# DUAL OPTIMISTIC ASCENT (PI CONTROL) IS THE AUGMENTED LAGRANGIAN METHOD IN DISGUISE

**Juan Ramirez*** **Simon Lacoste-Julien‡**
Mila - Quebec AI Institute
DIRO, Université de Montréal
‡ Canada CIFAR AI Chair

## ABSTRACT

Constrained optimization is a powerful framework for enforcing requirements on neural networks. These constrained deep learning problems are typically solved using first-order methods on their min-max Lagrangian formulation, but such approaches often suffer from oscillations and can fail to find all local solutions. While the Augmented Lagrangian method (ALM) addresses these issues, practitioners often favor dual optimistic ascent schemes (PI control) on the standard Lagrangian, which perform well empirically but lack formal guarantees. In this paper, we establish a previously unknown equivalence between these approaches: dual optimistic ascent on the Lagrangian is equivalent to gradient descent-ascent on the Augmented Lagrangian. This finding allows us to transfer the robust theoretical guarantees of the ALM to the dual optimistic setting, proving it converges linearly to all local solutions. Furthermore, the equivalence provides principled guidance for tuning the optimism hyper-parameter. Our work closes a critical gap between the empirical success of dual optimistic methods and their theoretical foundation.

## 1 INTRODUCTION

Machine learning problems frequently lead to constrained optimization formulations, where the objective function encodes the learning task, and the constraints guide optimization toward models with desirable properties such as sparsity (Gallego-Posada et al., 2022), fairness (Cotter et al., 2019a), or safety (Dai et al., 2024). Alternatively, learning can be incorporated directly into the constraints, as in SVMs and Feasible Learning (Ramirez et al., 2025b).

A widely adopted approach for solving constrained optimization problems involving deep neural networks is to formulate their associated Lagrangian and seek its min–max points (Elenter et al., 2022; Hashemizadeh et al., 2024; Hounie et al., 2023b; Kotary and Fioretto, 2024). This is typically done through gradient descent–ascent: descent on the problem (*primal*) variables—often with adaptive variants such as Adam (Kingma and Ba, 2015)—and ascent on the Lagrange multipliers (*dual* variables). This first-order Lagrangian approach is popular in constrained deep learning because it empirically handles non-convex problems and scales efficiently to models with billions of parameters.

However, gradient descent-ascent on the Lagrangian suffers from two major limitations: ① in the non-convex setting, it may fail to converge to certain local solutions of the problem—convergence is guaranteed only for solutions that correspond to local min–max points of the Lagrangian (Bertsekas, 2016, Prop. 5.4.2); and ② it often exhibits oscillations, where iterates repeatedly move in and out of the feasible set, slowing down convergence (Platt and Barr, 1987).

A seminal approach that addresses these limitations is the *Augmented Lagrangian* method, which incorporates a quadratic penalty term on the constraints (Hestenes, 1969; Powell, 1969; Rockafellar, 1973). With suitable hyperparameter choices, this method converges to all strict and regular local solutions of non-convex constrained problems, while mitigating oscillations.

Despite these advantages, the Augmented Lagrangian method is not the standard in constrained deep learning. Instead, the community largely relies on the original Lagrangian approach, mitigating its oscillatory behavior with alternative Lagrange multiplier updates such as the PI control strategy (Sohrabi

*Correspondence to: juan.ramirez@mila.quebec

et al., 2024; Stooke et al., 2020), also known as generalized optimistic gradient ascent (Mokhtari et al., 2020). These dual optimistic ascent methods have shown strong empirical performance in reducing oscillations, but their convergence properties remain largely underformalized.

This paper addresses this gap. We establish a simple yet previously unknown result: gradient descent–optimistic ascent on the Lagrangian is equivalent to gradient descent–ascent on the Augmented Lagrangian. For equality-constrained problems, the equivalence holds exactly at the level of matching primal iterates (Theorem 1). For general inequality constraints, we prove that both methods converge to the same set of points—namely, all local solutions of the original problem (Theorem 2).

In §4, we use this equivalence to formalize properties of the dual optimistic ascent (PI control) framework. We show that it strictly improves over standard GDA on the Lagrangian by recovering all local problem solutions (Theorem 3) and provide local convergence guarantees to these solutions (Corollaries 2 and 3). We further characterize the role of the optimism coefficient (proportional gain), formally showing that larger values reduce oscillations but may induce ill-conditioning (§4.3). Finally, in §4.4, we highlight how existing empirical results validate our findings.

**Scope.** All results are presented in the context of constrained optimization. Our claims about the optimistic gradient method are restricted to: ① its role in maximizing the dual player in a Lagrangian formulation (i.e., a linear player in a non-convex–linear min-max problem), and ② its application *only* to that player; for the non-convex primal player, we generally assume a different method is used. Furthermore, our equivalence results apply solely to a *first-order*, *single-step* Augmented Lagrangian method. They do not extend to Augmented Lagrangian algorithms with multiple primal steps between dual updates, nor to cases in which the primal minimization algorithm is second-order or higher.

## 2 Background

### 2.1 Constrained Optimization

Let $f : \mathbb{R}^d \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^m$, and $\boldsymbol{h} : \mathbb{R}^d \to \mathbb{R}^n$ be twice-differentiable functions. We consider the following family of constrained optimization problems:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) \quad \text{subject to} \quad \boldsymbol{g}(\boldsymbol{x}) \preceq \mathbf{0}, \quad \boldsymbol{h}(\boldsymbol{x}) = \mathbf{0}, \tag{CMP}$$

where $\preceq$ denotes element-wise inequality. We refer to $f$ as the objective function, and to $\boldsymbol{g}$ and $\boldsymbol{h}$ as the inequality and equality constraints, respectively.

**Definitions**. A point $\boldsymbol{x}$ is *feasible* if it satisfies all constraints. A point $\boldsymbol{x}^*$ is a *local constrained minimizer* if it is feasible and there exists an $\epsilon > 0$ such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all feasible $\boldsymbol{x}$ with $\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \epsilon$. A *global constrained minimizer* is a feasible point that satisfies this inequality for all feasible $\boldsymbol{x}$. An inequality constraint $g_i(\boldsymbol{x}) \leq 0$ is *active* at $\boldsymbol{x}$ if it holds with equality, i.e., $g_i(\boldsymbol{x}) = 0$. We denote the set of active inequality constraint indices at $\boldsymbol{x}$ by $\mathcal{A}_{\boldsymbol{x}} = \{i \mid g_i(\boldsymbol{x}) = 0\}$. Equality constraints are always considered "active" when satisfied.

**Solving constrained optimization problems**. A general approach to solving differentiable constrained optimization problems is to find min-max points of their associated Lagrangian (Arrow et al., 1958), which correspond to solutions of the original problem (Bertsekas, 2016, Props. 4.3.1 & 4.3.2):

$$\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^* \in \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \max_{\boldsymbol{\lambda}\succeq\mathbf{0},\boldsymbol{\mu}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \triangleq f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\mu}^\top \boldsymbol{h}(\boldsymbol{x}), \tag{Lag}$$

where $\boldsymbol{\lambda} \succeq \mathbf{0}$ and $\boldsymbol{\mu}$ are the Lagrange multipliers associated with the inequality and equality constraints, respectively. We refer to $\boldsymbol{x}$ as the *primal* variables, and to $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ as the *dual* variables.

A simple algorithm for finding min-max points of Eq. (Lag) is alternating[1] (projected) gradient descent-ascent (GDA): descent on $\boldsymbol{x}$ and (projected) ascent on $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$:

$$\begin{aligned} \boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t + \eta_{\text{dual}}\, \boldsymbol{h}(\boldsymbol{x}_t), \qquad \boldsymbol{\lambda}_{t+1} \leftarrow \left[\boldsymbol{\lambda}_t + \eta_{\text{dual}}\, \boldsymbol{g}(\boldsymbol{x}_t)\right]_+, \\ \boldsymbol{x}_{t+1} &\leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}} \left[\nabla f(\boldsymbol{x}_t) + \boldsymbol{\lambda}_{t+1}^\top \nabla \boldsymbol{g}(\boldsymbol{x}_t) + \boldsymbol{\mu}_{t+1}^\top \nabla \boldsymbol{h}(\boldsymbol{x}_t)\right], \end{aligned} \tag{Lag-GDA}$$

[1]Alternating updates are often preferred over simultaneous updates for constrained optimization because they provide stronger convergence guarantees when the Lagrangian is strongly convex (Zhang et al., 2022), without adding computational overhead (Sohrabi et al., 2024).

where $[\,\cdot\,]_+$ denotes projection onto the non-negative orthant to enforce $\boldsymbol{\lambda} \succeq \mathbf{0}$, and $\eta_{\boldsymbol{x}}, \eta_{\text{dual}} > 0$ are the primal and dual step-sizes. The primal update direction is a linear combination of the objective and constraint gradients, weighted by the current multipliers, which are typically initialized to zero. In practice, the descent step is often replaced with adaptive first-order schemes such as Adam. We refer to Eq. (Lag-GDA) as the *first-order Lagrangian approach*, or simply the Lagrangian approach.

**Constrained deep learning**. Applying constraints to deep neural networks—where the parameters $\boldsymbol{x}$ can number in the billions—is challenging due to the non-convexity and scale of the resulting optimization problems. Classical approaches that exploit problem structure or require costly operations, such as methods involving projections (Goldstein, 1964; Levitin and Polyak, 1966), feasible directions (Frank and Wolfe, 1956; Zoutendijk, 1960), or standard SQPs (Han, 1977; Powell, 1978; Wilson, 1963), are either inapplicable to non-convex deep learning or impractical at this scale.

By contrast, the Lagrangian approach in Eq. (Lag-GDA) has become a dominant method for constrained deep learning, with applications across a wide range of domains (Cotter et al., 2019a; Elenter et al., 2022; Gallego-Posada et al., 2022; Hashemizadeh et al., 2024; Hounie et al., 2023a;b; Ramirez et al., 2025b; Robey et al., 2021; Zhang et al., 2025), and integration into popular frameworks such as PyTorch (Paszke et al., 2019) and TensorFlow (Abadi et al., 2016) via libraries like Cooper (Gallego-Posada et al., 2025) and TFCO (Cotter et al., 2019b).

Its popularity stems from its convergence properties and computational scalability. It provides local convergence guarantees to min–max points of the Lagrangian, even in the non-convex setting (Bertsekas, 2016, Prop. 5.4.2), while remaining computationally efficient (Gallego-Posada, 2024; Ramirez et al., 2025a). Its overhead relative to unconstrained deep learning is negligible, limited to storing and updating the multipliers.[2]

**Limitations**. In the non-convex regime, the Lagrangian approach faces several drawbacks. Not all local constrained minimizers $\boldsymbol{x}^*$ of Eq. (CMP)—together with their corresponding optimal multipliers $\boldsymbol{\lambda}^* \succeq \mathbf{0}$ and $\boldsymbol{\mu}^*$—necessarily correspond to min-max points of the Lagrangian $\mathcal{L}$. In particular, $\boldsymbol{x}^*$ need not minimize $\mathcal{L}(\,\cdot\,, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$,[3] and the Hessian $\nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ may fail to be positive definite. In fact, the Lagrangian can be *strictly concave* in some infeasible directions—whereas strict concavity precludes optimality in unconstrained problems, it can arise at constrained optima (Bertsekas, 2016, Prop. 4.3.2). Such non-convex stationary points cannot be reached via Eq. (Lag-GDA).

Despite this limitation, Eq. (Lag-GDA) has proven effective in practice for a wide range of non-convex constrained deep learning applications. However, another challenge limits its applicability: the optimization dynamics often induce oscillations in the multipliers and their associated constraints, causing the primal iterates to repeatedly move in and out of the feasible set (Platt and Barr, 1987). These oscillations slow convergence and can be critical in settings where infeasibility cannot be tolerated—for example, when violations disrupt physical systems or compromise safety.

These limitations have motivated two main strategies for improvement. The first is the *Augmented Lagrangian Method*, which makes the (Augmented) Lagrangian *strictly convex* at all strict and regular constrained minimizers of Eq. (CMP). The second replaces simple gradient ascent on the multipliers with *generalized optimistic gradient ascent*, which mitigates oscillations in the dynamics. These methods are discussed in §2.2 and §2.3, respectively. In §3, we prove that they are, in fact, equivalent.

### 2.2 The Augmented Lagrangian Method

The (HPR) Augmented Lagrangian function (Hestenes, 1969; Powell, 1969; Rockafellar, 1973) adds a quadratic penalty to the standard Lagrangian to penalize constraint violations:

$$\mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \triangleq f(\boldsymbol{x}) + \frac{1}{2c}\left[\|\boldsymbol{\mu} + c\,\boldsymbol{h}(\boldsymbol{x})\|_2^2 - \|\boldsymbol{\mu}\|_2^2 + \left\|[\boldsymbol{\lambda} + c\,\boldsymbol{g}(\boldsymbol{x})]_+\right\|_2^2 - \|\boldsymbol{\lambda}\|_2^2\right] \tag{AL}$$

where $c > 0$ is a penalty coefficient. Expanding the squared norms in Eq. (AL) reveals that constraint violations are penalized by both a linear term involving the multiplier and a quadratic term scaled by $c$. The projection $[\,\cdot\,]_+$ creates a *one-sided* penalty for inequalities: it penalizes violations ($g_i(\boldsymbol{x}) > 0$)

[2]This overhead is negligible for two reasons: the number of constraints—and thus multipliers—is typically much smaller than the number of model parameters, adding minimal *memory* overhead; and the dual gradients correspond to constraint values, requiring no extra backward passes and thus negligible *computational* overhead.

[3]Even solutions that minimize $\mathcal{L}$—albeit not strictly—may not correspond to limit points of Eq. (Lag-GDA).

but avoids pushing iterates deeper into the feasible region once a constraint is sufficiently satisfied ($g_i(\boldsymbol{x}) \leq -\lambda_i/c$). This ensures that optimization focuses only on violated or nearly-active constraints.

**The set of min-max points of $\mathcal{L}_c$.** The Augmented Lagrangian preserves and enhances the solution set of the standard Lagrangian. Since the two functions share the same set of stationary points (Rockafellar, 1973, Cor. 3.4), they have an identical pool of candidate solutions. Furthermore, every local min-max point of $\mathcal{L}$ remains a local min-max point of $\mathcal{L}_c$, ensuring that any solution attainable with the standard Lagrangian approach remains attainable within the Augmented Lagrangian framework.

More importantly, under regularity assumptions, *every* strict local constrained minimizer $\boldsymbol{x}^*$ admits a threshold $\bar{c} \geq 0$ such that, for all $c \geq \bar{c}$, the Augmented Lagrangian $\mathcal{L}_c$ is *strictly convex* at $\boldsymbol{x}^*$, even if the original Lagrangian $\mathcal{L}$ is not convex there (Bertsekas, 2016, §4.2.1). Consequently, all strict local constrained minimizers of Eq. (CMP) are upgraded to local min-max points of $\mathcal{L}_c$.

Two important algorithmic consequences follow: ① *all* strict local constrained minimizers become strict local min-max points of $\mathcal{L}_c$, for which GDA enjoys local linear convergence guarantees (Bertsekas, 2016, Prop. 5.4.2); and ② the sets of strict local constrained minimizers of Eq. (CMP) and strict local min-max points of $\mathcal{L}_c$ coincide, ensuring that GDA on the Augmented Lagrangian converges to—and *only* to—stationary points corresponding to solutions of Eq. (CMP). Further benefits of the Augmented Lagrangian approach—including its ability to mitigate oscillations—are discussed in §4.

**The Augmented Lagrangian Method** seeks min-max points of the Augmented Lagrangian function:

$$\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^* \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} \max_{\boldsymbol{\lambda} \succeq \mathbf{0}, \boldsymbol{\mu}} \mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}). \tag{1}$$

The classic algorithm for solving this problem—the Method of Multipliers—iteratively performs:

$$\boldsymbol{x}_{t+1} \in \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} \mathcal{L}_{c_t}(\boldsymbol{x}, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t), \quad \boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + c_t\, \boldsymbol{h}(\boldsymbol{x}_{t+1}), \quad \boldsymbol{\lambda}_{t+1} = [\boldsymbol{\lambda}_t + c_t\, \boldsymbol{g}(\boldsymbol{x}_{t+1})]_+ . \tag{2}$$

In this method, the primal minimization is typically performed approximately, either for a fixed number of optimization steps or until a numerical tolerance is reached. The subsequent dual updates correspond to gradient ascent steps on the standard Lagrangian $\mathcal{L}$ with step-sizes $c_t$. The penalty coefficients are chosen as a non-decreasing sequence $c_{t+1} \geq c_t > 0$, often updated according to a schedule—for example, increasing $c_t$ when constraint violations are not reduced sufficiently across successive iterations (see, e.g., Birgin and Martínez (2014, Eq. 4.9 in Alg. 4.1)).

The Augmented Lagrangian method is a seminal approach for solving non-convex constrained optimization problems, with variants implemented in numerous optimization libraries across different programming languages (Conn et al., 2013; Gallego-Posada et al., 2025; Johnson; Pas et al., 2022). Its success spans a broad range of scientific problems, and it has become a key tool in constrained deep learning, applied in works ranging from seminal contributions to recent advances (Brouillard et al., 2020; Kotary and Fioretto, 2024; Lokhande et al., 2020; Platt and Barr, 1987; Shi et al., 2023).

In deep learning, where full (or even approximate) minimization of $\mathcal{L}_c$ is computationally infeasible, it is often replaced by one or more steps of a first-order optimizer, such as Adam. Accordingly, in this paper we consider the following family of Augmented Lagrangian algorithms: alternating gradient descent–(projected) gradient ascent on $\mathcal{L}_c$, with the primal update performed first.

$$\begin{aligned} \boldsymbol{x}_{t+1} &\leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}} \left[ \nabla f(\boldsymbol{x}_t) + \left[\boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_t)\right]_+^\top \nabla \boldsymbol{g}(\boldsymbol{x}_t) + \left(\boldsymbol{\mu}_t + c\, \boldsymbol{h}(\boldsymbol{x}_t)\right)^\top \nabla \boldsymbol{h}(\boldsymbol{x}_t) \right] \\ \boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t + \eta_{\text{dual}}\, \boldsymbol{h}(\boldsymbol{x}_{t+1}) \qquad \boldsymbol{\lambda}_{t+1} \leftarrow \left(1 - \tfrac{\eta_{\text{dual}}}{c}\right) \boldsymbol{\lambda}_t + \tfrac{\eta_{\text{dual}}}{c} \left[\boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_{t+1})\right]_+ \end{aligned} \tag{AL-GDA}$$

where $0 < \eta_{\text{dual}} \leq c$ and $\eta_{\boldsymbol{x}} > 0$ are fixed step-sizes. The algorithm first takes a single primal gradient descent step on $\mathcal{L}_c$, then performs a dual gradient ascent step on the multipliers with step-size $\eta_{\text{dual}}$, which is independent of the penalty $c$. Notably, the $\boldsymbol{\lambda}$-update is a convex combination of the previous iterate $\boldsymbol{\lambda}_t$ and a new estimate $[\boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_{t+1})]_+$. Setting $\eta_{\text{dual}} = c$ recovers the standard dual updates of the Method of Multipliers in Eq. (2). This method is therefore more specific than Eq. (2) by replacing the full primal minimization with a single gradient step and more general, as it decouples the dual step-size from $c$. For a full derivation of these updates, see Section B.

The primal update in Eq. (AL-GDA) can be viewed as a gradient descent step on the standard Lagrangian, using multiplier estimates that differ from their latest iterates. These estimates simulate a (projected) ascent step on $\mathcal{L}$, acting as a proactive "lookahead" to calibrate the primal updates. Recognizing this is the key insight behind the equivalence with dual optimistic ascent shown in §3.

### 2.3 DUAL OPTIMISTIC ASCENT

The oscillatory behavior inherent in the dynamics of Eq. (Lag-GDA) has motivated alternatives to plain gradient ascent for updating the dual variables. A prominent class of such methods is based on *optimism* (Rakhlin and Sridharan, 2013). In this work, we adopt the generalized optimistic gradient method of Mokhtari et al. (2020), applying it solely to $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. Combining this with standard gradient descent on $\boldsymbol{x}$ and alternating updates—where the multipliers are updated first—yields:

$$\begin{aligned}
\boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t + \eta_{\text{dual}}\, \boldsymbol{h}(\boldsymbol{x}_t) + \omega\big[\boldsymbol{h}(\boldsymbol{x}_t) - \boldsymbol{h}(\boldsymbol{x}_{t-1})\big], \\
\boldsymbol{\lambda}_{t+1} &\leftarrow \Big[\boldsymbol{\lambda}_t + \eta_{\text{dual}}\, \boldsymbol{g}(\boldsymbol{x}_t) + \omega\big[\boldsymbol{g}(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_{t-1})\big]\Big]_+, \\
\boldsymbol{x}_{t+1} &\leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}} \left[\nabla f(\boldsymbol{x}_t) + \boldsymbol{\lambda}_{t+1}^\top \nabla \boldsymbol{g}(\boldsymbol{x}_t) + \boldsymbol{\mu}_{t+1}^\top \nabla \boldsymbol{h}(\boldsymbol{x}_t)\right]
\end{aligned} \tag{Lag-GD-OA}$$

where $\omega > 0$ is the optimism coefficient. The first dual update, which would otherwise require evaluating "$\boldsymbol{h}(\boldsymbol{x}_{-1})$" and "$\boldsymbol{g}(\boldsymbol{x}_{-1})$," is typically replaced by a (projected) gradient ascent step.

While optimism is often applied to both players in general min-max optimization (i.e., optimistic gradient descent–ascent; OGDA), we follow the more general approach of Sohrabi et al. (2024); Stooke et al. (2020)[4] for constrained problems by applying optimism only to the dual variables. This preserves flexibility in the choice of the primal update, allowing to retain task-specific training pipelines—such as standard Adam without optimism. Such flexibility is especially important in deep learning, where model training pipelines are often highly specialized, making it undesirable to modify them solely to accommodate alternative dual optimization algorithms.

Dual (generalized) optimistic updates have been applied to stabilize optimization dynamics of constrained deep learning problems across reinforcement learning (Moskovitz et al., 2023; Stooke et al., 2020), unsupervised learning (Shao et al., 2022), and supervised learning (Sohrabi et al., 2024).

Intuitively, the optimistic update in Eq. (Lag-GD-OA) dampens oscillations by adjusting dual ascent steps with information from past constraint violations; specifically, by adding the optimistic term $\omega\big(\boldsymbol{g}(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_{t-1})\big)$. When violations shrink across iterations, this correction term is negative, acting as a brake: the multiplier increases more conservatively, avoiding an "overshoot" beyond its optimum. This prevents over-penalization of the constraints in later primal steps, thereby stabilizing the dynamics and reducing oscillations. For further discussion, see Sohrabi et al. (2024, §4.3).

**Literature gaps**. While the properties of the Augmented Lagrangian method are well understood, dual optimistic ascent on the Lagrangian remains largely underformalized. It lacks convergence guarantees and is supported exclusively by empirical demonstrations of its ability to mitigate oscillations.

Despite the vast literature on optimistic methods, most existing results do not directly apply to the dual optimistic ascent algorithm in Eq. (Lag-GD-OA). For general min-max games, OGDA is known to expand the set of stationary points to which it can converge relative to standard GDA (Daskalakis and Panageas, 2018). While this set can include the global solution in bilinear games, the nature of these additional points remains largely uncharacterized in general non-convex–non-concave games. Moreover, the established convergence guarantees for OGDA—such as global linear convergence under strong convexity–strong concavity (Gidel et al., 2019), sublinear convergence under convexity–concavity (Gorbunov et al., 2022), and sublinear convergence for non-convex–(strongly) concave settings (Mahdavinia et al., 2022)—are derived in frameworks incompatible with our own.

The inapplicability of these findings stems from two key issues. First, the cited results concern a different algorithm that applies optimism to *both* players, not just the dual player as in Eq. (Lag-GD-OA).[5] Second, the underlying assumptions are mismatched—they are either too restrictive, requiring (strong) convexity for the minimization player, or too general, assuming a non-convex–concave game that fails to exploit the specific non-convex–*linear* structure of our problem.

[4]The generalized optimistic gradient method is a special case of the algorithms in both papers; specifically, it corresponds to a Proportional-Integral (PI) controller on the multipliers.

[5]Moreover, known convergence results for OGDA seldom consider the *generalized* case where $\omega \neq \eta_{\text{dual}}$, with the notable exception of Mokhtari et al. (2020) for bilinear games.

## 3 Dual Optimistic Ascent (PI Control) is the Augmented Lagrangian Method in Disguise

In this section, we show that the Augmented Lagrangian method (Eq. (AL-GDA) in §2.2) and dual optimistic ascent (Eq. (Lag-GD-OA) in §2.3) are equivalent for equality-constrained problems when $c = \omega$, in the sense that their primal iterates coincide (Theorem 1 in §3.1). This equivalence does not extend to problems with inequality constraints. Nevertheless, Theorem 2 in §3.2 establishes that both algorithms share the same set of locally stable stationary points. Consequently, they achieve the same fundamental objective: ensuring that all strict and regular local constrained minimizers of Eq. (CMP) are limit points of their dynamics. We discuss further implications of these results in §4.

We begin by proving that both algorithms share the same set of stationary (fixed) points:

**Proposition 1** (Equivalence of Stationary Points)**.** *The set of fixed points for algorithms Eqs.* (AL-GDA) *and* (Lag-GD-OA) *are the same, and correspond to the set of KKT points of Eq.* (CMP)*.*

*Proof.* See *Proof of Proposition 1* in Section C. □

### 3.1 Equality Constraints

We first show that, for equality-constrained problems, the primal iterates of Eq. (Lag-GD-OA) and Eq. (AL-GDA) coincide when $\omega = c$ and the initialization is chosen appropriately.

**Theorem 1** (Equivalence for equality-constrained problems)**.** *The primal iterates $\{\boldsymbol{x}_t\}_{t=0}^{\infty}$ generated by primal-first GDA on the Augmented Lagrangian (Eq.* (AL-GDA)*) and dual-first gradient descent–optimistic ascent on the Lagrangian (Eq.* (Lag-GD-OA)*) for an equality-constrained problem match, provided that: ① the penalty and optimism coefficients are constant and equal, $\omega = c > 0$; and ② their respective initializations are chosen as $(\boldsymbol{x}_0, \boldsymbol{\mu}_0)$ and $(\boldsymbol{x}_0, \boldsymbol{\mu}_0 + (c - \eta_{dual})\boldsymbol{h}(\boldsymbol{x}_0))$.*

*Proof.* See *Proof of Theorem 1* in Section C. □

Note that Theorem 1 applies to *any* first-order optimization algorithm used to minimize the (Augmented) Lagrangian, not only gradient descent. This holds because both algorithms generate the same sequence of primal gradients. In particular, the result extends to primal updates performed with first-order deep learning optimizers such as Adam.

### 3.2 General Constrained Problems with Inequality Constraints

Unlike the equality-constrained case, the iterates of Eq. (AL-GDA) and Eq. (Lag-GD-OA) do not coincide for inequality-constrained problems. This difference arises from the placement of projections: in Eq. (Lag-GD-OA), they are applied only once when updating the multipliers, whereas in Eq. (AL-GDA), they are applied twice—once in the multiplier update and again when computing the "lookahead" multiplier for the primal gradient. Nevertheless, as shown in Theorem 2, the two algorithms share the same sets of locally stable stationary points. In this sense, both are equally powerful—and strictly more so than Eq. (Lag-GDA). We first define local stability:

**Definition 1** (Local stability)**.** *Consider an algorithm*

$$\boldsymbol{x}_{t+1} \leftarrow F(\boldsymbol{x}_t, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t), \qquad \boldsymbol{\lambda}_{t+1} \leftarrow G(\boldsymbol{x}_t, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t), \qquad \boldsymbol{\mu}_{t+1} \leftarrow H(\boldsymbol{x}_t, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t). \tag{3}$$

*A fixed point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ of this algorithm is called a locally stable stationary point (LSSP) if the Jacobian $\mathcal{J}$ of the joint operator $[F, G, H]$, evaluated at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$, has spectral radius strictly less than one, i.e., $\rho(\mathcal{J}) < 1$.*

A standard result states that an algorithm exhibits local linear convergence to all of its LSSPs (Bertsekas, 2016, Prop. 5.4.1). To analyze the stability of our two methods, we therefore study their respective operator Jacobians: $\mathcal{J}_{\text{AL}}$ for the Augmented Lagrangian method Eq. (AL-GDA) and $\mathcal{J}_{\text{OG}}$ for dual optimistic ascent Eq. (Lag-GD-OA). These are derived in Lemmas 2 and 4 of Section C.

**Theorem 2** (Equivalence of LSSPs - inequality constraints)**.** *Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ be a stationary point of the Lagrangian for an inequality-constrained problem, satisfying strict complementary slackness (see Assumption 1 in Section A). Then Eq.* (AL-GDA) *with penalty coefficient $c$ converges locally to $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ if and only if Eq.* (Lag-GD-OA) *with optimism coefficient $\omega = c$ also converges locally to that point. Moreover, the spectral radii of the Jacobians of the two algorithms satisfy*

$$\rho(\mathcal{J}_{AL}) = \max\{\rho(\mathcal{J}_{OG}), 1 - \eta_{dual}/c\}. \tag{4}$$

*Proof.* See *Proof of Theorem 2* in Section C. □

We can also conclude the following result for general constrained optimization problems:

**Corollary 1** (Equivalence of LSSPs)**.** *The algorithms in Eqs.* (AL-GDA) *and* (Lag-GD-OA) *share the same set of locally stable stationary points satisfying Assumption 1.*

*Proof.* The result follows by applying the same logic as in the proof of Theorem 2, where equality constraints are treated as a special case of active inequality constraints. □

## 4 Implications and Discussion

This section builds on the equivalence results from §3 to establish formal properties of the dual optimistic ascent method (PI control). We emphasize that these results apply specifically to Eq. (Lag-GD-OA), which takes only a *single* primal descent step per iteration. The underlying equivalence, and therefore the subsequent results, do not hold if multiple primal steps occur between dual updates.

### 4.1 Characterizing the Stable Points of Dual Optimistic Ascent

We now show that dual optimistic ascent offers a principled approach to constrained optimization: it recovers all regular, strict local constrained minimizers of Eq. (CMP). Although the method has been applied to constrained problems before, this property has not been previously identified or formalized.

Subsequent results rely on three standard assumptions, which we state formally in Section A: strict complementary slackness (Assumption 1), the second-order sufficiency condition (Assumption 2), and the linear independence constraint qualification (Assumption 3). We now recall the following classical results for the Augmented Lagrangian method:

**Proposition 2.** *(Bertsekas, 2016, §4.2.1) Let $\boldsymbol{x}^*$ be a strict local constrained minimizer with corresponding Lagrange multipliers $\boldsymbol{\lambda}^* \geq \mathbf{0}$ and $\boldsymbol{\mu}^*$, satisfying Assumptions 1 to 3. Then there exists $\bar{c} > 0$ such that, for every $c \geq \bar{c}$, the Augmented Lagrangian $\mathcal{L}_c$ is strictly convex in $\boldsymbol{x}$ at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$.*

**Proposition 3.** *(ALM finds all constrained minimizers) For any strict local constrained minimizer $\boldsymbol{x}^*$ satisfying Assumptions 1 to 3, there exist $c > 0$ large enough and $\eta_{\boldsymbol{x}}, \eta_{dual} > 0$ small enough such that $\boldsymbol{x}^*$ is an LSSP of Eq.* (AL-GDA).

*Proof.* This is well-known for simultaneous GDA in equality-constrained problems (Bertsekas (2016, Prop. 5.4.2)). For a proof specific to Eq. (AL-GDA), see *Proof of Proposition 3* in Section C. □

The converse also holds: the LSSPs of ALM dynamics are precisely the local constrained minimizers of Eq. (CMP), indicating that Eq. (AL-GDA) does not converge to spurious stationary points of $\mathcal{L}_c$.

**Proposition 4.** *(ALM finds only constrained minimizers) Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be an LSSP of Eq.* (AL-GDA) *for some $\eta_{\boldsymbol{x}}, \eta_{dual}, c > 0$. Then $\boldsymbol{x}^*$ is a local constrained minimizer of Eq.* (CMP).

*Proof.* See *Proof of Proposition 4* in Section C. □

We now characterize the solution set of dual optimistic ascent.

**Theorem 3** (LSSPs of dual optimistic ascent)**.** *Let $\boldsymbol{x}^*$ satisfy Assumptions 1 and 3. Then $\boldsymbol{x}^*$ is a strict local constrained minimizer of Eq.* (CMP)*—that is, it satisfies Assumption 2—if and only if there exists an optimism coefficient $\bar{\omega} \geq 0$ such that for all $\omega \geq \bar{\omega}$, there exist learning rates $\eta_{\boldsymbol{x}}, \eta_{dual} > 0$ small enough for which $\boldsymbol{x}^*$ is an LSSP of the dual optimistic ascent dynamics Eq.* (Lag-GD-OA).

*Proof.* This follows directly from the equivalence results in Corollary 1 together with the properties of Eq. (AL-GDA) given in Propositions 3 and 4 above. □

Theorem 3 establishes dual optimistic ascent as a principled framework for constrained optimization. While it is known that optimistic methods converge to a broader set of stable points than standard GDA (Daskalakis and Panageas, 2018), our result refines this for the constrained setting by precisely characterizing *which* additional stable points arise from (dual) optimism: it recovers all strict and regular local constrained minimizers of Eq. (CMP) and *only* such solutions.[6] This makes it strictly more powerful than simple Eq. (Lag-GDA). For an analysis of how large $\omega = c$ must be to guarantee convergence to a given $\boldsymbol{x}^*$, see Bertsekas (2014, Prop. 2.7).

[6]This highlights a key difference with general min-max games (e.g., GANs), where new equilibria stabilized by optimism can be spurious. In constrained optimization, these "spurious" solutions are a feature, not a bug.

### 4.2 Convergence Results

**Non-convex problems**. Our equivalence results allow us to transfer the well-understood convergence guarantees of the Augmented Lagrangian method to the dual optimistic ascent framework.

**Corollary 2** (Local convergence rate of dual optimistic ascent)**.** *For appropriate hyperparameter choices, Eq.* (Lag-GD-OA) *exhibits local linear convergence to all strict local constrained minimizers of Eq.* (CMP) *satisfying Assumptions 1 to 3. Whenever $\eta_{dual}$ is sufficiently close to $\omega = c$ (so that $1 - \eta_{dual}/c < \rho(\mathcal{J}_{AL})$), the convergence rate of Eq.* (Lag-GD-OA) *matches that of Eq.* (AL-GDA)*.*

*Proof.* Local linear convergence follows directly from Theorem 3. The spectral norm relationship between $\rho(\mathcal{J}_{\text{OG}})$ and $\rho(\mathcal{J}_{\text{AL}})$ established in Theorem 2 links the convergence rates of the two algorithms for inequality-constrained problems. Note that equality constraints can be treated locally as active inequalities, and thus the same relationship between spectral norms holds in that case. □

This result also implies that for problems where a specific convergence rate is known for Eq. (Lag-GDA),[7] that rate transfers directly to Eq. (Lag-GD-OA), provided $\eta_{\text{dual}}$ is sufficiently close to $c$.

Global convergence, however, is more delicate. While the matching primal iterates for equality-constrained problems (Theorem 1) might suggest that convergence guarantees could transfer directly across algorithms, this is not the case for inequality constraints. There, our equivalence is only local—ensuring the same set of LSSPs (Corollary 1) but not the same global behavior.

Even for equality constraints, a key obstacle prevents us from proving global convergence of dual optimistic ascent using known results for the Augmented Lagrangian method. Classic ALM results, such as the Q-linear rate in Bertsekas (2014, Prop. 2.7), apply to the Method of Multipliers (Eq. (2)) and rely on each primal step yielding a *sufficiently accurate (local) minimizer* of $\mathcal{L}_c$. In contrast, our equivalence holds only for *single-step first-order* updates, which do not meet this requirement.

One might then consider performing multiple primal minimization steps per dual optimistic ascent update to meet this condition. However, this breaks the equivalence to Eq. (AL-GDA) and instead amounts to a multi-step minimization of the standard Lagrangian $\mathcal{L}$. Since $\mathcal{L}$ may fail to be locally convex near a solution, such a procedure lacks convergence guarantees.

It may still be possible to establish global convergence of Eq. (Lag-GD-OA) for non-convex problems directly—following, for example, strategies developed for OGDA in nonconvex–concave games (Mahdavinia et al., 2022). However, this lies beyond the scope of our work.

**Convex problems.** In the convex setting, our equivalence theorems yield global convergence guarantees for dual optimistic ascent on equality-constrained problems. Extending these guarantees to inequality constraints remains an open direction for future work.

**Corollary 3** (Global linear convergence for convex equality-constrained problems)**.** *Consider a problem with a convex, smooth objective $f(\boldsymbol{x})$ and affine equality constraints $\boldsymbol{h}(\boldsymbol{x}) = B\boldsymbol{x} + b$. Assume the problem has a unique solution $\boldsymbol{x}^*$ and that LICQ is satisfied (Assumption 3), i.e., $B$ has full row rank. Then, for any $\omega > 0$ and sufficiently small step-sizes $\eta_{dual}, \eta_{\boldsymbol{x}}$, Eq.* (Lag-GD-OA) *converges globally at a linear rate to the optimal pair $(\boldsymbol{x}^*, \boldsymbol{\mu}^*)$. The exact rate depends on the properties of $f$ and $B$; see Alghunaim and Sayed (2020, Theorem 1) for a precise characterization.*

*Proof.* This follows from the equivalence in Theorem 1, and Alghunaim and Sayed (2020, Theorem 1), which establishes global linear convergence for Eq. (AL-GDA). □

This global convergence guarantee hinges on the strong convexity of the Augmented Lagrangian $\mathcal{L}_c$ on $\boldsymbol{x}$, which is induced even when the original objective $f$ is merely convex. In contrast, the standard Lagrangian remains only convex in $\boldsymbol{x}$, and GDA is not guaranteed to converge on the resulting convex–concave problem (Gidel et al., 2019). Dual optimistic ascent avoids this failure mode by inheriting the favorable convergence properties of the Augmented Lagrangian method.

### 4.3 Practical Implications for Hyper-parameter Tuning

Our equivalence result provides practical guidance on choosing the optimism hyper-parameter $\omega$. This parameter introduces a critical trade-off analogous to the coefficient $c$ in the Augmented Lagrangian

[7] There is no general closed-form expression for the iteration complexity of the Augmented Lagrangian method in Eq. (AL-GDA); such rates can only be established under additional assumptions on $f$, $\boldsymbol{g}$, and $\boldsymbol{h}$.

method: larger values of $\omega$ are beneficial as they ① expand the set of attainable solutions and ② dampen oscillations in the optimization dynamics. However, ③ excessively large values can lead to an ill-conditioned problem, which may in turn slow convergence. We now formalize these points.

**Corollary 4** (Monotonic inclusion of solutions)**.** *The set of strict local minimizers of Eq.* (CMP) *that are LSSPs of the dual optimistic ascent dynamics is monotonically non-decreasing in $\omega$. For sufficiently large $\omega$, this set coincides with all local minimizers satisfying Assumptions 1 to 3.*

*Proof.* This follows from the equivalence result in Corollary 1 and Prop. 3. □

In practice, however, it is generally unclear what value of $\omega$ is sufficiently large to turn a given solution $\boldsymbol{x}^*$ into an LSSP; this would typically require knowing $\boldsymbol{x}^*$ itself. This practical uncertainty motivates a simple strategy: to ensure all solutions are attainable, one may consider taking $\omega \to \infty$.

**Dampening Oscillations.** A key motivation for moving beyond standard GDA on the Lagrangian $\mathcal{L}$ is the presence of oscillations. Whenever the Hessian $\nabla^2_{\boldsymbol{x}}\mathcal{L}$ is indefinite, the dynamics are characterized by imaginary eigenvalues, an effect observed in practice (Sohrabi et al., 2024; Stooke et al., 2020) and expected in theory (Benzi and Simoncini, 2006, Prop 2.5 & Remark 2.8).

The Augmented Lagrangian method is a classical solution: as the penalty coefficient $c \to \infty$, the eigenvalues of its dynamics become purely real, eliminating oscillations (Benzi and Simoncini, 2006, Remark 2.9). Our equivalence result (Corollary 1) shows that Eq. (Lag-GD-OA) inherits this damping property, formalizing the insight that larger values of $\omega$ produce a more heavily dampened system.

**Corollary 5** (Dual optimistic ascent dampens oscillations)**.** *As the optimism parameter $\omega \to \infty$, the eigenvalues of the Jacobian $\mathcal{J}_{OG}$ become real, thereby removing oscillations.*

*Proof.* This follows from Corollary 1 and Benzi and Simoncini (2006, Prop 2.5 & Remark 2.9). □

As before, the exact threshold for $\omega$ to fully remove oscillations depends on problem-specific properties at $\boldsymbol{x}^*$, making it unknown in practice. This again motivates selecting a generously large $\omega$.

**Conditioning trade-off**. However, while these observations may suggest that increasing the optimism coefficient is beneficial, we emphasize another aspect: Bertsekas (2014, Eq. 16 in §2.1) shows that as $c \to \infty$, $\mathcal{J}_{\text{AL}}$—and hence $\mathcal{J}_{\text{OG}}$—becomes increasingly ill-conditioned.

**Corollary 6** (Ill-conditioning of dual optimistic ascent)**.** *As $\omega \to \infty$, the condition number of the operator Jacobian $\mathcal{J}_{OG}$ also tends to infinity.*

*Proof.* This follows from our equivalence result (Corollary 1) and the known ill-conditioning of the Augmented Lagrangian for a large penalty $c$ (Bertsekas, 2014, see Eq. 16 in §2.1). □

This reveals a critical trade-off: the optimism coefficient $\omega$ must be large enough to recover all solutions, yet not so large that it induces ill-conditioning. Since the optimal value depends on local properties at an unknown solution $\boldsymbol{x}^*$ and may not generalize across the optimization landscape, $\omega$ must be tuned in practice. Our equivalence provides a principled path forward: practitioners can use well-established ALM penalty-scheduling techniques to dynamically tune the optimism coefficient.

**Negative optimism.** $\nu$PI, a generalization of dual optimistic ascent, has been shown to recover gradient ascent with momentum, but only when the optimism coefficient is negative ($\omega < 0$) (Sohrabi et al., 2024, Thm. 1). Since momentum is known to perform poorly in min-max settings—particularly when applied to linear players such as the Lagrange multipliers (Gidel et al., 2019)—this has been used as an argument against negative optimism. Although the dual optimistic ascent algorithm in Eq. (Lag-GD-OA) cannot reproduce momentum, our analysis provides a more direct justification: negative optimism can destabilize the very constrained minimizers we aim to recover.

**Proposition 5** (Negative optimism can destabilize minimizers)**.** *For every local constrained minimizer $\boldsymbol{x}^*$ of Eq.* (CMP) *satisfying Assumptions 1 to 3, there exists a sufficiently negative optimism coefficient $\omega$ such that $\boldsymbol{x}^*$ is no longer an LSSP of the dual optimistic ascent dynamics.*

*Proof.* This follows from considering the Augmented Lagrangian $\mathcal{L}_\omega$ with a sufficiently negative penalty coefficient $\omega < 0$, which renders $\mathcal{L}_\omega$ concave along directions in the null space of the active constraints, and thus $\boldsymbol{x}^*$ ceases to be an LSSP of the dynamics. □

### 4.4 Connection to Existing Empirical Findings

Our theoretical results provide a formal foundation for the strong empirical performance of dual optimistic ascent (PI control) reported in the literature. In particular, its ability to recover all strict and regular local solutions (Theorem 3) explains its success on a non-convex problem where standard Eq. (Lag-GDA) fails, as illustrated in Table 1 (App. C) of Ramirez et al. (2025a).

Furthermore, our work provides a theoretical basis (Corollary 5) for the key empirical finding that larger optimism coefficients increasingly dampen oscillations and stabilize training (Sohrabi et al., 2024; Stooke et al., 2020). This behavior is demonstrated across diverse settings, including safe reinforcement learning (Stooke et al., 2020, Fig. 4, §6.3) and sparsity-constrained ResNet-18 classification (Sohrabi et al., 2024, Figs. 2 and 9, §3 and §5.3). While our contribution is theoretical, the ability of our framework to explain these empirical behaviors strongly validates our findings.

## 5 Conclusion

In this paper, we establish that dual optimistic ascent (PI control) on the Lagrangian is the Augmented Lagrangian method in disguise. This equivalence bridges a crucial gap between theory and practice by transferring the formal guarantees of the ALM—including local linear convergence to all strict, regular constrained minimizers—to the widely used dual optimistic ascent method. Furthermore, our work provides a principled interpretation of the optimism parameter $\omega$, reframing it from a heuristic knob into a formal trade-off between solution accessibility and numerical conditioning.

Despite our findings, we stress that the Augmented Lagrangian method remains a more principled framework for constrained deep learning. Our results apply when using single-step, first-order methods to minimize the Lagrangian; beyond this setting, dual optimistic ascent is not guaranteed to recover all local solutions to constrained problems.

A natural direction for future work is to explore whether similar principles can be applied to optimistic methods in general min–max games, providing deeper insight into optimization in modern deep learning. In addition, this connection opens the door to transferring hyperparameter tuning techniques between proportional gains in control theory and penalty coefficients in constrained optimization.

## Acknowledgements and Disclosure of Funding

This work was supported by the Canada CIFAR AI Chair program (Mila), the NSERC Discovery Grant RGPIN-2025-05123, by an unrestricted gift from Google, and by Samsung Electronics Co., Ldt. Simon Lacoste-Julien is a CIFAR Associate Fellow in the Learning in Machines & Brains program.

We thank Jose Gallego-Posada and Ioannis Mitliagkas for insightful discussions that ultimately led to this work; their observation that dual optimistic ascent and the Augmented Lagrangian method exhibit similar stabilizing effects was a key motivation for this investigation.

## Statement on the Use of Large Language Models

Large Language Models (LLMs) were used to assist in the preparation of this manuscript. Their role was that of research assistants: aiding in the literature review, cross-checking proofs, and evaluating the soundness of technical arguments. They were also employed to improve the clarity of the text.

## References

M. Abadi et al. TensorFlow: A system for Large-Scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 2016. (Cit. on p. 3)

S. A. Alghunaim and A. H. Sayed. Linear Convergence of Primal-Dual Gradient Methods and their Performance in Distributed Optimization. *Automatica*, 2020. (Cit. on p. 8)

K. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-linear Programming*. Stanford University Press, 1958. (Cit. on p. 2)

M. Benzi and V. Simoncini. On the Eigenvalues of a Class of Saddle Point Matrices. *Numerische Mathematik*, 2006. (Cit. on p. 9)

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016. (Cit. on p. 1, 2, 3, 4, 6, 7, 15, 30)

D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic press, 2014. (Cit. on p. 7, 8, 9)

E. G. Birgin and J. M. Martínez. *Practical Augmented Lagrangian Methods for Constrained Optimization*. The SIAM series on Fundamentals of Algorithms, 2014. (Cit. on p. 4)

P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin. Differentiable Causal Discovery from Interventional Data. In *NeurIPS*, 2020. (Cit. on p. 4)

A. R. Conn, G. Gould, and P. L. Toint. *LANCELOT: a Fortran Package for Large-Scale Nonlinear Optimization*. Springer Science & Business Media, 2013. (Cit. on p. 4)

A. Cotter, H. Jiang, M. R. Gupta, S. Wang, T. Narayan, S. You, and K. Sridharan. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *JMLR*, 2019a. (Cit. on p. 1, 3)

A. Cotter et al. TensorFlow Constrained Optimization (TFCO). https://github.com/google-research/tensorflow_constrained_optimization, 2019b. (Cit. on p. 3)

J. Dai, X. Pan, R. Sun, J. Ji, X. Xu, M. Liu, Y. Wang, and Y. Yang. Safe RLHF: Safe Reinforcement Learning from Human Feedback. In *ICLR*, 2024. (Cit. on p. 1)

C. Daskalakis and I. Panageas. The Limit Points of (Optimistic) Gradient Descent in Min-Max Optimization. In *NeurIPS*, 2018. (Cit. on p. 5, 7)

J. Elenter, N. NaderiAlizadeh, and A. Ribeiro. A Lagrangian Duality Approach to Active Learning. In *NeurIPS*, 2022. (Cit. on p. 1, 3)

M. Frank and P. Wolfe. An Algorithm for Quadratic Programming. *Naval Research Logistics Quarterly*, 1956. (Cit. on p. 3)

J. Gallego-Posada. Constrained Optimization for Machine Learning: Algorithms and Applications. *PhD Thesis, University of Montreal*, 2024. (Cit. on p. 3)

J. Gallego-Posada, J. Ramirez, A. Erraqabi, Y. Bengio, and S. Lacoste-Julien. Controlled Sparsity via Constrained Optimization or: *How I Learned to Stop Tuning Penalties and Love Constraints*. In *NeurIPS*, 2022. (Cit. on p. 1, 3)

J. Gallego-Posada, J. Ramirez, M. Hashemizadeh, and S. Lacoste-Julien. Cooper: A Library for Constrained Optimization in Deep Learning. *arXiv preprint arXiv:2504.01212*, 2025. (Cit. on p. 3, 4)

G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A Variational Inequality Perspective on Generative Adversarial Networks. In *ICLR*, 2019. (Cit. on p. 5, 8, 9)

A. A. Goldstein. Convex Programming in Hilbert Space. *University of Washington*, 1964. (Cit. on p. 3)

E. Gorbunov, A. Taylor, and G. Gidel. Last-Iterate Convergence of Optimistic Gradient Method for Monotone Variational Inequalities. In *NeurIPS*, 2022. (Cit. on p. 5)

S.-P. Han. A globally convergent method for nonlinear programming. *Journal of optimization theory and applications*, 1977. (Cit. on p. 3)

M. Hashemizadeh, J. Ramirez, R. Sukumaran, G. Farnadi, S. Lacoste-Julien, and J. Gallego-Posada. Balancing Act: Constraining Disparate Impact in Sparse Models. In *ICLR*, 2024. (Cit. on p. 1, 3)

M. R. Hestenes. Multiplier and Gradient Methods. *Journal of Optimization Theory and Applications*, 1969. (Cit. on p. 1, 3)

I. Hounie, L. F. Chamon, and A. Ribeiro. Automatic Data Augmentation via Invariance-Constrained Learning. In *ICML*, 2023a. (Cit. on p. 3)

I. Hounie, J. Elenter, and A. Ribeiro. Neural Networks with Quantization Constraints. In *ICASSP*, 2023b. (Cit. on p. 1, 3)

S. G. Johnson. The NLopt nonlinear-optimization package). http://github.com/stevengj/nlopt. (Cit. on p. 4)

D. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. In *ICLR*, 2015. (Cit. on p. 1, 21)

J. Kotary and F. Fioretto. Learning constrained optimization with deep augmented lagrangian methods. *arXiv preprint arXiv:2403.03454*, 2024. (Cit. on p. 1, 4)

E. S. Levitin and B. T. Polyak. Constrained Minimization Methods. *USSR Computational mathematics and mathematical physics*, 1966. (Cit. on p. 3)

V. S. Lokhande, A. K. Akash, S. N. Ravi, and V. Singh. FairALM: Augmented Lagrangian Method for Training Fair Models with Little Regret. In *ECCV*, 2020. (Cit. on p. 4)

P. Mahdavinia, Y. Deng, H. Li, and M. Mahdavi. Tight Analysis of Extra-gradient and Optimistic Gradient Methods For Nonconvex Minimax Problems. In *NeurIPS*, 2022. (Cit. on p. 5, 8)

A. Mokhtari, A. Ozdaglar, and S. Pattathil. A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach. In *AISTATS*, 2020. (Cit. on p. 2, 5)

T. Moskovitz, B. O'Donoghue, V. Veeriah, S. Flennerhag, S. Singh, and T. Zahavy. ReLOAD: Reinforcement Learning with Optimistic Ascent-Descent for Last-Iterate Convergence in Constrained MDPs. In *ICML*, 2023. (Cit. on p. 5)

P. Pas, M. Schuurmans, and P. Patrinos. Alpaqa: A matrix-free solver for nonlinear MPC and large-scale nonconvex optimization. In *European Control Conference*, 2022. (Cit. on p. 4)

A. Paszke et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 2019. (Cit. on p. 3)

J. C. Platt and A. H. Barr. Constrained Differential Optimization. In *NeurIPS*, 1987. (Cit. on p. 1, 3, 4)

M. J. Powell. A Method for Nonlinear Constraints in Minimization Problems. *Optimization, Academic Press*, 1969. (Cit. on p. 1, 3)

M. J. Powell. The convergence of variable metric methods for nonlinearly constrained optimization calculations. In *Nonlinear programming 3*. Elsevier, 1978. (Cit. on p. 3)

A. Rakhlin and K. Sridharan. Online Learning with Predictable Sequences. In *CoLT*, 2013. (Cit. on p. 5)

J. Ramirez, M. Hashemizadeh, and S. Lacoste-Julien. Position: Adopt Constraints Over Penalties in Deep Learning. *arXiv preprint arXiv:2505.20628*, 2025a. (Cit. on p. 3, 10)

J. Ramirez, I. Hounie, J. Elenter, J. Gallego-Posada, M. Hashemizadeh, A. Ribeiro, and S. Lacoste-Julien. Feasible Learning. In *AISTATS*, 2025b. (Cit. on p. 1, 3)

A. Robey, L. Chamon, G. J. Pappas, H. Hassani, and A. Ribeiro. Adversarial Robustness with Semi-Infinite Constrained Learning. In *NeurIPS*, 2021. (Cit. on p. 3)

R. T. Rockafellar. A Dual Approach to Solving Nonlinear Programming Problems by Unconstrained Optimization. *Mathematical Programming*, 1973. (Cit. on p. 1, 3, 4)

H. Shao, Y. Yang, H. Lin, L. Lin, Y. Chen, Q. Yang, and H. Zhao. Rethinking Controllable Variational Autoencoders. In *CVPR*, 2022. (Cit. on p. 5)

X. Shi, Y. Xu, G. Chen, and Y. Guo. An Augmented Lagrangian-Based Safe Reinforcement Learning Algorithm for Carbon-Oriented Optimal Scheduling of EV Aggregators. *IEEE Transactions on Smart Grid*, 2023. (Cit. on p. 4)

M. Sohrabi, J. Ramirez, T. H. Zhang, S. Lacoste-Julien, and J. Gallego-Posada. On PI Controllers for Updating Lagrange Multipliers in Constrained Optimization. In *ICML*, 2024. (Cit. on p. 1, 2, 5, 9, 10)

A. Stooke, J. Achiam, and P. Abbeel. Responsive Safety in Reinforcement Learning by PID Lagrangian Methods. In *ICML*, 2020. (Cit. on p. 2, 5, 9, 10)

R. B. Wilson. A simplicial algorithm for concave programming. *PhD Thesis, Graduate School of Bussiness Administration*, 1963. (Cit. on p. 3)

B. Zhang, S. Li, I. Hounie, O. Bastani, D. Ding, and A. Ribeiro. Alignment of large language models with constrained learning. *arXiv preprint arXiv:2505.19387*, 2025. (Cit. on p. 3)

G. Zhang, Y. Wang, L. Lessard, and R. B. Grosse. Near-optimal Local Convergence of Alternating Gradient Descent-Ascent for Minimax Optimization. In *AISTATS*, 2022. (Cit. on p. 2)

G. Zoutendijk. *Methods of Feasible Directions: A Study in Linear and Non-linear Programming*. Elsevier Publishing Company, 1960. (Cit. on p. 3)

# Appendix

# A Assumptions and Further Results

This section provides the formal definitions for the key assumptions used in the main text. We also present an additional result showing that applying dual optimism directly to the Augmented Lagrangian function produces a compounding o-optimistic/penalty effect, rather than giving rise to a completely new method.

## A.1 Assumptions

**Assumption 1** (Strict complementary slackness)**.** A stationary point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ of the Lagrangian $\mathcal{L}$ is said to satisfy *strict complementary slackness* if:

$$\lambda_i^* > 0 \text{ and } g_i(\boldsymbol{x}^*) = 0 \quad \text{or} \quad \lambda_i^* = 0 \text{ and } g_i(\boldsymbol{x}^*) < 0, \qquad \forall i = 1, \dots, m. \tag{5}$$

**Assumption 2** (Second-order sufficiency condition)**.** A stationary point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ of $\mathcal{L}$ satisfies the *second-order sufficiency condition* if the Hessian of the Lagrangian is positive definite on the tangent space of the active constraints. That is,

$$\boldsymbol{d}^\top \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)\, \boldsymbol{d} > 0, \quad \forall \boldsymbol{d} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\} \text{ such that } \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*)\boldsymbol{d} = 0 \text{ and } \nabla \boldsymbol{h}(\boldsymbol{x}^*)\boldsymbol{d} = 0. \tag{6}$$

Note that any KKT point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ satisfying Assumption 2 is a local constrained minimizer of Eq. (CMP) (Bertsekas, 2016, Prop. 4.3.2).

**Assumption 3** (Linear independence constraint qualification)**.** A point $\boldsymbol{x} \in \mathbb{R}^d$ satisfies the *linear independence constraint qualification* if the matrix

$$\left[\nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x})^\top, \nabla \boldsymbol{h}(\boldsymbol{x})^\top\right]^\top \text{ has full row rank.} \tag{7}$$

## A.2 Dual Optimism on the Augmented Lagrangian

We further show that, for equality constraints, applying dual optimistic ascent to the Augmented Lagrangian produces a compounding optimistic effect in the dual updates, without introducing any new phenomena.

**Proposition 6** (Dual optimism on the Augmented Lagrangian - equalities)**.** *Consider applying dual optimism to the Augmented Lagrangian $\mathcal{L}_c$ via primal-first alternating gradient descent–optimistic ascent with optimism coefficient $\omega$. The resulting primal iterates are equivalent to those of Eq.* (AL-GDA) *on $\mathcal{L}_{c+\omega}$ and to those of Eq.* (Lag-GD-OA) *with an optimism coefficient of $c + \omega$.*

*Proof.* See *Proof of Proposition 6* in Section C. □

This result demonstrates that the two frameworks are fully interchangeable for equality-constrained problems; one can reproduce the dynamics of the other by adjusting the hyperparameters accordingly.

While this equivalence is formally established only for equality constraints, we hypothesize that a similar relationship holds for problems with inequality constraints. A formal analysis of this more general case is left for future work.

## B THE AUGMENTED LAGRANGIAN METHOD

For the reader's convenience, in this section we provide computations of the gradient, Hessian, and stationary points of the Augmented Lagrangian function (Eq. (AL)).

Recall the Augmented Lagrangian function:

$$\mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \frac{1}{2c}\left[\|\boldsymbol{\mu} + c\,\boldsymbol{h}(\boldsymbol{x})\|_2^2 - \|\boldsymbol{\mu}\|_2^2 + \left\|[\boldsymbol{\lambda} + c\,\boldsymbol{g}(\boldsymbol{x})]_+\right\|_2^2 - \|\boldsymbol{\lambda}\|_2^2\right] \tag{8}$$

$$= f(\boldsymbol{x}) + \boldsymbol{\mu}^\top \boldsymbol{h}(\boldsymbol{x}) + \frac{c}{2}\|\boldsymbol{h}(\boldsymbol{x})\|_2^2 + \sum_{i=1}^m \begin{cases} \lambda_i g_i(\boldsymbol{x}) + \frac{c}{2} g_i^2(\boldsymbol{x}), & \text{if } \lambda_i + c\,g_i(\boldsymbol{x}) \geq 0 \\ -\lambda_i^2/2c, & \text{otherwise} \end{cases}. \tag{9}$$

### B.1 GRADIENTS OF $\mathcal{L}_c$

The primal gradient corresponds to:

$$\nabla_{\boldsymbol{x}} \mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \nabla f(\boldsymbol{x}) + \big(\boldsymbol{\mu}_t + c\,\boldsymbol{h}(\boldsymbol{x})\big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}) + \sum_{i=1}^m \begin{cases} [\lambda_i + c\,g_i(\boldsymbol{x})]\nabla g_i(\boldsymbol{x}), & \text{if } \lambda_i + c\,g_i(\boldsymbol{x}) \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

$$= \nabla f(\boldsymbol{x}) + \big(\boldsymbol{\mu}_t + c\,\boldsymbol{h}(\boldsymbol{x})\big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}) + \big[\boldsymbol{\lambda}_t + c\,\boldsymbol{g}(\boldsymbol{x})\big]_+^\top \nabla \boldsymbol{g}(\boldsymbol{x}). \tag{11}$$

The dual gradients are:

$$\nabla_{\boldsymbol{\mu}} \mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \boldsymbol{h}(\boldsymbol{x}), \tag{12}$$

and

$$\nabla_{\boldsymbol{\lambda}} \mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{i=1}^m \begin{cases} g_i(\boldsymbol{x}), & \text{if } \lambda_i + c\,g_i(\boldsymbol{x}) \geq 0 \\ -\lambda_i/c, & \text{otherwise} \end{cases} \tag{13}$$

$$= \frac{1}{c}\sum_{i=1}^m \begin{cases} c\,g_i(\boldsymbol{x}), & \text{if } \lambda_i + c\,g_i(\boldsymbol{x}) \geq 0 \\ -\lambda_i, & \text{otherwise} \end{cases} \tag{14}$$

$$= -\frac{\boldsymbol{\lambda}}{c} + \frac{1}{c}\sum_{i=1}^m \begin{cases} \lambda_i + c\,g_i(\boldsymbol{x}), & \text{if } \lambda_i + c\,g_i(\boldsymbol{x}) \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

$$= -\frac{\boldsymbol{\lambda}}{c} + \frac{1}{c}[\boldsymbol{\lambda} + c\,\boldsymbol{g}(\boldsymbol{x})]_+. \tag{16}$$

Gathering these together, we get:

$$\begin{aligned} \nabla_{\boldsymbol{x}} \mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \nabla f(\boldsymbol{x}) + \big(\boldsymbol{\mu}_t + c\,\boldsymbol{h}(\boldsymbol{x})\big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}) + \big[\boldsymbol{\lambda}_t + c\,\boldsymbol{g}(\boldsymbol{x})\big]_+^\top \nabla \boldsymbol{g}(\boldsymbol{x}) \\ \nabla_{\boldsymbol{\mu}} \mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \boldsymbol{h}(\boldsymbol{x}) \\ \nabla_{\boldsymbol{\lambda}} \mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &= \sum_{i=1}^m \begin{cases} g_i(\boldsymbol{x}), & \text{if } \lambda_i + c\,g_i(\boldsymbol{x}) \geq 0 \\ -\lambda_i/c, & \text{otherwise} \end{cases} \\ &= -\frac{\boldsymbol{\lambda}}{c} + \frac{1}{c}[\boldsymbol{\lambda} + c\,\boldsymbol{g}(\boldsymbol{x})]_+. \end{aligned} \tag{17}$$

### B.2 THE HESSIAN OF $\mathcal{L}_c$

As can be observed in Eq. (17) in Section B.1, the gradients of $\mathcal{L}_c$ with respect to $\boldsymbol{x}$ and $\boldsymbol{\lambda}$ are not differentiable whenever some $\lambda_i + c\,g_i(\boldsymbol{x})$ is exactly 0. Therefore, we will compute the following Hessian derivations exclusively for $(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ tuples with $\lambda_i + c\,g_i(\boldsymbol{x}) \neq 0$, for every $i = 1, \ldots, m$.

Note that at any stationary (KKT) tuple $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ of $\mathcal{L}_c$ satisfying *strict* complimentary slackness (Assumption 1) falls in the set of points with a well defined Hessian, for all $c > 0$.

The Hessian of $\mathcal{L}_c$ can be represented as a $3 \times 3$ block matrix:

$$\nabla^2 \mathcal{L}_c = \begin{pmatrix} \nabla^2_{\boldsymbol{x}} \mathcal{L}_c & \nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}} \mathcal{L}_c & \nabla^2_{\boldsymbol{x}\boldsymbol{\mu}} \mathcal{L}_c \\ \nabla^2_{\boldsymbol{\lambda}\boldsymbol{x}} \mathcal{L}_c & \nabla^2_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \mathcal{L}_c & \nabla^2_{\boldsymbol{\lambda}\boldsymbol{\mu}} \mathcal{L}_c \\ \nabla^2_{\boldsymbol{\mu}\boldsymbol{x}} \mathcal{L}_c & \nabla^2_{\boldsymbol{\mu}\boldsymbol{\lambda}} \mathcal{L}_c & \nabla^2_{\boldsymbol{\mu}\boldsymbol{\mu}} \mathcal{L}_c \end{pmatrix}. \tag{18}$$

Let $I_{\mathcal{A}_{\boldsymbol{x}}}$ be the set of indices $i$ for which $\lambda_i + c\, g_i(\boldsymbol{x}) \geq 0$, and let $\mathbf{I}_{\mathcal{A}_{\boldsymbol{x}}} = \texttt{diag}(I_{\mathcal{A}_{\boldsymbol{x}}})$ be the diagonal matrix containing these indicators.

**The Primal-Primal Block ($\nabla^2_{\boldsymbol{x}} \mathcal{L}_c$).** $\nabla^2_{\boldsymbol{x}} \mathcal{L}_c$ yields the Hessian of the standard Lagrangian evaluated at extrapolated multipliers, plus quadratic terms:

$$\nabla^2_{\boldsymbol{x}} \mathcal{L}_c = \nabla^2_{\boldsymbol{x}} \mathcal{L} \big|_{\boldsymbol{x}, [\boldsymbol{\lambda} + c\, \boldsymbol{g}(\boldsymbol{x})]_+, \boldsymbol{\mu} + c\, \boldsymbol{h}(\boldsymbol{x})} + c \left[ \nabla \boldsymbol{h}(\boldsymbol{x})^\top \nabla \boldsymbol{h}(\boldsymbol{x}) + \nabla \boldsymbol{g}(\boldsymbol{x})^\top \mathbf{I}_{\mathcal{A}_{\boldsymbol{x}}} \nabla \boldsymbol{g}(\boldsymbol{x}) \right], \tag{19}$$

where $\nabla^2_{\boldsymbol{x}} \mathcal{L}$ is the Hessian of the standard Lagrangian $\mathcal{L}$:

$$\nabla^2_{\boldsymbol{x}} \mathcal{L} \big|_{\boldsymbol{x}, [\boldsymbol{\lambda} + c\, \boldsymbol{g}(\boldsymbol{x})]_+, \boldsymbol{\mu} + c\, \boldsymbol{h}(\boldsymbol{x})} = \nabla^2 f(\boldsymbol{x}) + [\boldsymbol{\lambda} + c\, \boldsymbol{g}(\boldsymbol{x})]_+^\top \nabla^2 \boldsymbol{g}(\boldsymbol{x}) + (\boldsymbol{\mu} + c\, \boldsymbol{h}(\boldsymbol{x}))^\top \nabla^2 \boldsymbol{h}(\boldsymbol{x}). \tag{20}$$

**The Primal-Dual Inequality Blocks ($\nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}} \mathcal{L}_c$ and $\nabla^2_{\boldsymbol{\lambda}\boldsymbol{x}} \mathcal{L}_c$).** Differentiating $\nabla_{\boldsymbol{x}} \mathcal{L}_c$ with respect to $\boldsymbol{\lambda}$ yields the Jacobian of inequality constraints for those satisying the inequality $\lambda_i + c\, g_i(\boldsymbol{x}) \geq 0$, and 0 for the rest:

$$\nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}} \mathcal{L}_c = \nabla \boldsymbol{g}(\boldsymbol{x})^\top \mathbf{I}_{\mathcal{A}_{\boldsymbol{x}}}. \tag{21}$$

It also follows that

$$\nabla^2_{\boldsymbol{\lambda}\boldsymbol{x}} \mathcal{L}_c = \mathbf{I}_{\mathcal{A}_{\boldsymbol{x}}} \nabla \boldsymbol{g}(\boldsymbol{x}) = \left[ \nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}} \mathcal{L}_c \right]^\top. \tag{22}$$

**The Primal-Dual Equality Block ($\nabla^2_{\boldsymbol{x}\boldsymbol{\mu}} \mathcal{L}_c$ and $\nabla^2_{\boldsymbol{\mu}\boldsymbol{x}} \mathcal{L}_c$).** Yields the Jacobian of the equality constraints:

$$\nabla^2_{\boldsymbol{x}\boldsymbol{\mu}} \mathcal{L}_c = \nabla \boldsymbol{h}(\boldsymbol{x})^\top, \tag{23}$$

with

$$\nabla^2_{\boldsymbol{\mu}\boldsymbol{x}} \mathcal{L}_c = \nabla \boldsymbol{h}(\boldsymbol{x}) = \left[ \nabla^2_{\boldsymbol{x}\boldsymbol{\mu}} \mathcal{L}_c \right]^\top. \tag{24}$$

**The Dual-Dual Inequality Block ($\nabla^2_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \mathcal{L}_c$).** Differentiating $\nabla_{\boldsymbol{\lambda}} \mathcal{L}_c$ with respect to $\boldsymbol{\lambda}$ gives:

$$\nabla^2_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \mathcal{L}_c = \frac{1}{c}(\mathbf{I}_{\mathcal{A}_{\boldsymbol{x}}} - I) = -\frac{1}{c}(I - \mathbf{I}_{\mathcal{A}_{\boldsymbol{x}}}). \tag{25}$$

**All other blocks are zero** as the corresponding gradients do not depend on the other dual variables:

$$\nabla^2_{\boldsymbol{\lambda}\boldsymbol{\mu}} \mathcal{L}_c = \nabla^2_{\boldsymbol{\mu}\boldsymbol{\lambda}} \mathcal{L}_c = \nabla^2_{\boldsymbol{\mu}\boldsymbol{\mu}} \mathcal{L}_c = \mathbf{0}. \tag{26}$$

Combining all the blocks gives the full Hessian of the Augmented Lagrangian function:

$$\nabla^2 \mathcal{L}_c(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \begin{pmatrix} \nabla^2_{\boldsymbol{x}} \mathcal{L}_c & \nabla \boldsymbol{g}(\boldsymbol{x})^\top \mathbf{I}_{\mathcal{A}_{\boldsymbol{x}}} & \nabla \boldsymbol{h}(\boldsymbol{x})^\top \\ \mathbf{I}_{\mathcal{A}_{\boldsymbol{x}}} \nabla \boldsymbol{g}(\boldsymbol{x}) & -\frac{1}{c}(I - \mathbf{I}_{\mathcal{A}_{\boldsymbol{x}}}) & \mathbf{0} \\ \nabla \boldsymbol{h}(\boldsymbol{x}) & \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{27}$$

### B.3 Primal-first alternating gradient descent-ascent on $\mathcal{L}_c$

Primal-first alternating gradient descent-projected gradient ascent on the Augmented Lagrangian $\mathcal{L}_c$ has the following structure:

$$\begin{aligned} \boldsymbol{x}_{t+1} &\leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} \mathcal{L}_c(\boldsymbol{x}_t, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t) \\ \boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t + \eta_{\text{dual}} \nabla_{\boldsymbol{\mu}} \mathcal{L}_c(\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t) \\ \boldsymbol{\lambda}_{t+1} &\leftarrow \left[ \boldsymbol{\lambda}_t + \eta_{\text{dual}} \nabla_{\boldsymbol{\lambda}} \mathcal{L}_c(\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_t, \boldsymbol{\mu}_t) \right]_+. \end{aligned} \tag{28}$$

Using the gradients computed in Section B.1, we can produce the primal gradient descent update:

$$\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}} \left[ \nabla f(\boldsymbol{x}_t) + \big(\boldsymbol{\mu}_t + c\, \boldsymbol{h}(\boldsymbol{x}_t)\big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}_t) + \big[\boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_t)\big]_+^\top \nabla \boldsymbol{g}(\boldsymbol{x}_t) \right]. \tag{29}$$

Moreover, the dual ascent updates correspond to:

$$\boldsymbol{\mu}_{t+1} \leftarrow \boldsymbol{\mu}_t + \eta_{\text{dual}} \boldsymbol{h}(\boldsymbol{x}_{t+1}), \tag{30}$$

and to

$$\begin{aligned}
\boldsymbol{\lambda}_{t+1} &\leftarrow \left[ \boldsymbol{\lambda}_t + \eta_{\text{dual}} \left[ -\frac{\boldsymbol{\lambda}_t}{c} + \frac{1}{c} [\boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_{t+1})]_+ \right] \right]_+ && (31) \\
&\leftarrow \left[ \left(1 - \frac{\eta_{\text{dual}}}{c}\right) \boldsymbol{\lambda}_t + \frac{\eta_{\text{dual}}}{c} \Big[ \boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_{t+1}) \Big]_+ \right]_+ && (32) \\
&\leftarrow \left(1 - \frac{\eta_{\text{dual}}}{c}\right) \boldsymbol{\lambda}_t + \frac{\eta_{\text{dual}}}{c} \Big[ \boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_{t+1}) \Big]_+ . && (33)
\end{aligned}$$

The last step follows from the fact that the update before the projection consists of a convex combination of two non-negative terms whenever $\eta_{\text{dual}} \leq c$, and thus it itself remains non-negative.

Note that the ascent update with respect to $\boldsymbol{\lambda}$ reduces to the usual Augmented Lagrangian Method ascent update whenever $\eta_{\text{dual}} = c$:

$$\boldsymbol{\lambda}_{t+1} \leftarrow \Big[ \boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_{t+1}) \Big]_+ . \tag{34}$$

Gathering these together, we get the gradient descent-ascent updates on $\mathcal{L}_c$ from Eq. (AL-GDA),:

$$\begin{aligned}
\boldsymbol{x}_{t+1} &\leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}} \left[ \nabla f(\boldsymbol{x}_t) + \big(\boldsymbol{\mu}_t + c\, \boldsymbol{h}(\boldsymbol{x}_t)\big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}_t) + \big[\boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_t)\big]_+^\top \nabla \boldsymbol{g}(\boldsymbol{x}_t) \right] && (35) \\
\boldsymbol{\mu}_{t+1} &\leftarrow \boldsymbol{\mu}_t + \eta_{\text{dual}} \boldsymbol{h}(\boldsymbol{x}_{t+1}) && (36) \\
\boldsymbol{\lambda}_{t+1} &\leftarrow \left(1 - \frac{\eta_{\text{dual}}}{c}\right) \boldsymbol{\lambda}_t + \frac{\eta_{\text{dual}}}{c} \Big[ \boldsymbol{\lambda}_t + c\, \boldsymbol{g}(\boldsymbol{x}_{t+1}) \Big]_+ . && (37)
\end{aligned}$$

## C PROOFS

**Lemma 1.** *Let $k > 0$ be a constant. A tuple $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ with $\boldsymbol{\lambda}^* \succeq \mathbf{0}$ satisfies:*

$$\boldsymbol{\lambda}^* = \left[\boldsymbol{\lambda}^* + k\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+ \tag{38}$$

*if and only if $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ satisfies primal feasibility and complimentary slackness. That is,*

$$\boldsymbol{g}(\boldsymbol{x}^*) \preceq \mathbf{0} \quad \text{and} \quad \boldsymbol{\lambda}^* \odot \boldsymbol{g}(\boldsymbol{x}^*) = \mathbf{0}. \tag{39}$$

*Proof.* "⇒." Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ satisfy $\boldsymbol{\lambda}^* = \left[\boldsymbol{\lambda}^* + k\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+$. We can analyze this relationship component-wise, $\lambda_i^* = \max\{0, \lambda_i^* + k\,g_i(\boldsymbol{x}^*)\}$:

- If $\lambda_i^* > 0$, then the equation requires $\lambda_i^* = \lambda_i^* + k\,g_i(\boldsymbol{x}^*)$, which means $g_i(\boldsymbol{x}^*) = 0$.
- If $\lambda_i^* = 0$, then the equation becomes $0 = \max\{0, k\,g_i(\boldsymbol{x}^*)\}$, which requires $k\,g_i(\boldsymbol{x}^*) \leq 0$, and thus $g_i(\boldsymbol{x}^*) \leq 0$.

Combining these two cases, we have that $g_i(\boldsymbol{x}^*) \leq 0$ for all $i$ (primal feasibility) and that for each $i$, either $\lambda_i^* = 0$ or $g_i(\boldsymbol{x}^*) = 0$ or both. This is precisely the definition of complementary slackness.

"⇐." Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ satisfy primal feasibility and complementary slackness. We analyze the relationship component-wise for each $i$:

- If $\lambda_i^* > 0$, then complementary slackness ($\lambda_i^* g_i(\boldsymbol{x}^*) = 0$) implies that $g_i(\boldsymbol{x}^*) = 0$. In this case, the right-hand side of the target equation becomes $\max\{0, \lambda_i^* + k \cdot 0\} = \max\{0, \lambda_i^*\} = \lambda_i^*$. The equality holds.
- If $\lambda_i^* = 0$, then primal feasibility implies that $g_i(\boldsymbol{x}^*) \leq 0$. Since $k > 0$, we have $k\,g_i(\boldsymbol{x}^*) \leq 0$. In this case, the right-hand side becomes $\max\{0, 0 + k\,g_i(\boldsymbol{x}^*)\} = \max\{0, k\,g_i(\boldsymbol{x}^*)\} = 0 = \lambda_i^*$. The equality also holds.

Since the equality $\lambda_i^* = \max\{0, \lambda_i^* + k\,g_i(\boldsymbol{x}^*)\}$ holds for all components $i$, the vector equation $\boldsymbol{\lambda}^* = \left[\boldsymbol{\lambda}^* + k\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+$ is satisfied. □

***Proof of Proposition 1.*** We will now show that stationary (fixed) points of algorithms Eq. (AL-GDA) and Eq. (Lag-GD-OA) match, regardless of their hyper-parameter choices.

"⇒." Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be a fixed point of the Augmented Lagrangian dynamics of Eq. (AL-GDA). It follows that:

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}^* + \eta_{\text{dual}} \boldsymbol{h}(\boldsymbol{x}^*) \tag{40}$$

$$\boldsymbol{\lambda}^* = \left(1 - \frac{\eta_{\text{dual}}}{c}\right) \boldsymbol{\lambda}^* + \frac{\eta_{\text{dual}}}{c} \left[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+. \tag{41}$$

These equations imply primal feasibility and complimentary slackness. In fact, from the first equation, $\boldsymbol{\mu}^* = \boldsymbol{\mu}^* + \eta_{\text{dual}} \boldsymbol{h}(\boldsymbol{x}^*)$ implies that $\eta_{\text{dual}} \boldsymbol{h}(\boldsymbol{x}^*) = 0$, and since $\eta_{\text{dual}} > 0$, we must have $\boldsymbol{h}(\boldsymbol{x}^*) = 0$. This is primal feasibility for the equality constraints.

For the second equation, we can simplify:

$$\boldsymbol{\lambda}^* = \left(1 - \frac{\eta_{\text{dual}}}{c}\right) \boldsymbol{\lambda}^* + \frac{\eta_{\text{dual}}}{c} \left[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+ \tag{42}$$

$$\frac{\eta_{\text{dual}}}{c} \boldsymbol{\lambda}^* = \frac{\eta_{\text{dual}}}{c} \left[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+ \tag{43}$$

$$\boldsymbol{\lambda}^* = \left[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+ \tag{44}$$

As a consequence of Lemma 1, this condition implies both primal feasibility for the inequality constraints ($\boldsymbol{g}(\boldsymbol{x}^*) \preceq \mathbf{0}$) and complementary slackness ($\lambda_i^* g_i(\boldsymbol{x}^*) = 0$ for all $i$).

Primal feasibility can be used to argue:

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}^* + c\,\boldsymbol{h}(\boldsymbol{x}^*), \tag{45}$$

since $c > 0$. Together with the fact that $\boldsymbol{\lambda}^* = \left[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+$, it follows that:

$$\boldsymbol{x}^* = \boldsymbol{x}^* - \eta_{\boldsymbol{x}} \left[\nabla f(\boldsymbol{x}^*) + \left[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+^\top \nabla \boldsymbol{g}(\boldsymbol{x}^*) + \left(\boldsymbol{\mu}^* + c\,\boldsymbol{h}(\boldsymbol{x}^*)\right)^\top \nabla \boldsymbol{h}(\boldsymbol{x}^*)\right] \tag{46}$$

$$= \boldsymbol{x}^* - \eta_{\boldsymbol{x}} \left[\nabla f(\boldsymbol{x}^*) + (\boldsymbol{\lambda}^*)^\top \nabla \boldsymbol{g}(\boldsymbol{x}^*) + (\boldsymbol{\mu}^*)^\top \nabla \boldsymbol{h}(\boldsymbol{x}^*)\right], \tag{47}$$

which corresponds to the primal stationarity condition of dual optimistic ascent in Eq. (Lag-GD-OA); and also simply means:

$$\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = 0. \tag{48}$$

We have thus established that $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is a KKT point of Eq. (CMP).

Finally, we show that the dual variables are also stationary under the optimistic ascent dynamics. At the fixed point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$, we have $\boldsymbol{x}_t = \boldsymbol{x}_{t-1} = \boldsymbol{x}^*$. The dual updates from Eq. (Lag-GD-OA) therefore become:

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}^* + \eta_{\text{dual}} \boldsymbol{h}(\boldsymbol{x}^*) + \omega\big(\boldsymbol{h}(\boldsymbol{x}^*) - \boldsymbol{h}(\boldsymbol{x}^*)\big) \tag{49}$$

$$\boldsymbol{\lambda}^* = \Big[\boldsymbol{\lambda}^* + \eta_{\text{dual}}\, \boldsymbol{g}(\boldsymbol{x}^*) + \omega\big(\boldsymbol{g}(\boldsymbol{x}^*) - \boldsymbol{g}(\boldsymbol{x}^*)\big)\Big]_+ \tag{50}$$

The update for $\boldsymbol{\mu}^*$ simplifies to $\boldsymbol{\mu}^* = \boldsymbol{\mu}^* + \eta_{\text{dual}} \boldsymbol{h}(\boldsymbol{x}^*)$. As primal feasibility requires $\boldsymbol{h}(\boldsymbol{x}^*) = 0$, this becomes $\boldsymbol{\mu}^* = \boldsymbol{\mu}^*$, which is trivially satisfied.

The update for a stationary $\boldsymbol{\lambda}^*$ simplifies to $\boldsymbol{\lambda}^* = \left[\boldsymbol{\lambda}^* + \eta_{\text{dual}}\, \boldsymbol{g}(\boldsymbol{x}^*)\right]_+$. This condition holds due to Lemma 1 and the combination of primal feasibility and complementary slackness proved above.

Since the stationarity conditions for the primal variable $\boldsymbol{x}$ and the dual variables $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are all met, the tuple $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is a fixed point of the dual-first optimistic ascent dynamics in Eq. (Lag-GD-OA).

**"⇐."** Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be a fixed point of the dual-first optimistic ascent dynamics of Eq. (Lag-GD-OA). At a stationary point, the optimistic term $\boldsymbol{g}(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_{t-1})$ vanishes. Therefore, the fixed-point conditions are:

$$\boldsymbol{\mu}^* = \boldsymbol{\mu}^* + \eta_{\text{dual}} \boldsymbol{h}(\boldsymbol{x}^*) \tag{51}$$

$$\boldsymbol{\lambda}^* = \Big[\boldsymbol{\lambda}^* + \eta_{\text{dual}}\, \boldsymbol{g}(\boldsymbol{x}^*)\Big]_+ \tag{52}$$

$$\boldsymbol{x}^* = \boldsymbol{x}^* - \eta_{\boldsymbol{x}} \left[\nabla f(\boldsymbol{x}^*) + (\boldsymbol{\lambda}^*)^\top \nabla \boldsymbol{g}(\boldsymbol{x}^*) + (\boldsymbol{\mu}^*)^\top \nabla \boldsymbol{h}(\boldsymbol{x}^*)\right] \tag{53}$$

From these three conditions, we can deduce that the tuple $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ must satisfy the KKT conditions:

1. From the $\boldsymbol{\mu}$ condition, we get $\boldsymbol{h}(\boldsymbol{x}^*) = 0$.
2. The $\boldsymbol{\lambda}$ update condition, $\boldsymbol{\lambda}^* = \left[\boldsymbol{\lambda}^* + \eta_{\text{dual}}\, \boldsymbol{g}(\boldsymbol{x}^*)\right]_+$, together with Lemma 1, implies both primal feasibility for inequalities and complementary slackness.
3. The $\boldsymbol{x}$ update implies that the gradient of the Lagrangian is zero: $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = 0$.

Now we will show that this KKT point is also a fixed point of the Augmented Lagrangian dynamics in Eq. (AL-GDA).

- **For $\boldsymbol{\mu}$**: Primal feasibility implies $\boldsymbol{\mu}^* = \boldsymbol{\mu}^* + \eta_{\text{dual}} \boldsymbol{h}(\boldsymbol{x}^*)$.
- **For $\boldsymbol{\lambda}$**: Primal feasibility and complimentary slackness imply $\boldsymbol{\lambda}^* = \left[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\right]_+$ (Lemma 1). This in turn implies:

$$\boldsymbol{\lambda}^* = \left(1 - \frac{\eta_{\text{dual}}}{c}\right) \boldsymbol{\lambda}^* + \frac{\eta_{\text{dual}}}{c} \boldsymbol{\lambda}^* = \left(1 - \frac{\eta_{\text{dual}}}{c}\right) \boldsymbol{\lambda}^* + \frac{\eta_{\text{dual}}}{c} \Big[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\Big]_+. \tag{54}$$

The primal update for the Augmented Lagrangian method is stationary if the gradient term is zero:

$$\nabla f(\boldsymbol{x}^*) + \big[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\big]_+^\top \nabla \boldsymbol{g}(\boldsymbol{x}^*) + \big(\boldsymbol{\mu}^* + c\,\boldsymbol{h}(\boldsymbol{x}^*)\big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}^*) = 0 \tag{55}$$

But as argued before, $\boldsymbol{\lambda}^* = \big[\boldsymbol{\lambda}^* + c\,\boldsymbol{g}(\boldsymbol{x}^*)\big]_+$ and $\boldsymbol{\mu}^* = \boldsymbol{\mu}^* + c\,\boldsymbol{h}(\boldsymbol{x}^*)$. Therefore:

$$\nabla f(\boldsymbol{x}^*) + (\boldsymbol{\lambda}^*)^\top \nabla \boldsymbol{g}(\boldsymbol{x}^*) + (\boldsymbol{\mu}^*)^\top \nabla \boldsymbol{h}(\boldsymbol{x}^*) = \nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) \tag{56}$$

From the Eq. (Lag-GD-OA) fixed-point conditions, we know this term is zero. Thus, the primal update is stationary.

Since the stationarity conditions for the primal and dual variables are all met, the point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is also a fixed point of the Augmented Lagrangian dynamics. This completes the proof.

□

### C.1 Equivalence Proof for Equality-constrained Problems

***Proof of Theorem 1.*** Consider the following primal-first alternating gradient descent–ascent updates on the Augmented Lagrangian:

$$\begin{aligned} \boldsymbol{x}_{t+1} &\leftarrow \texttt{MinStep}\Big(\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) + \big(\boldsymbol{\mu}_t^{(\mathrm{ALM})} + c\,\boldsymbol{h}(\boldsymbol{x}_t)\big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}_t)\Big), \\ \boldsymbol{\mu}_{t+1}^{(\mathrm{ALM})} &\leftarrow \boldsymbol{\mu}_t^{(\mathrm{ALM})} + \eta_{\text{dual}}\,\boldsymbol{h}(\boldsymbol{x}_{t+1}). \end{aligned} \tag{57}$$

These are analogous to Eq. (AL-GDA), but stated for a generic first-order primal minimization step `MinStep`. This step could, for instance, be replaced with Adam (Kingma and Ba, 2015), and the remainder of the proof would still hold.

Now consider the following dual-first alternating gradient descent–optimistic ascent updates on the Lagrangian, similar to Eq. (Lag-GD-OA):

$$\begin{aligned} \boldsymbol{\mu}_{t+1}^{(\mathrm{OGA})} &\leftarrow \boldsymbol{\mu}_t^{(\mathrm{OGA})} + \eta_{\text{dual}}\,\boldsymbol{h}(\boldsymbol{x}_t) + \omega\big(\boldsymbol{h}(\boldsymbol{x}_t) - \boldsymbol{h}(\boldsymbol{x}_{t-1})\big), \\ \boldsymbol{x}_{t+1} &\leftarrow \texttt{MinStep}\Big(\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) + \Big(\boldsymbol{\mu}_{t+1}^{(\mathrm{OGA})}\Big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}_t)\Big). \end{aligned} \tag{58}$$

Suppose $\omega = c$. We show by induction that the primal iterates $\{\boldsymbol{x}_t\}_{t=0}^{\infty}$ generated by the primal-first ALM algorithm and the dual-first OGA algorithm are identical.

**Inductive Step**. Define:

$$\boldsymbol{\mu}_{t+1}^{(\mathrm{OGA})} = \boldsymbol{\mu}_t^{(\mathrm{ALM})} + c\,\boldsymbol{h}(\boldsymbol{x}_t). \tag{59}$$

With this substitution, the primal update in Eq. (57) becomes

$$\boldsymbol{x}_{t+1} = \texttt{MinStep}\Big(\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) + \Big(\boldsymbol{\mu}_{t+1}^{(\mathrm{OGA})}\Big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}_t)\Big) \tag{60}$$

$$= \texttt{MinStep}\Big(\boldsymbol{x}_t, \nabla_{\boldsymbol{x}} \mathcal{L}\Big(\boldsymbol{x}_t, \boldsymbol{\mu}_{t+1}^{(\mathrm{OGA})}\Big)\Big), \tag{61}$$

i.e., a descent step on the Lagrangian with multipliers $\boldsymbol{\mu}_{t+1}^{(\mathrm{OGA})}$.

For the dual update, we have

$$\boldsymbol{\mu}_{t+1}^{(\mathrm{OGA})} = \boldsymbol{\mu}_t^{(\mathrm{ALM})} + c\,\boldsymbol{h}(\boldsymbol{x}_t) \tag{62}$$

$$= \boldsymbol{\mu}_{t-1}^{(\mathrm{ALM})} + \eta_{\text{dual}}\,\boldsymbol{h}(\boldsymbol{x}_t) + c\,\boldsymbol{h}(\boldsymbol{x}_t) \tag{63}$$

$$= \boldsymbol{\mu}_t^{(\mathrm{OGA})} - c\,\boldsymbol{h}(\boldsymbol{x}_{t-1}) + \eta_{\text{dual}}\,\boldsymbol{h}(\boldsymbol{x}_t) + c\,\boldsymbol{h}(\boldsymbol{x}_t) \tag{64}$$

$$= \boldsymbol{\mu}_t^{(\mathrm{OGA})} + \eta_{\text{dual}}\,\boldsymbol{h}(\boldsymbol{x}_t) + c\big(\boldsymbol{h}(\boldsymbol{x}_t) - \boldsymbol{h}(\boldsymbol{x}_{t-1})\big), \tag{65}$$

which matches the optimistic ascent update in Eq. (123) with $\omega = c$. Combining this dual update with Eq. (60) yields exactly the algorithm in Eq. (123). Thus, if the primal iterates generated by both algorithms match up until step $t$, they will match at step $t+1$ under the change of variable $\boldsymbol{\mu}_{t+1}^{(\mathrm{OGA})} = \boldsymbol{\mu}_t^{(\mathrm{ALM})} + c\,\boldsymbol{h}(\boldsymbol{x}_t)$.

**Base Case**. We establish the base case for the induction by showing that the first primal iterates, $\boldsymbol{x}_1$, are identical under the specified initializations.

The primal descent update for Eq. (122) corresponds to a descent step on the Lagrangian with an effective multiplier of $\boldsymbol{\mu}_0^{(\text{ALM})} + c\,\boldsymbol{h}(\boldsymbol{x}_0)$. The primal descent update for Eq. (123) corresponds to a descent step on the Lagrangian with an effective multiplier of $\boldsymbol{\mu}_1^{(\text{OGA})}$. For the first step, the difference term in Eq. (123) is typically set to zero (as $\boldsymbol{h}(\boldsymbol{x}_{-1})$ is undefined), so that

$$\boldsymbol{\mu}_1^{(\text{OGA})} = \boldsymbol{\mu}_0^{(\text{OGA})} + \eta_{\text{dual}}\,\boldsymbol{h}(\boldsymbol{x}_0). \tag{66}$$

For the primal iterates to be identical, their effective multipliers must satisfy

$$\boldsymbol{\mu}_0^{(\text{ALM})} + c\,\boldsymbol{h}(\boldsymbol{x}_0) = \boldsymbol{\mu}_0^{(\text{OGA})} + \eta_{\text{dual}}\,\boldsymbol{h}(\boldsymbol{x}_0). \tag{67}$$

This equality is ensured by initializing the OGA method relative to the ALM method as

$$\boldsymbol{\mu}_0^{(\text{OGA})} = \boldsymbol{\mu}_0^{(\text{ALM})} + (c - \eta_{\text{dual}})\,\boldsymbol{h}(\boldsymbol{x}_0). \tag{68}$$

With this initialization, both effective multipliers for the first primal update are identical, yielding the same first iterate for both algorithms:

$$\boldsymbol{x}_1 = \texttt{MinStep}\Big(\boldsymbol{x}_0, \nabla f(\boldsymbol{x}_0) + \big(\boldsymbol{\mu}_0^{(\text{ALM})} + c\,\boldsymbol{h}(\boldsymbol{x}_0)\big)^\top \nabla \boldsymbol{h}(\boldsymbol{x}_0)\Big). \tag{69}$$

Combined with the recurrence in Eq. (60), this initial condition implies by induction that all subsequent primal iterates $\boldsymbol{x}_t$ also coincide. Therefore, the two algorithms are equivalent.

□

### C.2 Equivalence Proof for Inequality-constrained Problems

This subsection provides the full proof of Theorem 2, which establishes the equivalence of the locally stable stationary points (LSSPs) for the two algorithms in inequality-constrained problems. Our proof strategy is as follows: in Lemmas 2 and 4, we first derive the Jacobians of the update operators defined by Eq. (AL-GDA) and Eq. (Lag-GD-OA), respectively. We then derive their corresponding characteristic polynomials in Lemmas 3 and 5. Finally, in the main proof of Theorem 2, we establish a relationship between the roots of these polynomials, which govern local convergence when $\omega = c$.

**Lemma 2.** *Consider primal-first alternating gradient descent–projected gradient ascent on the Augmented Lagrangian from Eq. (AL-GDA), applied to a problem with inequality constraints only:*

$$\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}} \left[\nabla f(\boldsymbol{x}_t) + \big[\boldsymbol{\lambda}_t + c\,\boldsymbol{g}(\boldsymbol{x}_t)\big]_+^\top \nabla \boldsymbol{g}(\boldsymbol{x}_t)\right], \tag{70}$$

$$\boldsymbol{\lambda}_{t+1} \leftarrow \left(1 - \frac{\eta_{dual}}{c}\right)\boldsymbol{\lambda}_t + \frac{\eta_{dual}}{c}\Big[\boldsymbol{\lambda}_t + c\,\boldsymbol{g}(\boldsymbol{x}_{t+1})\Big]_+, \tag{71}$$

*with $\eta_{dual} < c$.*

*Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ be a stationary (KKT) point of Eq. (CMP) satisfying Assumption 1. The Jacobian of the update operator in Eq. (70), evaluated at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ and acting on the partitioned state*

$$(\boldsymbol{x}_t, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I}) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_{t+1,A}, \boldsymbol{\lambda}_{t+1,I}),$$

*is*

$$\mathcal{J}_{AL} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}}(A + c\,B^\top B) & -\eta_{\boldsymbol{x}} B^\top & 0 \\ \eta_{dual} B\,(\mathbb{I} - \eta_{\boldsymbol{x}}(A + c\,B^\top B)) & \mathbb{I} - \eta_{\boldsymbol{x}}\eta_{dual} BB^\top & 0 \\ 0 & 0 & (1 - \eta_{dual}/c)\mathbb{I} \end{pmatrix}. \tag{72}$$

*Here, $\boldsymbol{\lambda}_t = (\boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I})$ partitions the multipliers into active components*

$$\boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}} = \{\,\lambda_i \mid g_i(\boldsymbol{x}^*) = 0,\ \lambda_i^* > 0\,\},$$

*and inactive ones*

$$\boldsymbol{\lambda}_{t,I} = \{\,\lambda_i \mid g_i(\boldsymbol{x}^*) < 0,\ \lambda_i^* = 0\,\}.$$

*Under the same partition, applied to the constraints $\boldsymbol{g}(\boldsymbol{x}) = [\boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x})^\top, \boldsymbol{g}_I(\boldsymbol{x})^\top]^\top$, the matrices are*

$$A = \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \nabla^2 f(\boldsymbol{x}^*) + \sum_{i \in \mathcal{A}_{\boldsymbol{x}}} \lambda_i^* \nabla^2 g_i(\boldsymbol{x}^*), \tag{73}$$

$$B = \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*). \tag{74}$$

*Proof.* The Augmented Lagrangian method in Eq. (70) alternates between primal and dual updates, beginning with the primal variables. It can therefore be written as the composition of two operators, one for each step. The Jacobian of the full update is

$$\mathcal{J}_{\text{AL}} = \mathcal{J}_{\boldsymbol{\lambda}} \mathcal{J}_{\boldsymbol{x}}, \tag{75}$$

where $\mathcal{J}_{\boldsymbol{x}}$ and $\mathcal{J}_{\boldsymbol{\lambda}}$ denote the Jacobians of the primal and dual steps, respectively.

The Jacobian of the primal step $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_t)$ is

$$\mathcal{J}_{\boldsymbol{x}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}} \nabla^2_{\boldsymbol{x}} \mathcal{L}_c & -\eta_{\boldsymbol{x}} \nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}} \mathcal{L}_c \\ 0 & \mathbb{I} \end{pmatrix}, \tag{76}$$

and the Jacobian of the dual step $(\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_t) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_{t+1})$ is

$$\mathcal{J}_{\boldsymbol{\lambda}} = \begin{pmatrix} \mathbb{I} & 0 \\ \eta_{\text{dual}} \nabla^2_{\boldsymbol{\lambda}\boldsymbol{x}} \mathcal{L}_c & \mathbb{I} + \eta_{\text{dual}} \nabla^2_{\boldsymbol{\lambda}} \mathcal{L}_c \end{pmatrix}. \tag{77}$$

The primal Jacobian is well defined for any $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ with $\boldsymbol{\lambda}_t \succeq \mathbf{0}$ whenever $\lambda_{t,i} + c\, g_i(\boldsymbol{x}_t) \neq 0$ for all constraints $i = 1, \ldots, m$. Similarly, the dual jacobian is well defined whenever $\lambda_{t,i} + c\, g_i(\boldsymbol{x}_{t+1}) \neq 0$. In particular, both Jacobians are well defined at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ due to strict complimentary slackness.

Multiplying gives

$$\mathcal{J}_{\text{AL}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}} \nabla^2_{\boldsymbol{x}} \mathcal{L}_c & -\eta_{\boldsymbol{x}} \nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}} \mathcal{L}_c \\ \eta_{\text{dual}} \nabla^2_{\boldsymbol{\lambda}\boldsymbol{x}} \mathcal{L}_c \, (\mathbb{I} - \eta_{\boldsymbol{x}} \nabla^2_{\boldsymbol{x}} \mathcal{L}_c) & \mathbb{I} + \eta_{\text{dual}} \nabla^2_{\boldsymbol{\lambda}} \mathcal{L}_c - \eta_{\boldsymbol{x}} \eta_{\text{dual}} (\nabla^2_{\boldsymbol{\lambda}\boldsymbol{x}} \mathcal{L}_c)(\nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}} \mathcal{L}_c) \end{pmatrix}. \tag{78}$$

To recover the structure of $\mathcal{J}_{\text{AL}}$, we evaluate $\nabla^2_{\boldsymbol{x}} \mathcal{L}_c$, $\nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}} \mathcal{L}_c$, and $\nabla^2_{\boldsymbol{\lambda}} \mathcal{L}_c$ at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$. Explicit forms of these terms at generic $(\boldsymbol{x}, \boldsymbol{\lambda})$ were already derived in Section B.2.

We now partition the state $(\boldsymbol{x}_t, \boldsymbol{\lambda}_t)$ into $(\boldsymbol{x}_t, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I})$, where $\boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}$ corresponds to active constraints at $\boldsymbol{x}^*$ and $\boldsymbol{\lambda}_{t,I}$ to inactive ones. This partition is well defined under strict complementarity.

**The primal Hessian ($\nabla^2_{\boldsymbol{x}} \mathcal{L}_c$).** Complementary slackness implies $\boldsymbol{\lambda}^* = [\boldsymbol{\lambda}^* + c\, \boldsymbol{g}(\boldsymbol{x}^*)]_+$, both for active and inactive constraints (Lemma 1).

Substituting into the primal Hessian (see Eq. (19)) yields

$$\nabla^2_{\boldsymbol{x}} \mathcal{L}_c \big|_{\boldsymbol{x}^*, \boldsymbol{\lambda}^*} = \nabla^2 f(\boldsymbol{x}^*) + (\boldsymbol{\lambda}^*)^\top \nabla^2 \boldsymbol{g}(\boldsymbol{x}^*) + c\, \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*)^\top \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*) \tag{79}$$

$$= A + c\, B^\top B. \tag{80}$$

**The primal–dual Hessians for *active* constraints** are

$$\nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}_{\mathcal{A}_{\boldsymbol{x}}}} \mathcal{L}_c \big|_{\boldsymbol{x}^*, \boldsymbol{\lambda}^*} = \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*)^\top = B^\top, \qquad \nabla^2_{\boldsymbol{\lambda}_{\mathcal{A}_{\boldsymbol{x}}} \boldsymbol{x}} \mathcal{L}_c \big|_{\boldsymbol{x}^*, \boldsymbol{\lambda}^*} = \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*) = B. \tag{81}$$

**The primal–dual Hessians for *inactive* constraints** vanish:

$$\nabla^2_{\boldsymbol{x}\boldsymbol{\lambda}_I} \mathcal{L}_c \big|_{\boldsymbol{x}^*, \boldsymbol{\lambda}^*} = 0, \quad \nabla^2_{\boldsymbol{\lambda}_I \boldsymbol{x}} \mathcal{L}_c \big|_{\boldsymbol{x}^*, \boldsymbol{\lambda}^*} = 0. \tag{82}$$

**The dual–dual Hessian for active constraints** vanishes:

$$\nabla^2_{\boldsymbol{\lambda}_{\mathcal{A}_{\boldsymbol{x}}}} \mathcal{L}_c \big|_{\boldsymbol{x}^*, \boldsymbol{\lambda}^*} = 0. \tag{83}$$

**The dual–dual Hessian for inactive constraints** is

$$\nabla^2_{\boldsymbol{\lambda}_I} \mathcal{L}_c \,\big|_{\boldsymbol{x}^*, \boldsymbol{\lambda}^*} = -\frac{1}{c}\,\mathbb{I}. \tag{84}$$

Substituting all these expressions, the Jacobian $\mathcal{J}_{\text{AL}}$, evaluated at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = (\boldsymbol{x}^*, \boldsymbol{\lambda}^*_{\mathcal{A}_{\boldsymbol{x}}}, \mathbf{0})$ and acting on the partitioned state $(\boldsymbol{x}_t, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I})$, is

$$\mathcal{J}_{\text{AL}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}}(A + c\,B^\top B) & -\eta_{\boldsymbol{x}} B^\top & 0 \\ \eta_{\text{dual}} B(\mathbb{I} - \eta_{\boldsymbol{x}}(A + c\,B^\top B)) & \mathbb{I} - \eta_{\boldsymbol{x}}\eta_{\text{dual}} B B^\top & 0 \\ 0 & 0 & (1 - \eta_{\text{dual}}/c)\mathbb{I} \end{pmatrix}. \tag{85}$$

□

**Lemma 3.** *The characteristic polynomial of the Jacobian $\mathcal{J}_{AL}$ at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ from Lemma 2 is*

$$\chi_{\mathcal{J}_{AL}}(\sigma) = \Big(1 - \frac{\eta_{dual}}{c} - \sigma\Big)^{m - |\mathcal{A}_{\boldsymbol{x}}|} \det\left((1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma)A - \eta_{\boldsymbol{x}}(c(1-\sigma) - \eta_{dual}\sigma)B^\top B\right). \tag{86}$$

*Proof.* The Jacobian $\mathcal{J}_{\text{AL}}$ is block lower-triangular with respect to the partition between the active subsystem $(\boldsymbol{x}, \boldsymbol{\lambda}_{\mathcal{A}_{\boldsymbol{x}}})$ and the inactive subsystem $\boldsymbol{\lambda}_I$. Its characteristic polynomial is therefore the product of the characteristic polynomials of these two diagonal blocks.

**Inactive subsystem.** The bottom-right block corresponding to the dual variables of inactive inequality constraints is

$$\mathcal{J}_{\text{AL},I} = \Big(1 - \frac{\eta_{\text{dual}}}{c}\Big)\,\mathbb{I},$$

whose characteristic polynomial is

$$\chi_{\mathcal{J}_{\text{AL},I}}(\sigma) = \Big(1 - \frac{\eta_{\text{dual}}}{c} - \sigma\Big)^{|I_I|}. \tag{87}$$

**Active subsystem.** The Jacobian of the active subsystem is

$$\mathcal{J}_{\text{AL},\mathcal{A}_{\boldsymbol{x}}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}}(A + c\,B^\top B) & -\eta_{\boldsymbol{x}} B^\top \\ \eta_{\text{dual}} B(\mathbb{I} - \eta_{\boldsymbol{x}}(A + c\,B^\top B)) & \mathbb{I} - \eta_{\boldsymbol{x}}\eta_{\text{dual}} B B^\top \end{pmatrix}. \tag{88}$$

This matrix is similar to

$$\mathcal{J}'_{\text{AL},\mathcal{A}_{\boldsymbol{x}}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}}A - \eta_{\boldsymbol{x}}(c + \eta_{\text{dual}})B^\top B & -\eta_{\boldsymbol{x}} B^\top \\ \eta_{\text{dual}} B & \mathbb{I} \end{pmatrix} \tag{89}$$

via the transformation matrix

$$P = \begin{pmatrix} \mathbb{I} & 0 \\ -\eta_{\text{dual}} B & I \end{pmatrix}. \tag{90}$$

The characteristic polynomial of the active subsystem is found by computing the determinant of $\mathcal{J}'_{\text{AL},\mathcal{A}_{\boldsymbol{x}}} - \sigma\mathbb{I}$:

$$\chi_{\mathcal{J}_{\text{AL},\mathcal{A}_{\boldsymbol{x}}}}(\sigma) = \det \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}}A - \eta_{\boldsymbol{x}}(c + \eta_{\text{dual}})B^\top B - \sigma\mathbb{I} & -\eta_{\boldsymbol{x}} B^\top \\ \eta_{\text{dual}} B & (1-\sigma)\mathbb{I} \end{pmatrix}. \tag{91}$$

Let the blocks of the matrix be denoted $M_{11}, M_{12}, M_{21}, M_{22}$. The bottom two blocks, $M_{21} = \eta_{\text{dual}} B$ and $M_{22} = (1-\sigma)\mathbb{I}$, commute since the identity matrix commutes with any matrix. We can therefore use the block determinant identity $\det(M) = \det(M_{11}M_{22} - M_{12}M_{21})$.

We first compute the product $M_{11}M_{22}$:

$$M_{11}M_{22} = \left((1-\sigma)\mathbb{I} - \eta_{\boldsymbol{x}}A - \eta_{\boldsymbol{x}}(c + \eta_{\text{dual}})B^\top B\right)(1-\sigma)\mathbb{I} \tag{92}$$

$$= (1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma)A - \eta_{\boldsymbol{x}}(c + \eta_{\text{dual}})(1-\sigma)B^\top B. \tag{93}$$

Next, we compute the product $M_{12}M_{21}$:

$$M_{12}M_{21} = (-\eta_{\boldsymbol{x}} B^\top)(\eta_{\text{dual}} B) = -\eta_{\boldsymbol{x}}\eta_{\text{dual}} B^\top B. \tag{94}$$

Subtracting the second product from the first yields the matrix $M_{11}M_{22} - M_{12}M_{21}$:

$$(1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma)A - \left[\eta_{\boldsymbol{x}}(c+\eta_{\text{dual}})(1-\sigma) - \eta_{\boldsymbol{x}}\eta_{\text{dual}}\right] B^\top B. \tag{95}$$

We simplify the coefficient of the $B^\top B$ term:

$$\eta_{\boldsymbol{x}}(c+\eta_{\text{dual}})(1-\sigma) - \eta_{\boldsymbol{x}}\eta_{\text{dual}} = \eta_{\boldsymbol{x}}\left[c(1-\sigma) + \eta_{\text{dual}}(1-\sigma) - \eta_{\text{dual}}\right] = \eta_{\boldsymbol{x}}(c(1-\sigma) - \eta_{\text{dual}}\sigma). \tag{96}$$

Thus, the characteristic polynomial of the active subsystem is

$$\chi_{\mathcal{J}_{\text{AL}},\mathcal{A}_{\boldsymbol{x}}}(\sigma) = \det\left((1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma)A - \eta_{\boldsymbol{x}}(c(1-\sigma) - \eta_{\text{dual}}\sigma)B^\top B\right). \tag{97}$$

**Conclusion.** The characteristic polynomial of the full Jacobian $\mathcal{J}_{\text{AL}}$ is the product of the polynomials for the active and inactive subsystems, $\chi_{\mathcal{J}_{\text{AL}}}(\sigma) = \chi_{\mathcal{J}_{\text{AL},I}}(\sigma) \cdot \chi_{\mathcal{J}_{\text{AL},\mathcal{A}_{\boldsymbol{x}}}}(\sigma)$, which gives the final result.

□

**Lemma 4.** *Consider dual-first alternating gradient descent–projected optimistic ascent on the Lagrangian, analogous to Eq. (Lag-GD-OA), applied to a problem with inequality constraints only:*

$$\boldsymbol{\lambda}_{t+1} \leftarrow \left[\boldsymbol{\lambda}_t + \eta_{dual}\, \boldsymbol{g}(\boldsymbol{x}_t) + \omega\left(\boldsymbol{g}(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_{t-1})\right)\right]_+ \tag{98}$$

$$\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}}\left[\nabla f(\boldsymbol{x}_t) + \boldsymbol{\lambda}_{t+1}^\top \nabla \boldsymbol{g}(\boldsymbol{x}_t)\right]. \tag{99}$$

*Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ be a stationary (KKT) point of Eq. (CMP) satisfying Assumption 1. The Jacobian of the update operator in Eq. (98), evaluated at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ and acting on the state*

$$(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I}) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1,A}, \boldsymbol{\lambda}_{t+1,I}),$$

*is*

$$\mathcal{J}_{OG} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}}A - \eta_{\boldsymbol{x}}(\eta_{dual}+\omega)B^\top B & \eta_{\boldsymbol{x}}\omega\, B^\top B & -\eta_{\boldsymbol{x}}B^\top & 0 \\ \mathbb{I} & 0 & 0 & 0 \\ (\eta_{dual}+\omega)B & -\omega\, B & \mathbb{I} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{100}$$

*Here, as in Lemmas 2 and 3, the multipliers are partitioned as $\boldsymbol{\lambda}_t = (\boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I})$, where the active components are*

$$\boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}} = \{\, \lambda_i \mid g_i(\boldsymbol{x}^*) = 0,\ \lambda_i^* > 0 \,\},$$

*and the inactive components are*

$$\boldsymbol{\lambda}_{t,I} = \{\, \lambda_i \mid g_i(\boldsymbol{x}^*) < 0,\ \lambda_i^* = 0 \,\}.$$

*Under this same partition of the constraints $\boldsymbol{g}(\boldsymbol{x}) = [\boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x})^\top, \boldsymbol{g}_I(\boldsymbol{x})^\top]^\top$, the matrices are once more defined as*

$$A = \nabla^2_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = \nabla^2 f(\boldsymbol{x}^*) + \sum_{i\in\mathcal{A}_{\boldsymbol{x}}} \lambda_i^* \nabla^2 g_i(\boldsymbol{x}^*), \tag{101}$$

$$B = \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*). \tag{102}$$

*Proof.* The dual–first optimistic ascent method in Eq. (98) alternates between updates of the dual and primal variables, beginning with the dual step. Accordingly, the overall update operator can be written as the composition of the two corresponding operators. Its Jacobian is thus

$$\mathcal{J}_{\text{OG}} = \mathcal{J}_{\boldsymbol{x}}\mathcal{J}_{\boldsymbol{\lambda}}, \tag{103}$$

where $\mathcal{J}_{\boldsymbol{x}}$ and $\mathcal{J}_{\boldsymbol{\lambda}}$ denote the Jacobians of the primal and dual steps, respectively. Since the dual update depends on both $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$, we evaluate these Jacobians on the augmented state

$$(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_t) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1}). \tag{104}$$

The primal Jacobian is well defined for any $(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_{t+1})$ with $\boldsymbol{\lambda}_{t+1} \succeq \mathbf{0}$. In contrast, the dual Jacobian is well defined only when

$$\lambda_{t,i} + \eta_{\text{dual}}\, g_i(\boldsymbol{x}_t) + \omega \left(g_i(\boldsymbol{x}_t) - g_i(\boldsymbol{x}_{t-1})\right) \neq 0. \tag{105}$$

At $(\boldsymbol{x}^*, \boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ this condition holds due to strict complementarity. Indeed,

$$\lambda_i^* + \eta_{\text{dual}}\, g_i(\boldsymbol{x}^*) + \omega \left(g_i(\boldsymbol{x}^*) - g_i(\boldsymbol{x}^*)\right) = \lambda_i^* + \eta_{\text{dual}}\, g_i(\boldsymbol{x}^*) \neq 0. \tag{106}$$

At this point, two cases arise:

- **Active constraints** ($\lambda_i^* > 0$ and $g_i(\boldsymbol{x}^*) = 0$). Then
$$[\lambda_i^* + \eta_{\text{dual}}\, g_i(\boldsymbol{x}^*)]_+ = [\lambda_i^*]_+ = \lambda_i^*. \tag{107}$$
- **Inactive constraints** ($\lambda_i^* = 0$ and $g_i(\boldsymbol{x}^*) < 0$). Then
$$[\lambda_i^* + \eta_{\text{dual}}\, g_i(\boldsymbol{x}^*)]_+ = [\omega\, g_i(\boldsymbol{x}^*)]_+ = 0 = \lambda_i^*. \tag{108}$$

In both cases, the projection condition reduces to whether $\lambda_i^* > 0$, i.e., whether the constraint is active at $\boldsymbol{x}^*$. Thus, we adopt the same partition $\boldsymbol{\lambda}_t = (\boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I})$ as in Lemma 2, and reuse the definitions of $A$ and $B$.

It follows that the Jacobian of the dual update, acting on the augmented and partitioned state $(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I}) \mapsto (\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_{t+1,A}, \boldsymbol{\lambda}_{t+1,I})$, and evaluated at $(\boldsymbol{x}^*, \boldsymbol{x}^*, \boldsymbol{\lambda}^*_{\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}^*_I)$, is

$$\mathcal{J}_{\boldsymbol{\lambda}} = \begin{pmatrix} \mathbb{I} & 0 & 0 & 0 \\ 0 & \mathbb{I} & 0 & 0 \\ (\eta_{\text{dual}} + \omega)B & -\omega\, B & \mathbb{I} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{109}$$

while the Jacobian of the primal update,

$$(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_{t+1,A}, \boldsymbol{\lambda}_{t+1,I}) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1,A}, \boldsymbol{\lambda}_{t+1,I}), \tag{110}$$

evaluated at $(\boldsymbol{x}^*, \boldsymbol{x}^*, \boldsymbol{\lambda}^*_{\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}^*_I)$, is

$$\mathcal{J}_{\boldsymbol{x}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}} A & 0 & -\eta_{\boldsymbol{x}} B^\top & 0 \\ \mathbb{I} & 0 & 0 & 0 \\ 0 & 0 & \mathbb{I} & 0 \\ 0 & 0 & 0 & \mathbb{I} \end{pmatrix}. \tag{111}$$

Multiplying the two blocks yields

$$\mathcal{J}_{\text{OG}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}} A - \eta_{\boldsymbol{x}}(\eta_{\text{dual}} + \omega) B^\top B & \eta_{\boldsymbol{x}} \omega\, B^\top B & -\eta_{\boldsymbol{x}} B^\top & 0 \\ \mathbb{I} & 0 & 0 & 0 \\ (\eta_{\text{dual}} + \omega) B & -\omega\, B & \mathbb{I} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{112}$$

□

**Lemma 5.** *The characteristic polynomial of the Jacobian $\mathcal{J}_{OG}$ at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ from Lemma 4 is*

$$\chi_{\mathcal{J}_{OG}}(\sigma) = (-\sigma)^{m - |\mathcal{A}_{\boldsymbol{x}}| + d} \det\left((1-\sigma)^2 \mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma) A - \eta_{\boldsymbol{x}}(\omega(1-\sigma) - \eta_{dual}\sigma) B^\top B\right). \tag{113}$$

*Proof.* The Jacobian $\mathcal{J}_{\text{OG}}$ is block lower-triangular with respect to the partition between the active subsystem $(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}})$ and the inactive subsystem $\boldsymbol{\lambda}_{t,I}$. Its characteristic polynomial is therefore the product of the characteristic polynomials of these two diagonal blocks.

**Inactive subsystem.** The bottom-right block corresponding to the dual variables of inactive inequality constraints is $\mathcal{J}_{\text{OG},I} = 0$. It's characteristic polynomial is

$$\chi_{\mathcal{J}_{\text{OG},I}}(\sigma) = \det(0 - \sigma\mathbb{I}) = \det(-\sigma\mathbb{I}_{|I_I|}) = (-\sigma)^{|I_I|}. \tag{114}$$

**Active Subsystem**

The Jacobian of the active subsystem, which maps $(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1,A})$, is

$$\mathcal{J}_{\text{OG},\mathcal{A}_{\boldsymbol{x}}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}} A - \eta_{\boldsymbol{x}}(\eta_{\text{dual}} + \omega)B^\top B & \eta_{\boldsymbol{x}}\omega\, B^\top B & -\eta_{\boldsymbol{x}} B^\top \\ \mathbb{I} & 0 & 0 \\ (\eta_{\text{dual}} + \omega)B & -\omega\, B & \mathbb{I} \end{pmatrix}. \tag{115}$$

We seek its characteristic polynomial, $\chi_{\mathcal{J}_{\text{OG},\mathcal{A}_{\boldsymbol{x}}}}(\sigma) = \det(\mathcal{J}_{\text{OG},\mathcal{A}_{\boldsymbol{x}}} - \sigma\mathbb{I})$. Let $M = \mathcal{J}_{\text{OG},\mathcal{A}_{\boldsymbol{x}}} - \sigma\mathbb{I}$, with blocks $M_{ij}$:

$$M = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}} A - \eta_{\boldsymbol{x}}(\eta_{\text{dual}} + \omega)B^\top B - \sigma\mathbb{I} & \eta_{\boldsymbol{x}}\omega\, B^\top B & -\eta_{\boldsymbol{x}} B^\top \\ \mathbb{I} & -\sigma\mathbb{I} & 0 \\ (\eta_{\text{dual}} + \omega)B & -\omega\, B & (1-\sigma)\mathbb{I} \end{pmatrix}. \tag{116}$$

To simplify the determinant calculation, we apply a determinant-preserving column operation, $C_2 \leftarrow C_2 + \sigma C_1$, which zeros out the $(2,2)$ block:

$$\det(M) = \det \begin{pmatrix} M_{11} & M_{12} + \sigma M_{11} & M_{13} \\ \mathbb{I} & -\sigma\mathbb{I} + \sigma\mathbb{I} & 0 \\ M_{31} & M_{32} + \sigma M_{31} & M_{33} \end{pmatrix} = \det \begin{pmatrix} M_{11} & M'_{12} & M_{13} \\ \mathbb{I} & 0 & 0 \\ M_{31} & M'_{32} & M_{33} \end{pmatrix}, \tag{117}$$

where $M'_{12} = M_{12} + \sigma M_{11}$ and $M'_{32} = M_{32} + \sigma M_{31}$. Swapping the first two block rows (corresponding to $\boldsymbol{x}_{t+1}$ and $\boldsymbol{x}_t$) multiplies the determinant by $(-1)^d$, yielding a block upper-triangular matrix:

$$\chi_{\mathcal{J}_{\text{OG},\mathcal{A}_{\boldsymbol{x}}}}(\sigma) = (-1)^d \det \begin{pmatrix} \mathbb{I} & 0 & 0 \\ M_{11} & M'_{12} & M_{13} \\ M_{31} & M'_{32} & M_{33} \end{pmatrix} = (-1)^d \det(\mathbb{I}) \det \begin{pmatrix} M'_{12} & M_{13} \\ M'_{32} & M_{33} \end{pmatrix}. \tag{118}$$

Since $M_{33} = (1-\sigma)\mathbb{I}$ is a multiple of the identity matrix, it commutes with all other blocks. We can thus compute the determinant of the remaining $2 \times 2$ block matrix as $\det(M'_{12}M_{33} - M_{13}M'_{32})$. Let's compute this product:

$$\begin{aligned}
&M'_{12}M_{33} - M_{13}M'_{32} \\
&= (M_{12} + \sigma M_{11})M_{33} - M_{13}(M_{32} + \sigma M_{31}) \\
&= (1-\sigma)(M_{12} + \sigma M_{11}) - M_{13}(M_{32} + \sigma M_{31}) \\
&= (1-\sigma)\big(\eta_{\boldsymbol{x}}\omega B^\top B + \sigma(\mathbb{I} - \eta_{\boldsymbol{x}} A - \eta_{\boldsymbol{x}}(\eta_{\text{dual}} + \omega)B^\top B - \sigma\mathbb{I})\big) \\
&\quad - (-\eta_{\boldsymbol{x}} B^\top)\big(-\omega B + \sigma(\eta_{\text{dual}} + \omega)B\big) \\
&= \sigma(1-\sigma)(1-\sigma)\mathbb{I} - \eta_{\boldsymbol{x}}\sigma(1-\sigma)A \\
&\quad + \big[(1-\sigma)\eta_{\boldsymbol{x}}\omega - \sigma(1-\sigma)\eta_{\boldsymbol{x}}(\eta_{\text{dual}} + \omega) - \eta_{\boldsymbol{x}}\omega + \sigma\eta_{\boldsymbol{x}}(\eta_{\text{dual}} + \omega)\big]B^\top B \\
&= \sigma(1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}\sigma(1-\sigma)A \\
&\quad + \eta_{\boldsymbol{x}}\big[\omega - \sigma\omega - \sigma\eta_{\text{dual}} - \sigma\omega + \sigma^2\eta_{\text{dual}} + \sigma^2\omega - \omega + \sigma\eta_{\text{dual}} + \sigma\omega\big]B^\top B \\
&= \sigma(1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}\sigma(1-\sigma)A + \eta_{\boldsymbol{x}}\big[\sigma^2\eta_{\text{dual}} - \sigma\omega(1-\sigma)\big]B^\top B \\
&= \sigma(1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}\sigma(1-\sigma)A - \eta_{\boldsymbol{x}}\sigma\big(\omega(1-\sigma) - \eta_{\text{dual}}\sigma\big)B^\top B.
\end{aligned}$$

The determinant of this $d \times d$ matrix is:

$$\det(M'_{12}M_{33} - M_{13}M'_{32}) = \sigma^d \det\left((1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma)A - \eta_{\boldsymbol{x}}\big(\omega(1-\sigma) - \eta_{\text{dual}}\sigma\big)B^\top B\right). \tag{119}$$

Substituting this back into the expression for $\chi_{\mathcal{J}_{\text{OG}},\mathcal{A}_{\boldsymbol{x}}}(\sigma)$:

$$\chi_{\mathcal{J}_{\text{OG}},\mathcal{A}_{\boldsymbol{x}}}(\sigma) = (-1)^d \cdot \sigma^d \det(\dots) \tag{120}$$

$$= (-\sigma)^d \det\left((1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma)A - \eta_{\boldsymbol{x}}\big(\omega(1-\sigma) - \eta_{\text{dual}}\sigma\big)B^\top B\right). \tag{121}$$

The characteristic polynomial of the full Jacobian $\mathcal{J}_{\text{OG}}$ is the product of the polynomials for the active and inactive subsystems, $\chi_{\mathcal{J}_{\text{OG}}}(\sigma) = \chi_{\mathcal{J}_{\text{OG},I}}(\sigma) \cdot \chi_{\mathcal{J}_{\text{OG}},\mathcal{A}_{\boldsymbol{x}}}(\sigma)$, which gives the final result. □

***Proof of Theorem** 2.* Consider the following algorithms for solving inequality-constrained problems:

① Eq. (AL-GDA): Primal-first alternating gradient descent–projected gradient ascent on the Augmented Lagrangian,

$$\begin{aligned}
\boldsymbol{x}_{t+1} &\leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}}\left[\nabla f(\boldsymbol{x}_t) + \left[\boldsymbol{\lambda}_t + c\,\boldsymbol{g}(\boldsymbol{x}_t)\right]_+^\top \nabla \boldsymbol{g}(\boldsymbol{x}_t)\right] \\
\boldsymbol{\lambda}_{t+1} &\leftarrow \left(1 - \frac{\eta_{\text{dual}}}{c}\right)\boldsymbol{\lambda}_t + \frac{\eta_{\text{dual}}}{c}\left[\boldsymbol{\lambda}_t + c\,\boldsymbol{g}(\boldsymbol{x}_{t+1})\right]_+,
\end{aligned} \tag{122}$$

where $0 < \eta_{\text{dual}} \le c$.

② Eq. (Lag-GD-OA): dual-first alternating gradient descent–projected optimistic ascent on the Lagrangian,

$$\begin{aligned}
\boldsymbol{\lambda}_{t+1} &\leftarrow \left[\boldsymbol{\lambda}_t + \eta_{\text{dual}}\,\boldsymbol{g}(\boldsymbol{x}_t) + \omega\big(\boldsymbol{g}(\boldsymbol{x}_t) - \boldsymbol{g}(\boldsymbol{x}_{t-1})\big)\right]_+ \\
\boldsymbol{x}_{t+1} &\leftarrow \boldsymbol{x}_t - \eta_{\boldsymbol{x}}\left[\nabla f(\boldsymbol{x}_t) + \boldsymbol{\lambda}_{t+1}^\top \nabla \boldsymbol{g}(\boldsymbol{x}_t)\right],
\end{aligned} \tag{123}$$

where $\omega > 0$.

We will analyze the local convergence properties of both algorithms at a stationary (KKT) point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$, satisfying Assumption 1, and operating on a partition of the $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_I)$ state into those corresponding to active and inactive constraints at $\boldsymbol{x}^*$:

$$\boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}} = \{\,\lambda_i \mid g_i(\boldsymbol{x}^*) = 0,\ \lambda_i^* > 0\,\},$$

and inactive ones

$$\boldsymbol{\lambda}_{t,I} = \{\,\lambda_i \mid g_i(\boldsymbol{x}^*) < 0,\ \lambda_i^* = 0\,\}.$$

Let the matrices $A$ and $B$ be defined at the stationary point $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*)$ as

$$A = \nabla_{\boldsymbol{x}}^2 \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) \quad \text{and} \quad B = \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*), \tag{124}$$

where $A$ is the **Hessian of the Lagrangian** with respect to $\boldsymbol{x}$ and $B$ is the **Jacobian of the active constraints**.

From Lemmas 2 and 3, we know that the Jacobian $\mathcal{J}_{\text{AL}}$ of the Augmented Lagrangian updates in Eq. (122), acting on the partitioned state $(\boldsymbol{x}_t, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I}) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_{t+1,A}, \boldsymbol{\lambda}_{t+1,I})$ and evaluated at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = (\boldsymbol{x}^*, \boldsymbol{\lambda}^*_{\mathcal{A}_{\boldsymbol{x}}}, \mathbf{0})$, is

$$\mathcal{J}_{\text{AL}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}}(A + c\,B^\top B) & -\eta_{\boldsymbol{x}} B^\top & 0 \\ \eta_{\text{dual}} B(\mathbb{I} - \eta_{\boldsymbol{x}}(A + c\,B^\top B)) & \mathbb{I} - \eta_{\boldsymbol{x}}\eta_{\text{dual}} BB^\top & 0 \\ 0 & 0 & (1 - \eta_{\text{dual}}/c)\mathbb{I} \end{pmatrix}; \tag{125}$$

and that the characteristic polynomial for this Jacobian is:

$$\chi_{\mathcal{J}_{\text{AL}}}(\sigma) = \left(1 - \frac{\eta_{\text{dual}}}{c} - \sigma\right)^{m - |\mathcal{A}_{\boldsymbol{x}}|} \det\left((1-\sigma)^2\mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma)A - \eta_{\boldsymbol{x}}(c(1-\sigma) - \eta_{\text{dual}}\sigma)B^\top B\right). \tag{126}$$

Moreover, from Lemmas 4 and 5, we know that for the same choices of $A$ and $B$ matrices, it follows that the Jacobian $\mathcal{J}_{\text{OG}}$ of the dual optimistic ascent updates in Eq. (123), acting on the partitioned state

$(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I}) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{x}_t, \boldsymbol{\lambda}_{t+1,A}, \boldsymbol{\lambda}_{t+1,I})$ and evaluated at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = (\boldsymbol{x}^*, \boldsymbol{x}^*, \boldsymbol{\lambda}^*_{\mathcal{A}_{\boldsymbol{x}}}, \mathbf{0})$, is

$$\mathcal{J}_{\text{OG}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}} A - \eta_{\boldsymbol{x}}(\eta_{\text{dual}} + \omega) B^\top B & \eta_{\boldsymbol{x}} \omega\, B^\top B & -\eta_{\boldsymbol{x}} B^\top & 0 \\ \mathbb{I} & 0 & 0 & 0 \\ (\eta_{\text{dual}} + \omega) B & -\omega\, B & \mathbb{I} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{127}$$

and that the characteristic polynomial for this Jacobian is:

$$\chi_{\mathcal{J}_{\text{OG}}}(\sigma) = (-\sigma)^{|I_I|+d} \det\left((1-\sigma)^2 \mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma) A - \eta_{\boldsymbol{x}}(\omega(1-\sigma) - \eta_{\text{dual}}\sigma) B^\top B\right), \tag{128}$$

**Remark on the $\sigma = 1$ case** A subtle but important point is that, although the structure of these characteristic polynomials might suggest $\sigma = 1$ could be a root if $\det(B^\top B) = 0$, it can be shown that $\sigma = 1$ is in fact not an eigenvalue of either $\mathcal{J}_{\text{AL}}$ or $\mathcal{J}_{\text{OG}}$ under Assumptions 1 to 3—conditions that are generally necessary, though not sufficient, for convergence.

We establish the equivalence of the local convergence properties of the two algorithms by setting the optimistic parameter $\omega = c$ and comparing the characteristic polynomials of their Jacobians.

Upon setting $\omega = c$ in $\chi_{\mathcal{J}_{\text{OG}}}(\sigma)$, the non-trivial determinant term becomes identical to that in $\chi_{\mathcal{J}_{\text{AL}}}(\sigma)$:

$$\det\left((1-\sigma)^2 \mathbb{I} - \eta_{\boldsymbol{x}}(1-\sigma) A - \eta_{\boldsymbol{x}}(c(1-\sigma) - \eta_{\text{dual}}\sigma) B^\top B\right). \tag{129}$$

The roots of this determinant define the set of non-trivial eigenvalues, which are shared between both methods. Local convergence of either algorithm depends on whether all these shared eigenvalues lie within the unit circle.

The remaining eigenvalues are given by the roots of the other factors in each polynomial:

- For the Augmented Lagrangian method, the factor $\left(1 - \frac{\eta_{\text{dual}}}{c} - \sigma\right)^{|I_I|}$ yields $|I_I|$ eigenvalues at $\sigma = 1 - \eta_{\text{dual}}/c$. Given the condition $0 < \eta_{\text{dual}} \leq c$, these eigenvalues lie in the interval $(0, 1)$, and are thus stable.
- For the optimistic ascent method, the factor $(-\sigma)^{|I_I|+d}$ yields $|I_I| + d$ eigenvalues at $\sigma = 0$, which are also stable.

**Conclusion**. Since the non-trivial eigenvalues governing convergence are identical for both algorithms, and the remaining trivial eigenvalues for both are stable, the spectral radius $\rho(\mathcal{J}_{\text{AL}}) < 1$ if and only if $\rho(\mathcal{J}_{\text{OG}}) < 1$. Therefore, with the choice of $\omega = c$, one algorithm converges locally if and only if the other one does.

Moreover, it holds that:

$$\rho(\mathcal{J}_{\text{AL}}) = \max\{\rho(\mathcal{J}_{\text{OG}}), 1 - \eta_{\text{dual}}/c\}. \tag{130}$$

This implies that if the shared eigenvalues dictate the convergence rate (i.e., when $1 > \rho(\mathcal{J}_{\text{OG}}) \geq 1 - \eta_{\text{dual}}/c$), both algorithms converge at the same local rate. □

### C.3 Further Proofs

This subsection includes the proofs for Prop. 3 (ALM finds all constrained minimizers) and Prop. 4 (ALM finds only constrained minimizers).

***Proof of Proposition 3.*** Let $\boldsymbol{\lambda}^* \succeq \mathbf{0}$ and $\boldsymbol{\mu}^*$ be the optimal Lagrange multipliers for $\boldsymbol{x}^*$ such that $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ form a KKT point of Eq. (CMP). Since $\boldsymbol{x}^*$ satisfies Assumption 3, these multipliers exist and are unique.

By Lemma 2, the Jacobian of algorithm Eq. (AL-GDA), acting on the partitioned state $(\boldsymbol{x}_t, \boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\lambda}_{t,I}) \mapsto (\boldsymbol{x}_{t+1}, \boldsymbol{\lambda}_{t+1,A}, \boldsymbol{\lambda}_{t+1,I})$ and evaluated at $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*) = (\boldsymbol{x}^*, \boldsymbol{\lambda}^*_{\mathcal{A}_{\boldsymbol{x}}}, \mathbf{0})$, is

$$\mathcal{J}_{\text{AL}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}}(A + c\, B^\top B) & -\eta_{\boldsymbol{x}} B^\top & 0 \\ \eta_{\text{dual}} B(\mathbb{I} - \eta_{\boldsymbol{x}}(A + c\, B^\top B)) & \mathbb{I} - \eta_{\boldsymbol{x}} \eta_{\text{dual}} B B^\top & 0 \\ 0 & 0 & (1 - \eta_{\text{dual}}/c)\mathbb{I} \end{pmatrix}, \tag{131}$$

where

$$A = \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*), \qquad B = \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*). \tag{132}$$

This Jacobian is well-defined thanks to Assumption 1.

We now proceed by grouping equality constraints with active inequality constraints, considering the Jacobian on the state partitioned as $(\boldsymbol{x}_t, (\boldsymbol{\lambda}_{t,\mathcal{A}_{\boldsymbol{x}}}, \boldsymbol{\mu}_t), \boldsymbol{\lambda}_{t,I})$. This yields the same Jacobian $\mathcal{J}_{\text{AL}}$, except $B$ is redefined as

$$A = \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*), \qquad B = \begin{pmatrix} \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*) \\ \nabla \boldsymbol{h}(\boldsymbol{x}^*) \end{pmatrix}. \tag{133}$$

We will now prove that under Assumptions 2 and 3, there exist $c > 0$ large enough and $\eta_{\boldsymbol{x}}, \eta_{\text{dual}} > 0$ small enough such that all eigenvalues of $\mathcal{J}_{\text{AL}}$ lie within the unit circle.

$\mathcal{J}_{\text{AL}}$ is block lower-triangular, so its eigenvalues are the union of the eigenvalues of its blocks. The eigenvalues of the inactive inequality block are $\sigma = 1 - \eta_{\text{dual}}/c$. Under the assumption of Eq. (AL-GDA) that $\eta_{\text{dual}} \leq c$, these satisfy $|\sigma| < 1$.

Setting $\eta = \eta_{\boldsymbol{x}} = \eta_{\text{dual}}$ for simplicity, the upper-left block is

$$\mathcal{J}_{\text{AL},\mathcal{A}_{\boldsymbol{x}}} = \begin{pmatrix} \mathbb{I} - \eta(A + c\, B^\top B) & -\eta B^\top \\ \eta B(\mathbb{I} - \eta(A + c\, B^\top B)) & \mathbb{I} - \eta^2 B B^\top \end{pmatrix}. \tag{134}$$

As in Lemma 3, $\mathcal{J}_{\text{AL},\mathcal{A}_{\boldsymbol{x}}}$ is similar to

$$\mathcal{J}'_{\text{AL},\mathcal{A}_{\boldsymbol{x}}} = \begin{pmatrix} \mathbb{I} - \eta A - \eta(c+\eta) B^\top B & -\eta B^\top \\ \eta B & \mathbb{I} \end{pmatrix} \tag{135}$$

$$= \mathbb{I} - \eta \begin{pmatrix} A + (c+\eta) B^\top B & B^\top \\ -B & 0 \end{pmatrix} \tag{136}$$

$$= \mathbb{I} - \eta\, M. \tag{137}$$

By Prop. 2, under Assumptions 2 and 3, there exists $c > 0$ such that $A + cB^\top B$ is positive definite. Hence $A + (c+\eta) B^\top B$ is also positive definite. By Bertsekas (2016, Prop. 5.4.2), $M$ has eigenvalues with positive real parts, so for sufficiently small $\eta > 0$, all eigenvalues of $\mathcal{J}'_{\text{AL},\mathcal{A}_{\boldsymbol{x}}}$ satisfy $|\sigma| < 1$.

Combining this with the inactive block, we conclude $\rho(\mathcal{J}_{\text{AL}}) < 1$, completing the proof. □

***Proof of Proposition 4.*** Let $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ be an LSSP of the algorithm in Eq. (AL-GDA). By Prop. 1, $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ is a KKT point of Eq. (CMP). Moreover, we know that $\rho(\mathcal{J}_{\text{AL}}) < 1$, where

$$\mathcal{J}_{\text{AL}} = \begin{pmatrix} \mathbb{I} - \eta_{\boldsymbol{x}}(A + c\, B^\top B) & -\eta_{\boldsymbol{x}} B^\top & 0 \\ \eta_{\text{dual}} B(\mathbb{I} - \eta_{\boldsymbol{x}}(A + c\, B^\top B)) & \mathbb{I} - \eta_{\boldsymbol{x}} \eta_{\text{dual}} B B^\top & 0 \\ 0 & 0 & (1 - \eta_{\text{dual}}/c)\mathbb{I} \end{pmatrix}, \tag{138}$$

and

$$A = \nabla^2_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*), \qquad B = \begin{pmatrix} \nabla \boldsymbol{g}_{\mathcal{A}_{\boldsymbol{x}}}(\boldsymbol{x}^*) \\ \nabla \boldsymbol{h}(\boldsymbol{x}^*) \end{pmatrix}, \tag{139}$$

as in the *Proof of Proposition 3*. The assumption $\rho(\mathcal{J}_{\text{AL}}) < 1$ implies that $\rho(\mathcal{J}'_{\text{AL},\mathcal{A}_{\boldsymbol{x}}}) < 1$, where

$$\mathcal{J}'_{\text{AL},\mathcal{A}_{\boldsymbol{x}}} = \begin{pmatrix} \mathbb{I} - \eta A - \eta(c+\eta) B^\top B & -\eta B^\top \\ \eta B & \mathbb{I} \end{pmatrix} \tag{140}$$

is similar to the upper-left block of $\mathcal{J}_{\text{AL}}$.

We proceed by contradiction. Suppose that $\boldsymbol{x}^*$ is not a local constrained minimizer of Eq. (CMP). Then $(\boldsymbol{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)$ must violate the second-order sufficiency conditions of Assumption 2 (Bertsekas, 2016, Prop. 4.3.2).

This implies the existence of a nonzero vector $\boldsymbol{d} \in \mathbb{R}^d$ in the null space of $B$ such that

$$\boldsymbol{d}^\top A \boldsymbol{d} \leq 0. \tag{141}$$

By the Rayleigh-Ritz theorem, choose $\boldsymbol{d}$ as a normalized eigenvector ($\|\boldsymbol{d}\| = 1$) corresponding to the smallest eigenvalue, $\alpha$, of $A$ restricted to the subspace $\{\boldsymbol{d} \mid B\boldsymbol{d} = 0\}$. By the assumption, $\alpha \leq 0$.

Consider the action of $\mathcal{J}'_{\text{AL},\mathcal{A}_\infty}$ on the vector $[\boldsymbol{d}^\top, 0]^\top$:

$$\mathcal{J}'_{\text{AL},\mathcal{A}_\infty} \begin{pmatrix} \boldsymbol{d} \\ 0 \end{pmatrix} = \begin{pmatrix} (\mathbb{I} - \eta A - \eta(c+\eta)B^\top B)\boldsymbol{d} \\ \eta B \boldsymbol{d} \end{pmatrix} = \begin{pmatrix} \boldsymbol{d} - \eta A \boldsymbol{d} \\ 0 \end{pmatrix}, \tag{142}$$

using $Bd = 0$.

Then

$$\left\| \mathcal{J}'_{\text{AL},\mathcal{A}_\infty} \begin{pmatrix} \boldsymbol{d} \\ 0 \end{pmatrix} \right\|^2 = \|\boldsymbol{d} - \eta A \boldsymbol{d}\|^2 = \|\boldsymbol{d}\|^2 - 2\eta(\boldsymbol{d}^\top A \boldsymbol{d}) + \eta^2 \|A\boldsymbol{d}\|^2 = 1 - 2\eta\alpha + \eta^2 \|A\boldsymbol{d}\|^2 \geq 1, \tag{143}$$

since $\alpha \leq 0$ and $\|\boldsymbol{d}\|^2 = 1$.

Thus, we have found a vector $[\boldsymbol{d}^\top, 0]^\top$ with unit norm for which $\mathcal{J}'_{\text{AL},\mathcal{A}_\infty}$ is non-contractive, contradicting the assumption $\rho(\mathcal{J}_{\text{AL}}) < 1$.

Therefore, our initial assumption that Assumption 2 is violated must be false.

□