neural Information Processing Systems (NeurIPS).

Moskovitz, T., O'Donoghue, B., Veeriah, V., Flennerhag, S., Singh, S., Zahavy, T., 2023. Reload: Reinforcement learning with optimistic ascent-descent for last-iterate convergence in constrained mdps, in: Proceedings of the International Conference on Machine Learning (ICML), PMLR. pp. 25303–25336.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J.D., Razaviyayn, M., 2019. Solving a class of non-convex min-max games using iterative first order methods. Advances in Neural Information Processing Systems (NeurIPS) 32.

Papini, M., Pirotta, M., Restelli, M., 2022. Smoothing policies and safe policy gradients. Machine Learning 111, 4081–4137.

Paternain, S., Chamon, L.F.O., Calvo-Fullana, M., Ribeiro, A., 2019. Constrained reinforcement learning has zero duality gap, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 7553–7563.

Peters, J., Schaal, S., 2006. Policy gradient methods for robotics, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE. pp. 2219–2225.

Rozada, S., Ding, D., Marques, A.G., Ribeiro, A., 2025. Deterministic policy gradient primal-dual methods for continuous-space constrained mdps, in: AAAI Conference on Artificial Intelligence, pp. 20200–20208.

Sehnke, F., Osendorfer, C., Rückstieß, T., Graves, A., Peters, J., Schmidhuber, J., 2010. Parameter-exploring policy gradients. Neural Networks 23, 551–559.

Silver, D., Lever, G., Heess, N.M.O., Degris, T., Wierstra, D., Riedmiller, M.A., 2014. Deterministic policy gradient algorithms, in: Proceedings of the International Conference on Machine Learning (ICML).

Stooke, A., Achiam, J., Abbeel, P., 2020. Responsive safety in reinforcement learning by PID lagrangian methods, in: Proceedings of the International Conference on Machine Learning (ICML), PMLR. pp. 9133–9143.

Sutton, R.S., Barto, A.G., 2018. Reinforcement learning: an introduction. MIT Press.

Tessler, C., Mankowitz, D.J., Mannor, S., 2019. Reward constrained policy optimization.

Todorov, E., Erez, T., Tassa, Y., 2012. Mujoco: A physics engine for model-based control, in: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 5026–5033.

Uchibe, E., Doya, K., 2007. Constrained reinforcement learning from intrinsic and extrinsic rewards, in: IEEE International Conference on Development and Learning, IEEE. pp. 163–168.

Vaswani, S., Yang, L., Szepesvári, C., 2022. Near-optimal sample complexity bounds for constrained mdps, in: Advances in Neural Information Processing Systems (NeurIPS).

Vázquez-Abad, F.J., Krishnamurthy, V., Martin, K., Baltcheva, I., 2002. Self learning control of constrained markov chains - a gradient approach, in: IEEE Conference on Decision and Control (CDC), IEEE. pp. 1940–1945.

Williams, R.J., 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning 8, 229–256.

Xiong, H., Xu, T., Zhao, L., Liang, Y., Zhang, W., 2022. Deterministic policy gradient: Convergence analysis, in: Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI), PMLR. pp. 2159–2169.

Xu, T., Liang, Y., Lan, G., 2021. CRPO: A new approach for safe reinforcement learning with convergence guarantee, in: Proceedings of the International Conference on Machine Learning (ICML), PMLR. pp. 11480–11491.

Yang, J., Kiyavash, N., He, N., 2020. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 1153–1165.

Ying, D., Ding, Y., Lavaei, J., 2022. A dual approach to constrained markov decision processes with entropy regularization, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR. pp. 1887–1909.

Yu, M., Yang, Z., Kolar, M., Wang, Z., 2019. Convergent policy optimization for safe reinforcement learning, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 3121–3133.

Yuan, R., Gower, R.M., Lazaric, A., 2022. A general sample complexity analysis of vanilla policy gradient, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR. pp. 3332–3380.

Zhao, F., You, K., 2021. Primal-dual learning for the model-free risk-constrained linear quadratic regulator, in: Proceedings of the Conference on Learning for Dynamics and Control (CDC), PMLR. pp. 702–714.

Zheng, T., You, P., Mallada, E., 2022. Constrained reinforcement learning via dissipative saddle flow dynamics, in: Asilomar Conference on Signals, Systems, and Computers (ACSSC), IEEE. pp. 1362–1366.