A Projection-Free Homeomorphic Lagrangian Method for Optimization with Convex Inequality and General Equality Constraints

Anonymous authors
Paper under double-blind review

Abstract

Constrained optimization with convex inequality and general equality constraints arises throughout machine learning, power systems, and scientific computing. Classical first-order methods require either expensive projections onto the feasible region or optimization oracles, whereas recent projection-free reparameterization methods focus on problems without equality constraints. We propose **Hom-PALM**, a Homeomorphic Proximal Augmented Lagrangian Method that bridges this gap by reparameterizing inequality constraints onto the unit ball via a homeomorphism while handling equality constraints through an augmented Lagrangian framework in the original domain. This decomposition enables efficient ball-constrained subproblem solving while preserving the structure required for Lagrangian methods. A key challenge is that a naive combination of reparameterization with the Lagrangian framework yields suboptimal rates: each Lagrangian iteration requires solving a subproblem using an accelerated-based method; yet the homeomorphic mapping renders subproblems nonconvex with only hidden convexity, precluding standard accelerated rates. As our main technical contribution, we extend first-order acceleration to the projection-free setting without optimization oracles and integrate it within Hom-PALM, achieving **optimal** convergence rates for convex problems and best-known rates for nonconvex problems. For instance, Hom-PALM attains a rate of $\widetilde{\mathcal{O}}(\epsilon^{-3})$ for nonconvex objectives with convex inequality and *nonlinear* equality constraints, matching the best-known complexity but with per-iteration cost of only $\widetilde{\mathcal{O}}(n^2)$ without optimization oracles. We validate Hom-PALM on convex QCQP and Stiefel manifold optimization, demonstrating competitive performance.

1 Introduction

Constrained optimization problems with convex inequality constraints and nonlinear equality constraints arise in various applications such as optimal power flow in power system operation, semidefinite relaxations, robust optimization, and adversarial learning. Inequality constraints naturally encode resource budgets, safety margins, and regularization bounds, while nonlinear equalities arise from physical conservation laws, manifold geometry (e.g., orthogonality), and coupling conditions in multi-component systems. The simultaneous presence of both constraint types characterizes many problems of practical importance but poses substantial algorithmic challenges.

Classical projection-based gradient methods (Beck, 2017; Zhang et al., 2022; Zhang & Luo, 2022) offer fast convergence rates but require solving a quadratic program at each iteration, which is computationally expensive unless the constraint set admits simple closed-form projections (e.g., Euclidean balls or boxes). Frank-Wolfe methods (Frank et al., 1956; Mu et al., 2016; Tajima et al., 2021; Mhammedi, 2022; Maskan et al., 2025) replace projections with a linear optimization oracle, yet this oracle can still be demanding for complex feasible sets, and these methods tend to exhibit oscillatory behavior and slow convergence near

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

Table 1: Summary of reparameterization methods and ALM for solving constrained optimization.

Reference	Settings: Obj.	Ctr.	Key Assumption	Algorithm	Subproblem Inner Solver[3]	Per-iteration Complexity	Convergence Rate[4]
(Liu et al., 2025)	C	C ineq.	ND Minimizer	Homeomorphism + Regular PGD	-	MO	$\mathcal{O}(\epsilon^{-1})$
	SC	w/o eq.	-				$\mathcal{O}(\lvert\log\epsilon\rvert)$
	NC		-				$\mathcal{O}(\epsilon^{-2})$
(Li & Qu, 2021)	C	C ineq. + L eq.	Strong duality	Proximal ALM	Accelerated-based	PO	$\tilde{\mathcal{O}}(\epsilon^{-1})$
(Kong et al., 2023b)	NC	C ineq. + C eq.	Slater's condition		ACG		$\tilde{\mathcal{O}}(\epsilon^{-3})$
(Melo et al., 2024)		C ineq. + L eq.					$\tilde{\mathcal{O}}(\epsilon^{-2.5})$
Thm. 1	SC	C ineq.	-	**Hom-PALM**: Homeomorphism + Proximal ALM	Hom-APG over ball (Sec. 4.2.1)	MO	$\tilde{\mathcal{O}}(\epsilon^{-0.5})$
	C	+ L eq.	-				$\tilde{\mathcal{O}}(\epsilon^{-1})$
Thm. 2	NC	C ineq. + L eq.	Slater's condition + WC obj.		APG over ball		$\tilde{\mathcal{O}}(\epsilon^{-2.5})$
		C ineq. + NL eq.	KL regularity + WC eq./obj.				$\mathcal{O}(\epsilon^{-3})$

[1] **Basis Abbreviations**: C = convex, NC = non-convex, DC = difference of convex, SC = strongly convex, (N)L = (non-)linear, WC = weakly convex, Obj = objective, Ctr = constraint, ND = non-degenerate, PO = projection oracle, MO = membership oracle.
[2] **Algorithm Abbreviations**: ALM = augmented Lagrangian methods, FW = Frank-Wolfe, PGD = projected gradient descent, APG = accelerated proximal gradient, ACG = accelerated composite gradient, AdapAPG = adaptive APG.
[3] **Inner solvers**: AdapAPG, ACG, and Hom-APG, despite different names and formulations, essentially share the same algorithmic steps, including a proximal-gradient step and an acceleration step.
[4] **Convergence Rate**: iterations for finding an ϵ-optimum for convex problems or an ϵ-stationary point for non-convex problems.

optimal solutions (Bomze et al., 2024; Freund & Grigas, 2016). Penalty-based approaches (Bertsekas, 1976; Smith et al., 1997; Lin et al., 2022) avoid explicit feasibility enforcement but often become ill-conditioned as penalty parameters increase. More recent projection-free formulations based on reparameterization (Li et al., 2023; Tang & Toh, 2024; Chok & Vasil, 2025; Liu et al., 2025) and radial-dual reformulations (Grimmer, 2024a;b) show promising theoretical properties, though they are typically limited to structured inequality constraints and may require impractical step-size schedules.

The presence of nonlinear equality constraints introduces additional non-convexity, rendering many of the above methods ineffective. The Augmented Lagrangian Method (ALM) (Powell, 1969; Hestenes, 1969) provides a principled approach by incorporating equality constraints into the objective via dual variables and penalty terms, converting the problem into a sequence of subproblems with only inequality constraints. However, solving each subproblem typically requires projections or proximal steps onto the inequality-constrained set, resulting in high per-iteration costs despite ALM's theoretical appeal.

We refer readers to Appendix A for a detailed discussion of related work on per-iteration cost and convergence comparison for different approaches. To date, a fundamental question remains: *Can we design a scalable optimization method that handles both equality and inequality constraints while achieving fast convergence and low per-iteration complexity?*

In this paper, we answer this question by proposing **Hom-PALM**, a first-order projection-free framework for constrained optimization with both equality and inequality constraints. Hom-PALM combines homeomorphic reparameterization with the proximal ALM, mapping convex inequality constraints to the unit ball and introducing dual variables for the equality constraints. This transformation enables solving primal non-convex subproblems over the unit ball, avoiding expensive projections onto the original constraint set, followed by dual variable updates.

$\triangleright$ For originally convex problems, we establish solution set equivalence, global optimality, and strong duality for the reformulation. We prove *optimal* convergence rates of $\tilde{\mathcal{O}}(\epsilon^{-1/2})$ for strongly convex objectives and $\tilde{\mathcal{O}}(\epsilon^{-1})$ for general convex objectives, matching the complexity lower bound (Ouyang & Xu, 2021) for convex-concave bilinear saddle point problems in Lagrangian formulation. The central algorithmic difficulty is that the homeomorphic transformation makes the primal subproblems nonconvex in the transformed coordinates, even when the original problem is convex. We address this difficulty with **Hom-APG**, which forms regularized models and extrapolates in the original coordinates while solving the

transformed models over the ball. Section 3.4 introduces this mechanism, and Section 4.2.1 analyzes its role in the total complexity of Hom-PALM.

$\triangleright$ For originally non-convex problems, we derive convergence rates of $\tilde{\mathcal{O}}(\epsilon^{-2.5})$ for nonconvex objectives with linear equality constraints and $\tilde{\mathcal{O}}(\epsilon^{-3})$ for nonlinear equality constraints, matching the state-of-the-art results in Table 1. All results feature $\tilde{\mathcal{O}}(n^2)$ per-iteration cost for the projection-free inner solver.

We evaluate the framework on convex QCQP and nonlinear Stiefel Manifold optimization, and compare convergence and solution quality against projection-based, penalty-based, and augmented-Lagrangian baselines, showing competitive convergence speed and solution quality.

2 Problem Statement and Homeomorphic Reparameterization

We consider the following constrained optimization problem:

$$\min_{\mathbf{x}} \ f(\mathbf{x}), \quad \text{s.t.} \ \ \mathbf{x} \in \mathcal{C}, \ \boldsymbol{\varphi}(\mathbf{x}) = 0. \tag{\textbf{P}}$$

Here $\mathbf{x} \in \mathbb{R}^n$ is the decision variable, f is the objective function, $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) \leq \mathbf{0}\}$ is a compact full-dimensional convex set defined by convex inequality constraints $\mathbf{g} = (g_1, \ldots, g_m)$, and $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p) : \mathbb{R}^n \to \mathbb{R}^p$ collects the equality constraints. We denote the feasible set by $\mathcal{K} := \{\mathbf{x} \in \mathcal{C} \mid \boldsymbol{\varphi}(\mathbf{x}) = 0\}$. Throughout the paper, $\mathbf{x}$ denotes variables in the original space for (**P**).

For problem (**P**), state-of-the-art first-order methods based on (proximal) augmented Lagrangian frameworks (Li & Qu, 2021; Melo et al., 2024; Zhang & Luo, 2022; Pu et al., 2024) have established sharp iteration complexities in both convex and nonconvex regimes. Recent advances further improve inner-loop efficiency through linearization, acceleration, and stochastic or manifold extensions (Deng et al., 2025). Nevertheless, within this ALM line, (Zhang & Luo, 2022) attains an $\mathcal{O}(\epsilon^{-2})$ rate for nonconvex problems with linear inequality constraints, but each iteration requires projecting onto the inequality-feasible set (typically a QP); (Pu et al., 2024) extends to nonlinear inequalities at the same rate, yet equality constraints are not handled directly and incorporating them via primal feasibility or nested ALM/penalty loops can undermine the desired complexity. These observations suggest that, even at the current frontier, two bottlenecks persist for (**P**): **(i)** the cost of projections or proximal steps onto the general convex inequality set $\mathcal{C}$, and **(ii)** the difficulty of treating general, potentially nonlinear, equality constraints without expensive elimination or multi-loop schemes. We refer to Appendix A for further reading and a detailed comparison.

2.1 Reparameterization and Hom-PGD Background

To mitigate the projection bottleneck (i), a recent line of *reparameterization* methods replaces projections onto $\mathcal{C}$ by optimization over a simpler reference domain (Li et al., 2023; Tang & Toh, 2024; Poon & Peyré, 2023; Grimmer, 2024a;b). These approaches map a structured feasible region to a simpler geometry (e.g., a simplex, polyhedron, or ball) while preserving optimal solutions. However, their applicability is still limited: existing constructions typically require special structure on $\mathcal{C}$ and do not address the simultaneous presence of general equality constraints in (**P**).

Hom-PGD (Liu et al., 2025). The most closely related work is Hom-PGD (Liu et al., 2025), which maps a compact convex inequality set to a unit ball and achieves favorable rates for the inequality-only problem. Concretly, it focuses on optimization over a compact convex set described only by inequality constraints.It then constructs a homeomorphism Ψ that transforms ($\mathbf{P}_s$) into a ball-constrained problem ($\mathbf{H}_s$).

$$\min f(\mathbf{y}), \ \text{s.t.} \ \mathbf{y} \in \mathcal{C} := \{\mathbf{g}(\mathbf{y}) \leq 0\} \qquad (\mathbf{P}_s) \qquad\qquad \min h(\mathbf{z}), \ \text{s.t.} \ \mathbf{z} \in \mathcal{B} := \{\|\mathbf{z}\|_2 \leq 1\} \qquad (\mathbf{H}_s)$$

Here $h(\mathbf{z}) = f(\Psi(\mathbf{z}))$ and $\mathcal{C} = \Psi(\mathcal{B})$. The transformed problem preserves feasibility and optimal solutions, while replacing projections onto $\mathcal{C}$ by closed-form projections onto $\mathcal{B}$.

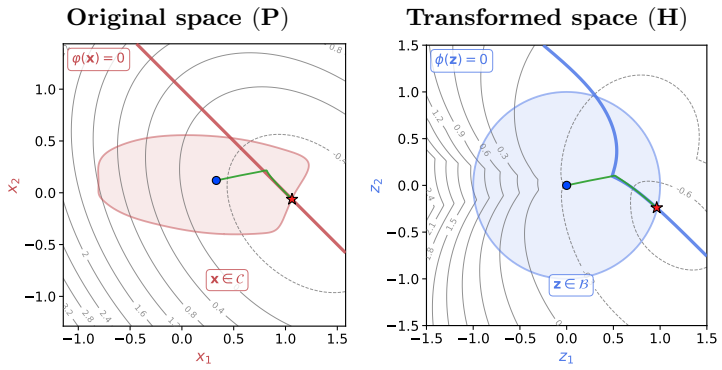


Figure 1: Optimization problem in original and transformed spaces via a homeomorphism (Def. 3.2). The convex inequality-defined set is mapped to a unit ball through $\mathcal{C} = \Psi(\mathcal{B})$. After the transformation, the objective and equality constraints generally become *nonconvex or nonlinear*, but the problem structure, such as landscape, feasible sets, and optimal solutions, remains in bijection.

Under convexity of $(\mathbf{P}_s)$, regular PGD on $(\mathbf{H}_s)$ (Alg. 4) enjoys favorable landscape properties and convergence speeds.

However, Hom-PGD does not treat general equalities in $(\mathbf{P})$. When equalities are present, prior reparameterization pipelines typically eliminate them beforehand—for example, via null-space or QR factorization for affine constraints—which incurs a one-time setup cost of $\mathcal{O}(n^3)$ in the dense unstructured case and still leaves nonlinear equalities outside the scope of the inequality-only reformulation. For problem $(\mathbf{P})$, the change of variables therefore leaves the additional condition $\boldsymbol{\varphi}(\Psi(\mathbf{z})) = 0$. Hom-PALM handles this condition through augmented-Lagrangian updates, as described in Section 3.

3 Homeomorphic Proximal Augmented Lagrangian Methods

Hom-PALM combines a change of variables for the convex inequality set with augmented-Lagrangian updates for the equalities. We first define the reformulation and its gauge construction, then describe the outer updates and the accelerated inner solver for the convex setting.

3.1 Homeomorphic Reparameterization

Definition 3.1 (Homeomorphic constrained optimization)**.** Given a homeomorphism $\Psi : \mathcal{B} \to \mathcal{C}$, we define the transformed optimization problem of $(\mathbf{P})$ as:

$$\min_{\mathbf{z}} \; h(\mathbf{z}), \quad \text{s.t.} \;\; \mathbf{z} \in \mathcal{B}, \;\; \boldsymbol{\phi}(\mathbf{z}) = 0 \tag{$\mathbf{H}$}$$

with objective function $h(\mathbf{z}) = f(\Psi(\mathbf{z}))$, equality constraint function $\boldsymbol{\phi}(\mathbf{z}) = \boldsymbol{\varphi}(\Psi(\mathbf{z}))$, and inequality constraint set $\mathcal{B} = \Psi^{-1}(\mathcal{C})$. Geometric illustrations of $\mathbf{P}$ and $(\mathbf{H})$ are in Fig. 1.

Remark. Although Ψ generally makes h and $\boldsymbol{\phi}$ nonconvex and nonlinear even when $(\mathbf{P})$ is convex, the reformulation replaces $\mathcal{C}$ with the unit ball, avoiding expensive projections while preserving a one-to-one correspondence between optimal solutions of $(\mathbf{P})$ and $(\mathbf{H})$.

3.2 Homeomorphism Construction: Gauge Mapping

We adopt the established gauge mapping as a closed-form homeomorphism for convex sets.

Definition 3.2 (Gauge mapping (Tabas & Zhang, 2022a;b; Li et al., 2025; Liu et al., 2025))**.** Let $\gamma_{\mathcal{C}}(\mathbf{x}, \mathbf{x}^\circ) = \inf\{\lambda \geq 0 \mid \mathbf{x} \in \lambda(\mathcal{C} - \mathbf{x}^\circ)\}$ be the Gauge/Minkowski function (Blanchini & Miani, 2008) given an interior point $\mathbf{x}^\circ \in \operatorname{int}(\mathcal{C})$. The gauge mapping $\Psi : \mathcal{B} \to \mathcal{C}$ is defined between a unit ball and a compact convex set:

$$\Psi(\mathbf{z}) = \frac{\|\mathbf{z}\|}{\gamma_{\mathcal{C}}(\mathbf{z}, \mathbf{x}^\circ)}\mathbf{z} + \mathbf{x}^\circ, \;\; \forall \mathbf{z} \in \mathcal{B} \setminus \{\mathbf{0}\}; \quad \Psi^{-1}(\mathbf{x}) = \frac{\gamma_{\mathcal{C}}(\mathbf{x} - \mathbf{x}^\circ, \mathbf{x}^\circ)}{\|\mathbf{x} - \mathbf{x}^\circ\|}(\mathbf{x} - \mathbf{x}^\circ), \;\; \forall \mathbf{x} \in \mathcal{C} \setminus \{\mathbf{x}^\circ\}.$$

Algorithm 1 Hom-PALM(θ)

Input: Initial primal and dual variables $\mathbf{z}_1 \in \mathcal{B}, \boldsymbol{\gamma}_1 \in \mathbb{R}^p$, coefficients $\rho_k, \beta_k > 0$.
for $k = 1$ **to** K **do**
  Primal: $\mathbf{z}_{k+1} \approx \arg\min_{\mathbf{z}\in\mathcal{B}} \mathbb{H}^{\theta}_{\rho_k,\beta_k}(\mathbf{z}; \boldsymbol{\gamma}_k)$
  Dual: $\boldsymbol{\gamma}_{k+1} = (1-\theta)\boldsymbol{\gamma}_k + \rho_k \boldsymbol{\phi}(\mathbf{z}_{k+1})$
end for
Output: $\mathbf{x}_{K+1} = \Psi(\mathbf{z}_{K+1})$

Algorithm 2 Hom-APG (concise version)

Input: Initial point $\mathbf{z}_0 \in \mathcal{B}$, $\mathbf{y}_0 \leftarrow \Psi(\mathbf{z}_0)$, parameters $\lambda_{\text{reg}} > 0$, q,ι_k
for $k = 1, 2, \ldots, K$ **do**
  $\widetilde{Q}_k(\mathbf{z}) := h(\mathbf{z}) + \frac{\lambda_{\text{reg}}}{2}\|\Psi(\mathbf{z}) - \mathbf{y}_{k-1}\|^2$
  Solve $\mathbf{z}_k \approx \min_{\mathbf{z}\in\mathcal{B}} \widetilde{Q}_k(\mathbf{z})$ via PGD
  $\mathbf{y}_k \leftarrow \Psi(\mathbf{z}_k) + \iota_k(\Psi(\mathbf{z}_k) - \Psi(\mathbf{z}_{k-1}))$
end for
Output: $\mathbf{z}_K$

We define the center points explicitly by $\Psi(\mathbf{0}) = \mathbf{x}^\circ$ and $\Psi^{-1}(\mathbf{x}^\circ) = \mathbf{0}$. The gauge mapping admits closed-form expressions for common convex sets and can be computed by bisection for general sets. Appendix B.1 gives its bi-Lipschitz bounds and computation details. The convergence analysis additionally requires a differentiable map with a nonsingular, Lipschitz-continuous Jacobian. These requirements are stronger than the almost-everywhere differentiability stated for the gauge construction; applying the theory to a smoothed implementation requires establishing those properties and accounting for the approximation error. We state the mapping assumptions explicitly in Section 4.

3.3 Hom-PALM Outer Updates

Building on the reformulation with gauge mapping, we apply a proximal augmented Lagrangian method to problem (**H**). The proximal term stabilizes the transformed non-convex subproblem, while the augmented-Lagrangian term handles the equality constraints without eliminating them. To align with recent PALM variants (Melo et al., 2024), we present the parameterized family PALM(θ) for $\theta \in [0, 1]$ associated with a generalized proximal term. Specifically, given a pair of primal and dual iterate $\mathbf{z}_k, \boldsymbol{\gamma}_k$, PALM($\theta$) updates the primal and dual variables as follows:

$$\textbf{Primal: } \mathbf{z}_{k+1} \approx \arg\min_{\mathbf{z}\in\mathcal{B}} \mathbb{H}^{\theta}_{\rho_k,\beta_k}(\mathbf{z}; \boldsymbol{\gamma}_k), \quad \textbf{Dual: } \boldsymbol{\gamma}_{k+1} = (1-\theta)\boldsymbol{\gamma}_k + \rho_k \boldsymbol{\phi}(\mathbf{z}_{k+1}). \tag{1}$$

We define the proximal augmented Lagrangian function $\mathbb{H}^{\theta}_{\rho,\beta}(\mathbf{z}; \boldsymbol{\gamma}) := \hat{h}^{\theta}_{\rho}(\mathbf{z}; \boldsymbol{\gamma}) + \beta\mathcal{P}(\mathbf{z}, \mathbf{z}_k)$, where $\hat{h}^{\theta}_{\rho}(\mathbf{z}; \boldsymbol{\gamma}) := h(\mathbf{z}) + (1-\theta)\langle \boldsymbol{\gamma}, \boldsymbol{\phi}(\mathbf{z})\rangle + \frac{\rho}{2}\|\boldsymbol{\phi}(\mathbf{z})\|^2$ is the θ-augmented Lagrangian with penalty parameter $\rho > 0$, and $\mathcal{P}(\cdot, \mathbf{z}_k) : \mathbb{R}^n \to \mathbb{R}$ is a generalized proximal term weighted by $\beta > 0$. The proximal function $\mathcal{P}$ admits several choices in this work: the standard quadratic penalty $\frac{1}{2}\|\mathbf{z} - \mathbf{z}_k\|^2$, or a homeomorphic proximal term $\frac{1}{2}\|\Psi(\mathbf{z}) - \Psi(\mathbf{z}_k)\|^2$, each inducing a geometry tailored to the problem structure. We refer to the integration of the homeomorphic transformation with this framework as **Hom-PALM**(θ), outlined in Algorithm 1. The parameter $\theta \in [0, 1]$ interpolates between two algorithmic paradigms: $\theta = 0$ recovers classical proximal ALM methods (Li & Qu, 2021; Melo et al., 2024; Ding et al., 2023; Hermans et al., 2022), while $\theta = 1$ yields a pure penalty method and its proximal variant (Lin et al., 2022). For notational convenience, we write Hom-PALM to denote the special case $\theta = 0$ unless otherwise stated.

Each primal subproblem is solved inexactly over $\mathcal{B}$, whose projection has a closed form. Mapping a ball iterate through Ψ enforces membership in $\mathcal{C}$; the equality residual is handled by the outer updates. The inner solver depends on the problem setting: Hom-APG is used for the convex analysis, while the nonconvex extensions use the APG-based routines described in Section 4.2.2.

3.4 Hom-APG: Accelerated Inner Solves

In the convex setting, composing the homeomorphic proximal subproblem with Ψ generally destroys convexity in $\mathbf{z}$, so standard acceleration cannot be applied directly. Hom-APG (Alg. 2) regularizes and extrapolates in the original coordinates while solving each model over the ball via Hom-PGD; parameter schedules and convergence guarantees are given in Appendix E.1 and Section 4.2.1.

4 Theoretical Guarantees and Complexity

We first state the mapping assumptions and the structural properties of the reformulation. We then analyze the Hom-APG inner solver and its contribution to the total inner-iteration complexity of Hom-PALM in the convex setting, followed by the nonconvex extensions and the computational cost.

General Assumptions (with details in Appendix C.2): The objective f and constrained functions g_i, ϕ_j $(i \in [m], j \in [p])$ in $(\mathbf{P})$ are continuously differentiable and smooth. The homeomorphic mapping Ψ is invertible with a non-singular Jacobian matrix and is (κ_1, κ_2)-bi-Lipschitz continuous. Additionally, the Jacobian matrix of Ψ (denoted as J_Ψ) exists and is Lipschitz continuous. Unless otherwise specified, the results below are conditional on a homeomorphism satisfying these assumptions. The gauge construction and its regularity boundary are discussed in Section 3 and Appendix B.1; almost-everywhere differentiability alone does not supply the Jacobian regularity assumed here.

4.1 Landscape Analysis

Although $(\mathbf{H})$ is nonconvex in $\mathbf{z}$, it inherits favorable landscape properties from convex $(\mathbf{P})$ through the homeomorphism Ψ—a phenomenon we term *hidden convexity*. When $(\mathbf{P})$ is convex with affine equalities, Slater's condition (Def. D.8), and a nonempty solution set, Appendix D establishes the following: **(i) Strong duality:** the augmented-Lagrangian dual of $(\mathbf{H})$ satisfies $\max_{\boldsymbol{\gamma}} \min_{\mathbf{z}\in\mathcal{B}} \hat{h}_\rho(\mathbf{z};\boldsymbol{\gamma}) = \min_{\mathbf{z}\in\mathcal{B}} \max_{\boldsymbol{\gamma}} \hat{h}_\rho(\mathbf{z};\boldsymbol{\gamma}) = h^* = f^*$. **(ii) KKT equivalence:** under strict complementary slackness (Def. D.6), $\mathbf{x}^*$ is a KKT point of $(\mathbf{P})$ if and only if $\mathbf{z}^* = \Psi^{-1}(\mathbf{x}^*)$ is a KKT point of $(\mathbf{H})$. **(iii) Global optimality:** every KKT stationary point of $(\mathbf{H})$ is a global optimum. Together, these results justify Hom-PALM: strong duality ensures dual ascent recovers the primal optimum, KKT equivalence preserves optimality under Ψ for mapping solutions from $(\mathbf{H})$ back to $(\mathbf{P})$, and global optimality guarantees that any KKT point found is optimal when $(\mathbf{P})$ is convex. The KKT correspondences also extend to nonconvex settings; see Appendix D.

4.2 Convergence Analysis with Inexact Inner Solvers

In this section, we present a theoretical convergence analysis of the Hom-PALM(θ) (Alg. 1) for solving problem $(\mathbf{P})$. Our main result shows that it achieves a convergence rate comparable to existing results in the convex and non-convex setting while retaining a projection-free structure for solving each subproblem.

4.2.1 Convex Setting

Consider the convex setting of problem $(\mathbf{P})$, with a convex objective and affine equalities. The analysis has two steps: bound the work required by the Hom-APG solver introduced in Section 3.4, and propagate its inexact-solve guarantees through the outer Hom-PALM iterations.

Inner-solver guarantee. We begin with the accelerated convergence statement for Hom-APG; the full statement and proof are given in Appendix E.1.

Proposition 4.1 (Accelerated convergence of Hom-APG, informal)**.** *Consider problem* $(\mathbf{P})$ *with μ_f-strongly convex and L_f-smooth objective, and the reformulated problem* $(\mathbf{H})$*. With appropriate parameters, Hom-APG (Alg. 2) achieves an accelerated linear convergence rate of $\mathcal{O}\left(\exp\left(-c\sqrt{\mu_f/L_f}\, s\right)\right)$, where c is a problem-dependent constant and s denotes the total number of inner iterations, i.e, the number of iterations to PGD.*

Proposition 4.1 improves upon the regular rate of $\mathcal{O}\left(\exp\left(-c(\mu_f/L_f)\, s\right)\right)$ for strongly convex problems achieved by the projection-free method Hom-PGD (Alg. 4) (Liu et al., 2025), closing the gap between homeomorphic methods and standard accelerated algorithms. Moreover, the resulting rate is comparable to APG methods (Beck & Teboulle, 2009; Tseng, 2008; Lin et al., 2018), but avoids the potentially expensive projection step at each iteration. See Appendix E.1 for complete proofs.

Overall complexity. Applying the inner-solver guarantee to the sequence of proximal augmented-Lagrangian subproblems yields the following bound on the total inner work. The outer parameter schedules and the required inner accuracies are specified in Appendix E.2.

Theorem 1. *Suppose strong duality holds for problem* (**P**). *Consider Hom-PALM with* $\theta = 0$ *and* $\mathcal{P}(\mathbf{z}, \mathbf{z}_k) = \frac{1}{2}\|\Psi(\mathbf{z}) - \Psi(\mathbf{z}_k)\|^2$, *implemented as Alg. 1 with appropriate parameters* β_k, ρ_k *defined in Appendix E.2. Then Hom-PALM returns an expected* ϵ*-optimal solution (Def. E.5) of* (**P**) *with total inner-iteration complexity* $\widetilde{\mathcal{O}}(1/\sqrt{\epsilon})$ *when* f *is strongly convex, and* $\widetilde{\mathcal{O}}(1/\epsilon)$ *when* f *is convex.*

Remarks. (a) The overall iteration complexity aligns with prior (proximal) ALM analyses for convex problems (Liu et al., 2019; Li & Qu, 2021; Xu, 2017; Patrascu et al., 2017), albeit under different assumptions. Notably, the convergence rate of Hom-PALM is *optimal* in the (strongly) convex setting, as it matches the complexity lower bound established in (Ouyang & Xu, 2021). **(b)** Our Hom-PALM methods are projection-free, as all iterations are performed over the unit ball. This yields significant computational efficiency compared to prior methods that require explicit projection or proximal operations over the original inequality constraint set. **(c)** The proof leverages the bijective correspondence between iteration trajectories in the primal space and the homeomorphic space. By exploiting the convexity of the original problem, we achieve the same convergence rate without expensive projections.

Proof novelty. The novelty of our proof techniques lies in resolving an obstacle that a direct combination with Li & Qu (2021) cannot handle. The PALM analysis therein assumes each inner subproblem is convex; after composing with Ψ, the subproblem $\min_{\mathbf{z}\in\mathcal{B}} L_\rho(\Psi(\mathbf{z}), \lambda)$ is nonconvex in $\mathbf{z}$ and possesses only *hidden* convexity, so standard APG and the inner-oracle framework no longer apply. Standard APG also requires a closed-form proximal step on the original feasible set, which is typically unavailable for general convex inequalities. We extend first-order acceleration to this hidden-convex, projection-free setting via Hom-APG (Sec. 3.4), which regularizes and extrapolates in the original coordinates while solving each inner model over the ball. Integrating Hom-APG into Hom-PALM recovers the optimal outer complexity without projections onto $\mathcal{C}$.

4.2.2 Non-Convex Setting

In this section, we consider problem (**P**) with a non-convex objective, convex inequality constraints, and (non-)linear equality constraints. Leveraging the geometry-preserving properties of the homeomorphic transformation, the convergence analysis of Hom-PALM(θ) follows existing results (Melo et al., 2024; Lin et al., 2022). Below, we state the convergence rates in a concise form for both linear and nonlinear equality constraints. Formal statements, including definitions of parameters, approximate stationary points, and assumptions, along with complete proofs, are provided in Appendix E.3 and E.4, respectively.

Theorem 2 (informal)**.** *Under appropriate regularity conditions, using APG-based methods as the inner solver (Melo et al., 2024; Lin et al., 2022), Hom-PALM*(θ) *with a quadratic proximal term converges to an* ϵ*-approximate stationary point (Def. E.6) of problem* (**H**) *with the following complexity bounds, where hidden constants depend on problem parameters but are independent of* ϵ*: (i)* ***Non-convex objective with linear equality constraints****: Hom-PALM*(θ) *(*$\theta \in (0, 1)$*) achieves* $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$ *complexity. (ii)* ***(Non-)convex objective with non-linear equality constraints****: Hom-PALM*(1) *achieves* $\widetilde{\mathcal{O}}(\epsilon^{-3})$ *complexity.*

Remarks. (a) The above complexity bounds match those of state-of-the-art methods for non-convex constrained optimization (Melo et al., 2024; Lin et al., 2022). **(b)** The inner solvers, whether the proposed Hom-APG in the convex setting or APG-based algorithms in the non-convex setting, share the same algorithmic spirit: combining proximal gradient steps with acceleration. Moreover, they operate over the unit ball in the transformed space, which significantly reduces computational cost compared to solving subproblems in standard ALM-based methods. **(c)** While Hom-APG can also be applied in the non-convex setting, it does not provide additional benefits due to the inherent non-convexity in the original space. Thus, we directly apply APG-based methods in the transformed space and achieve convergence rates matching those in the original space. **(d)** Theorem 2 also applies to

non-convex inequality sets that are homeomorphic to a unit ball, such as star-convex sets. By iterating over the transformed space, we not only avoid expensive projections but also achieve state-of-the-art convergence rates for optimization over non-convex constraint sets (Lin et al., 2022).

4.3 Complexity Analysis

Now, we provide the total run-time complexity of Hom-PALM(θ), including initialization complexity, AL subproblem complexity, and last-step complexity. All the omitted details can be found in Appendix B.3. Specifically, the overall complexity of Hom-PALM(θ) consists of: **(a)** *Initialization*: One-time interior point oracle call to find a feasible point in $\mathcal{C}$. **(b)** *Hom-PALM subproblems*: Each of the iterations costs $\tilde{\mathcal{O}}(n^2)$. **(c)** *Final mapping*: $\tilde{\mathcal{O}}(n)$ to map the solution back via $\mathbf{x}^* = \Psi(\mathbf{z}^*)$.

In total, the runtime is $\tilde{\mathcal{O}}(n^2 \times \text{CvR})$ where CvR denotes the total convergence complexity (see Sec. 4.2 for details), which is significantly lower than the $\mathcal{O}(n^3)$ per-iteration cost of second-order methods and many first-order methods needed to solve equality constraints, demonstrating the scalability of Hom-PALM(θ) to high-dimensional problems.

5 Empirical Study

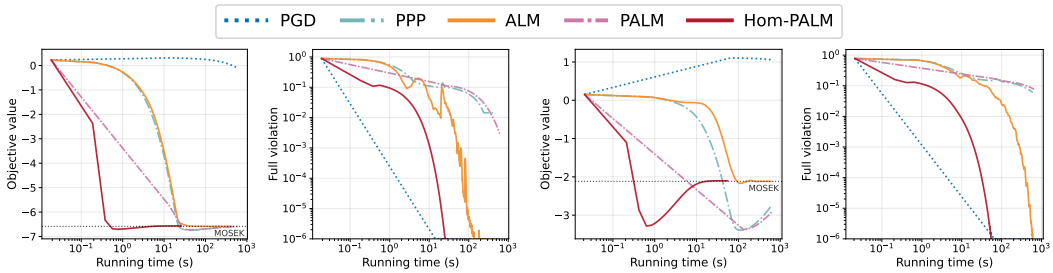
In this section, we evaluate the proposed Hom-PALM method through experiments on two problem classes: synthetic convex QCQP problems and Stiefel manifold optimization for principal component analysis. We denote each instance by `(n,p,m)`, where n is the number of decision variables, p is the number of equality constraints, and m is the number of inequality constraints.

Baselines. We compare against representative first-order methods: projected gradient descent (**PGD**), proximal-point penalty methods (**PPP**); ALM-type methods can handle both equality and inequality constraints in practice; however, in convergence analysis, inequality constraints are typically treated as indicator functions composed with the objective. To clarify this distinction, we adopt the following notation: **ALM** converts all constraints into penalty or augmented Lagrangian terms, while **PALM** treats inequality constraints as indicator functions and applies the augmented Lagrangian formulation only to equality constraints. Problem-specific solvers, such as **Retraction**-based methods for manifold optimization, are introduced in their respective subsections. We use commercial solvers—**MOSEK** for convex problems and **IPOPT** for non-convex problems—to obtain ground-truth solutions and solve constrained subproblems within the baseline methods. All baselines are executed with the same maximum (outer×inner) iteration budgets of `200,000` and a maximum running time of `600` seconds.

Metrics. For each method, we track the objective value, equality-constraint violation, inequality-constraint violation, total iteration counts, and wall-clock time. We report wall-clock convergence of objective value and maximum residual over equality and inequality constraints (i.e., full violation) in the main text and provide the complete experimental results in Appendix F.

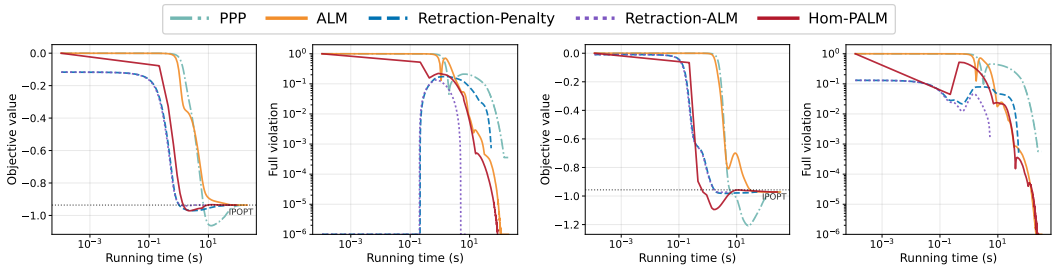
5.1 Convex Quadratically Constrained Quadratic Program

We consider convex QCQP instances with *linear equality constraints* and *convex quadratic inequalities*, following standard CVXPY constructions. This benchmark isolates a convex setting, but projection onto the full/inequality feasible set remains expensive due to the simultaneous presence of linear and quadratic constraints. We test two instances with identical variable counts but different constraint counts to examine scalability as the system becomes more restrictive. Figure 2 shows that Hom-PALM approaches the MOSEK reference objective while maintaining small constraint violations across both settings. PGD reduces feasibility residuals but stalls far from optimality, reflecting the limitation of projection without a primal-dual mechanism. ALM and PALM improve the objective but require longer runtime, especially as equality constraints increase. These results validate Hom-PALM's



(a) ((n,p,m)=(1000,100,2100)) (b) ((n,p,m)=(1000,500,2500))

Figure 2: Convex QCQP (complete visualizations are in Appendix F.6).



(a) ((n,p,m)=(192, 9, 385)) (b) ((n,p,m)=(2139, 9, 4279))

Figure 3: Stiefel Manifold Optimization (complete visualizations are in Appendix F.7).

design: the homeomorphic map eliminates expensive projections onto the convex inequality set, while the augmented Lagrangian structure handles equality constraints effectively.

5.2 Nonlinear Stiefel Manifold Optimization

We evaluate Hom-PALM on constrained PCA over the Stiefel manifold using the DIGITS and LFW datasets from Scikit-learn (Pedregosa et al., 2018). This setting complements the QCQP benchmark: the Stiefel constraint introduces *nonlinear equalities*, while group-budget constraints add *convex inequalities*. We also compare against problem-specific retraction-based methods that iterate on the Stiefel manifold while using penalty or ALM techniques to handle the remaining inequality constraints.

Figure 3 shows that Hom-PALM reaches objective values close to or better than the IPOPT reference while substantially reducing constraint violations. Retraction-based methods leverage manifold structure for fast objective descent but converge slowly in terms of inequality constraint satisfaction. Direct ALM converges more slowly overall, particularly on the larger LFW instance. These results demonstrate that Hom-PALM offers an efficient alternative for handling nonlinear equalities and convex inequalities without requiring problem-specific retractions.

6 Limitations

Our analysis reveals several **limitations**. First, the convergence rates depend on the regularity of the explicitly constructed homeomorphism; developing more structured alternatives beyond gauge mappings remains fundamental to advancing reparameterization-based methods. Second, our non-convex analysis inherits existing ALM/PPP proof techniques under additional assumptions. Strengthening proof techniques tailored to the transformed non-convex geometry is a natural direction for future work. From an application perspective, integrating such projection-free methods via GPU-based implementations for large-scale real-world problems presents another promising avenue.

References

P-A Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

Roberto Andreani, Kelvin R Couto, Orizon P Ferreira, and Gabriel Haeser. Constraint qualifications and strong global convergence properties of an augmented lagrangian method on riemannian manifolds. *SIAM Journal on Optimization*, 34(2):1799–1825, 2024.

Necdet S Aybat and Garud Iyengar. A first-order augmented lagrangian method for compressed sensing. *SIAM Journal on Optimization*, 22(2):429–459, 2012.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153):1–43, 2018.

Amir Beck. *Introduction to nonlinear optimization: Theory, algorithms, and applications with MATLAB*. SIAM, 2014.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. Deep frank-wolfe for neural network optimization. *arXiv preprint arXiv:1811.07591*, 2018.

Dimitri P Bertsekas. On penalty and multiplier methods for constrained minimization. *SIAM Journal on Control and Optimization*, 14(2):216–235, 1976.

Dimitri P Bertsekas. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334, 1997.

Franco Blanchini and Stefano Miani. *Set-theoretic methods in control*, volume 78. Springer, 2008.

Immanuel M Bomze, Francesco Rinaldi, and Damiano Zeffiro. Frank–wolfe and friends: a journey into projection-free first-order optimization methods. *Annals of Operations Research*, 343(2):607–638, 2024.

Lahcen El Bourkhissi and Ion Necoara. Complexity of linearized augmented lagrangian for optimization with nonlinear equality constraints. *arXiv preprint arXiv:2301.08345*, 2023.

Lahcen El Bourkhissi, Ion Necoara, Panagiotis Patrinos, and Quoc Tran-Dinh. Complexity of linearized perturbed augmented lagrangian methods for nonsmooth nonconvex optimization with nonlinear equality constraints. *arXiv preprint arXiv:2503.01056*, 2025.

Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Gábor Braun, Sebastian Pokutta, Dan Tu, and Stephen Wright. Blended conditonal gradients. In *International conference on machine learning*, pp. 735–743. PMLR, 2019.

Richard L Burden and J Douglas Faires. Numerical analysis, brooks, 1997.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International conference on machine learning*, pp. 654–663. PMLR, 2017.

Yair Carmon, Arun Jambulapati, Qijia Jiang, Yujia Jin, Yin Tat Lee, Aaron Sidford, and Kevin Tian. Acceleration with a ball optimization oracle. *Advances in Neural Information Processing Systems*, 33:19052–19063, 2020.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40: 120–145, 2011.

Xin Chen, Niao He, Yifan Hu, and Zikun Ye. Efficient algorithms for minimizing compositions of convex functions and random functions and its applications in network revenue management. *arXiv preprint arXiv:2205.01774*, 2022.

James Chok and Geoffrey M Vasil. Optimization over a probability simplex. *Journal of Machine Learning Research*, 26(73):1–35, 2025.

Diego Cifuentes. On the burer–monteiro method for general semidefinite programs. *Optimization Letters*, 15(6):2299–2309, 2021.

Cyrille Combettes and Sebastian Pokutta. Boosting frank-wolfe by chasing gradients. In *International conference on machine learning*, pp. 2111–2121. PMLR, 2020.

Cyrille W Combettes and Sebastian Pokutta. Complexity of linear minimization and projection on some sets. *Operations Research Letters*, 49(4):565–571, 2021.

Christopher Criscitiello and Nicolas Boumal. An accelerated first-order method for nonconvex optimization on manifolds. *Foundations of Computational Mathematics*, 23(4): 1433–1509, 2023.

Oscar Davis, Samuel Kessler, Mircea Petrache, Ismail Ilkan Ceylan, Michael Bronstein, and Avishek Joey Bose. Fisher flow matching for generative modeling over discrete data. *arXiv preprint arXiv:2405.14664*, 2024.

Vladimir Fedorovich Demianov and Aleksandr Moiseevich Rubinov. Approximate methods in optimization problems. *(No Title)*, 1970.

Kangkang Deng and Zheng Peng. A manifold inexact augmented lagrangian method for nonsmooth optimization on riemannian submanifolds in euclidean space. *IMA Journal of Numerical Analysis*, 43(3):1653–1684, 2023.

Kangkang Deng, Jiang Hu, Jiayuan Wu, and Zaiwen Wen. Oracle complexities of augmented lagrangian methods for nonsmooth manifold optimization. *arXiv preprint arXiv:2404.05121*, 2024.

Kangkang Deng, Rui Wang, Zhenyuan Zhu, Junyu Zhang, and Zaiwen Wen. The augmented lagrangian methods: Overview and recent advances. *arXiv preprint arXiv:2510.16827*, 2025.

Neil K Dhingra, Sei Zhen Khong, and Mihailo R Jovanović. The proximal augmented lagrangian method for nonsmooth composite optimization. *IEEE Transactions on Automatic Control*, 64(7):2861–2868, 2018.

Kuang-Yu Ding, Xin-Yee Lam, and Kim-Chuan Toh. On proximal augmented lagrangian based decomposition methods for dual block-angular convex composite programming problems. *Computational Optimization and Applications*, 86(1):117–161, 2023.

Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.

Joseph C Dunn. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization*, 17(2):187–211, 1979.

Ilyas Fatkhullin, Niao He, and Yifan Hu. Stochastic optimization under hidden convexity. *arXiv preprint arXiv:2401.00108*, 2023.

Mihai I Florea and Sergiy A Vorobyov. An accelerated composite gradient method for large-scale composite objective problems. *IEEE Transactions on Signal Processing*, 67(2): 444–459, 2018.

Marguerite Frank, Philip Wolfe, et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Robert M Freund and Paul Grigas. New analysis and results for the frank–wolfe method. *Mathematical Programming*, 155(1):199–230, 2016.

Dan Garber and Elad Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *International Conference on Machine Learning*, pp. 541–549. PMLR, 2015.

Dan Garber and Elad Hazan. A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3):1493–1528, 2016.

Caroline Geiersbach, Tim Suchan, and Kathrin Welker. Stochastic augmented lagrangian method in riemannian shape manifolds. *Journal of Optimization Theory and Applications*, 203(1):165–195, 2024.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

Benjamin Grimmer. Radial duality part i: foundations. *Mathematical Programming*, 205(1): 33–68, 2024a.

Benjamin Grimmer. Radial duality part ii: applications and algorithms. *Mathematical Programming*, 205(1):69–105, 2024b.

Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

Kaja Gruntkowska, Hanmin Li, Aadi Rane, and Peter Richtárik. The ball-proximal (=" broximal") point method: a new algorithm, convergence theory, and applications. *arXiv preprint arXiv:2502.02002*, 2025.

Jacques Guélat and Patrice Marcotte. Some comments on wolfe's away step. *Mathematical Programming*, 35(1):110–119, 1986.

Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4): 2927–2955, 2020.

Davood Hajinezhad and Mingyi Hong. Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization. *Mathematical Programming*, 176(1):207–245, 2019.

Bingsheng He and Xiaoming Yuan. On the acceleration of augmented lagrangian method for linearly constrained optimization. *Optimization online*, 3, 2010.

Ben Hermans, Andreas Themelis, and Panagiotis Patrinos. Qpalm: a proximal augmented lagrangian method for nonconvex quadratic programs. *Mathematical Programming Computation*, 14(3):497–541, 2022.

Magnus R Hestenes. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5):303–320, 1969.

Mingyi Hong and Zhi-Quan Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1):165–199, 2017.

Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *International Conference on Machine Learning*, pp. 1529–1538. PMLR, 2017.

Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *European conference on computer vision*, pp. 253–268. Springer, 2014.

Thomas Kerdreux, Lewis Liu, Simon Lacoste-Julien, and Damien Scieur. Affine invariant analysis of frank-wolfe on strongly convex sets. In *International conference on machine learning*, pp. 5398–5408. PMLR, 2021.

M Kojima. A primal-dual interior point algo-rithm for linear programming. *Progress in mathematical programming: Interior point and related methods/Springer-Verlag*, 1989.

Aritra Konar and Nicholas D Sidiropoulos. Hidden convexity in qcqp with toeplitz-hermitian quadratics. *IEEE Signal Processing Letters*, 22(10):1623–1627, 2015.

Weiwei Kong and Renato DC Monteiro. An accelerated inexact dampened augmented lagrangian method for linearly-constrained nonconvex composite optimization problems. *Computational Optimization and Applications*, 85(2):509–545, 2023.

Weiwei Kong, Jefferson G Melo, and Renato DC Monteiro. Fista and extensions–review and new insights. *arXiv preprint arXiv:2107.01267*, 2021.

Weiwei Kong, Jefferson G Melo, and Renato DC Monteiro. Iteration complexity of an inner accelerated inexact proximal augmented lagrangian method based on the classical lagrangian function. *SIAM Journal on Optimization*, 33(1):181–210, 2023a.

Weiwei Kong, Jefferson G Melo, and Renato DC Monteiro. Iteration complexity of a proximal augmented lagrangian method for solving nonconvex composite optimization problems with nonlinear convex constraints. *Mathematics of Operations Research*, 48(2):1066–1094, 2023b.

Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

Guanghui Lan and Renato DC Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. *Mathematical Programming*, 155(1):511–547, 2016.

Eitan Levin, Joe Kileel, and Nicolas Boumal. The effect of smooth parametrizations on nonconvex optimization landscapes. *Mathematical Programming*, pp. 1–49, 2024.

Evgeny S Levitin and Boris T Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 6(5):1–50, 1966.

Fei Li and Zheng Qu. An inexact proximal augmented lagrangian framework with arbitrary linearly convergent inner solver for composite convex optimization. *Mathematical Programming Computation*, 13(3):583–644, 2021.

Huan Li and Zhouchen Lin. Accelerated alternating direction method of multipliers: An optimal o (1/k) nonergodic analysis. *Journal of Scientific Computing*, 79(2):671–699, 2019.

Qiuwei Li, Daniel McKenzie, and Wotao Yin. From the simplex to the sphere: faster constrained optimization using the hadamard parametrization. *Information and Inference: A Journal of the IMA*, 12(3):1898–1937, 2023.

Xinpeng Li, Enming Liang, and Minghua Chen. Gauge flow matching for efficient constrained generative modeling over general convex set. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025.

Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Rate-improved inexact augmented lagrangian method for constrained nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2170–2178. PMLR, 2021.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18 (212):1–54, 2018.

Qihang Lin and Lin Xiao. An adaptive accelerated proximal gradient method and its homotopy continuation for sparse optimization. In *International Conference on Machine Learning*, pp. 73–81. PMLR, 2014.

Qihang Lin, Runchao Ma, and Yangyang Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational optimization and applications*, 82(1):175–224, 2022.

Zhouchen Lin, Huan Li, and Cong Fang. Accelerated optimization for machine learning. *Nature Singapore: Springer*, 2020.

Changshuo Liu and Nicolas Boumal. Simple algorithms for optimization on riemannian manifolds with constraints. *Applied Mathematics & Optimization*, 82(3):949–981, 2020.

Chenghao Liu, Enming Liang, and Minghua Chen. Fast projection-free approach (without optimization oracle) for optimization over compact convex set. *Advances in Neural Information Processing Systems*, 2025.

Chenghao Liu, Enming Liang, and Minghua Chen. Hom-pgd+: Fast reparameterized optimization over non-convex ball-homeomorphic set. In *Proceedings of the Forty-third International Conference on Machine Learning*. PMLR, 2026.

Ning Liu and Benjamin Grimmer. Gauges and accelerated optimization over smooth and/or strongly convex sets. *arXiv preprint arXiv:2303.05037*, 2023.

Ya-Feng Liu, Xin Liu, and Shiqian Ma. On the nonergodic convergence rate of an inexact augmented lagrangian framework for composite convex programming. *Mathematics of Operations Research*, 44(2):632–650, 2019.

Songtao Lu. A single-loop gradient descent and perturbed ascent algorithm for nonconvex functional constrained optimization. In *International conference on machine learning*, pp. 14315–14357. PMLR, 2022.

Zhaosong Lu and Zirui Zhou. Iteration-complexity of first-order augmented lagrangian methods for convex conic programming. *SIAM journal on optimization*, 33(2):1159–1190, 2023.

Zhou Lu, Nataly Brukhim, Paula Gradu, and Elad Hazan. Projection-free adaptive regret with membership oracles. In *International Conference on Algorithmic Learning Theory*, pp. 1055–1073. PMLR, 2023.

Zhi-Quan Luo and Paul Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72 (1):7–35, 1992.

Zhi-Quan Luo and Paul Tseng. On the convergence rate of dual ascent methods for linearly constrained convex minimization. *Mathematics of Operations Research*, 18(4):846–867, 1993a.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993b.

Hoomaan Maskan, Yikun Hou, Suvrit Sra, and Alp Yurtsever. Revisiting frank-wolfe for structured nonconvex optimization. *arXiv preprint arXiv:2503.08921*, 2025.

Sanjay Mehrotra. On the implementation of a primal-dual interior point method. *SIAM Journal on optimization*, 2(4):575–601, 1992.

Jefferson G Melo, Renato DC Monteiro, and Hairong Wang. A proximal augmented lagrangian method for linearly constrained nonconvex composite optimization problems. *Journal of Optimization Theory and Applications*, 202(1):388–420, 2024.

Zakaria Mhammedi. Efficient projection-free online convex optimization with membership oracle. In *Conference on Learning Theory*, pp. 5314–5390. PMLR, 2022.

Bamdev Mishra, Gilles Meyer, Silvere Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, 29:591–621, 2014.

Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of machine learning research*, 21(105):1–49, 2020.

Renato DC Monteiro, Camilo Ortiz, and Benar F Svaiter. An adaptive accelerated first-order method for convex optimization. *Computational Optimization and Applications*, 64(1): 31–73, 2016.

Cun Mu, Yuqian Zhang, John Wright, and Donald Goldfarb. Scalable robust matrix recovery: Frank–wolfe meets proximal methods. *SIAM Journal on Scientific Computing*, 38(5): A3291–A3317, 2016.

Ricardo Ñanculef, Emanuele Frandi, Claudio Sartori, and Héctor Allende. A novel frank–wolfe algorithm. analysis and applications to large-scale svm training. *Information Sciences*, 285:66–99, 2014.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

Yu E Nesterov. Introductory lectures on convex optimization. a basic course. 2004.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

Jong-Shi Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1):299–332, 1997.

Andrei Patrascu, Ion Necoara, and Quoc Tran-Dinh. Adaptive inexact fast augmented lagrangian methods for constrained convex optimization. *Optimization Letters*, 11(3): 609–626, 2017.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python, 2018.

Javier F Pena. Affine invariant convergence rates of the conditional gradient method. *SIAM Journal on Optimization*, 33(4):2654–2674, 2023.

Clarice Poon and Gabriel Peyré. Smooth over-parameterized solvers for non-smooth structured optimization. *Mathematical programming*, 201(1):897–952, 2023.

Michael JD Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pp. 283–298, 1969.

Wenqiang Pu, Kaizhao Sun, and Jiawei Zhang. Smoothed proximal lagrangian method for nonlinear constrained programs. *arXiv preprint arXiv:2408.15047*, 2024.

Akshay Ramachandran, Kevin Shu, and Alex L Wang. Hidden convexity, optimization, and algorithms on rotation matrices. *Mathematics of Operations Research*, 2024.

R Tyrrell Rockafellar. Augmented lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of operations research*, 1(2):97–116, 1976.

Mehmet Fatih Sahin, Ahmet Alacaoglu, Fabian Latorre, Volkan Cevher, et al. An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Thabo Samakhoana and Benjamin Grimmer. Scalable projection-free optimization methods via multiradial duality theory. *arXiv preprint arXiv:2403.13688*, 2024.

Qiankun Shi, Xiao Wang, and Hao Wang. Bregman linearized augmented lagrangian method for nonconvex constrained stochastic zeroth-order optimization. *arXiv preprint arXiv:2504.09409*, 2025a.

Qiankun Shi, Xiao Wang, and Hao Wang. A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. *Mathematics of Operations Research*, 2025b.

Qiankun Shi, Xiao Wang, and Hao Wang. A momentum-based linearized augmented lagrangian method for nonconvex constrained stochastic optimization. *Mathematics of Operations Research*, 51(1):92–133, 2026.

Alice E Smith, David W Coit, Thomas Baeck, David Fogel, and Zbigniew Michalewicz. Penalty functions. *Handbook of evolutionary computation*, 97(1):C5, 1997.

Arnesh Sujanani and Renato DC Monteiro. An adaptive superfast inexact proximal augmented lagrangian method for smooth nonconvex composite optimization problems. *Journal of Scientific Computing*, 97(2):34, 2023.

Daniel Tabas and Baosen Zhang. Computationally efficient safe reinforcement learning for power systems. In *2022 American Control Conference (ACC)*, pp. 3303–3310. IEEE, 2022a.

Daniel Tabas and Baosen Zhang. Safe and efficient model predictive control using neural networks: An interior point approach. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 1142–1147. IEEE, 2022b.

Kenya Tajima, Yoshihiro Hirohashi, Esmeraldo Ronnie Rey Zara, and Tsuyoshi Kato. Frank-wolfe algorithm for learning svm-type multi-category classifiers. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pp. 432–440. SIAM, 2021.

Tianyun Tang and Kim-Chuan Toh. Optimization over convex polyhedra via hadamard parametrizations. *Mathematical Programming*, pp. 1–41, 2024.

Jesus Tordesillas, Jonathan P How, and Marco Hutter. Rayen: Imposition of hard convex constraints on neural networks. *arXiv preprint arXiv:2307.08336*, 2023.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

Elias Wirth, Thomas Kerdreux, and Sebastian Pokutta. Acceleration of frank-wolfe algorithms with open-loop step-sizes. In *International Conference on Artificial Intelligence and Statistics*, pp. 77–100. PMLR, 2023.

Elias Wirth, Javier Pena, and Sebastian Pokutta. Fast convergence of frank-wolfe algorithms on polytopes. *arXiv preprint arXiv:2406.18789*, 2024.

Elias Wirth, Javier Pena, and Sebastian Pokutta. Accelerated affine-invariant convergence rates of the frank–wolfe algorithm with open-loop step-sizes. *Mathematical Programming*, pp. 1–45, 2025.

Philip Wolfe. Convergence theory in nonlinear programming. *Integer and nonlinear programming*, pp. 1–36, 1970.

Yue Xie and Stephen J Wright. Complexity of proximal augmented lagrangian for nonconvex optimization with nonlinear equality constraints. *arXiv preprint arXiv:1908.00131*, 2019.

Yangyang Xu. Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. *SIAM Journal on Optimization*, 27(3):1459–1484, 2017.

Yangyang Xu. First-order methods for constrained convex programming based on linearized augmented lagrangian function. *INFORMS Journal on Optimization*, 3(1):89–117, 2021.

Junfeng Yang and Xiaoming Yuan. Linearized augmented lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of computation*, 82(281):301–329, 2013.

Yinyu Ye. Combining binary search and newton's method to compute real roots for a class of real functions. *Journal of Complexity*, 10(3):271–280, 1994.

Yinyu Ye. Quadratic programming over an ellipsoid. In *Encyclopedia of Optimization*, pp. 2112–2116. Springer, 2001.

Jiawei Zhang and Zhi-Quan Luo. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 30(3): 2272–2302, 2020.

Jiawei Zhang and Zhi-Quan Luo. A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. *SIAM Journal on Optimization*, 32(3):2319–2346, 2022.

Jiawei Zhang, Wenqiang Pu, and Zhi-Quan Luo. On the iteration complexity of smoothed proximal alm for nonconvex optimization problem with convex constraints. *arXiv preprint arXiv:2207.06304*, 2022.

Zhenyuan Zhu, Fan Chen, Junyu Zhang, and Zaiwen Wen. On the optimal lower and upper complexity bounds for a class of composite optimization problems. *arXiv preprint arXiv:2308.06470*, 2023.

Contents

A Related Work **19**
- A.1 Classical First-Order Methods . . . 19
- A.2 Augmented Lagrangian Method . . . 20
- A.3 Accelerated Gradient Methods . . . 21
- A.4 Geometry-changing reformulations . . . 22

B Homeomorphism between a Convex set and a Ball via Gauge Mapping **24**
- B.1 Gauge Mapping and Properties . . . 24
- B.2 Computation of Gauge Mapping . . . 26
- B.3 Complexity Analysis . . . 26

C Preliminaries for Technical Proof **28**
- C.1 Basic Concepts . . . 28
- C.2 Basic Assumptions and Notations . . . 29
- C.3 Basic Facts . . . 29

D Landscape Analysis **30**
- D.1 Action of Homeomorphism on a Constrained Set . . . 30
- D.2 Properties of AL Functions . . . 31
- D.3 Global Optimality Property of Optimization Problem (**H**) . . . 32
- D.4 Strong Duality of Problem (**H**) . . . 34
- D.5 KKT Conditions of Problem (**P**) and (**H**) . . . 35
- D.6 Relationships of KKT Stationary Points between Problem (**P**) and (**H**) . . . 36

E Convergence Analysis **39**
- E.1 Hom-APG . . . 39
- E.2 Convex Case . . . 43
- E.3 Non-Convex Case: Non-Convex Objective with Linear Equality . . . 46
- E.4 Non-Convex Case: Non-Convex Objective with Non-Linear Equality . . . 48

F Experiments **52**
- F.1 Reporting Metric . . . 52
- F.2 (Baseline) Algorithms . . . 52
- F.3 Convex QCQP with Linear Equalities . . . 53
- F.4 Constrained PCA on the Stiefel Manifold . . . 53
- F.5 Implementation Details . . . 54
- F.6 Complete QCQP Convergence Plots . . . 54
- F.7 Complete Stiefel Convergence Plots . . . 56

A Related Work

We discuss related work on methods for reducing computational costs and achieving speedups in *constrained* optimization, organized into four parts: (i) classical first-order methods, (ii) (proximal) augmented Lagrangian methods, (iii) accelerated gradient methods, and (iv) geometry-transforming reformulations.

A.1 Classical First-Order Methods

Projection-based gradient methods. A standard approach for constrained optimization applies projected (proximal) gradient descent (PGD), which takes a gradient step and then projects back to feasibility, i.e.,

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{C}} \left(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k) \right)$$

(and similarly with a proximal step in the composite case). Under standard smoothness assumptions, PGD attains the usual first-order rates, see e.g., (Beck, 2014; 2017; Burden & Faires, 1997; Bertsekas, 1997): sublinear convergence $\mathcal{O}(1/K)$ in the convex setting, linear convergence $\mathcal{O}(Q^K)$ for some constant $Q < 1$ under additional strong convexity, and convergence $\mathcal{O}(1/\sqrt{K})$ to first-order stationary points in the nonconvex objective setting, where K is the iteration number. While these guarantees are well understood, their practicality hinges on the availability of *cheap* projections. For general convex sets $\mathcal{C}$ described by inequality systems, projections can require solving a nontrivial convex program (e.g., a QP), leading to high per-iteration cost.

To address the high computational cost of Euclidean projections onto complicated feasible sets, a large body of work develops alternatives that avoid projections or reduce their per-iteration burden. Despite their success, they still have various limitations.

Frank–Wolfe methods. Frank-Wolfe (FW) (a.k.a. conditional gradient) methods are projection-free first-order schemes that are simple to implement and enjoy affine-invariance (Lan, 2013; Kerdreux et al., 2021; Pena, 2023). A practical appeal is that FW iterates are convex combinations of extreme points, which often yields sparse/structured solutions and has made FW popular in machine learning applications (Ñanculef et al., 2014; Joulin et al., 2014; Berrada et al., 2018; Mokhtari et al., 2020; Tajima et al., 2021). The convergence theory of classical FW is well developed (Levitin & Polyak, 1966; Demianov & Rubinov, 1970; Dunn, 1979; Guélat & Marcotte, 1986). Nevertheless, FW has two well-known drawbacks: each step relies on an efficient linear minimization oracle (LMO) over the feasible set (Combettes & Pokutta, 2021), and the worst-case rates can be slower than projection-based counterparts (Bomze et al., 2024; Freund & Grigas, 2016). In particular, under Wolfe's lower-bound construction (Wolfe, 1970), the vanilla FW algorithm cannot be faster than $\mathcal{O}(\epsilon^{-1+\delta})$ for any $\delta > 0$. Improving upon this barrier typically requires strengthened assumptions or algorithmic variants (e.g., away-steps, blended/boosted and accelerated variants) (Garber & Hazan, 2015; 2016; Braun et al., 2019; Combettes & Pokutta, 2020; Wirth et al., 2023; 2025); see also (Pena, 2023; Wirth et al., 2024; Maskan et al., 2025) for more refined analyses.

Penalty methods. Penalty approaches remove explicit constraints by adding penalty terms, but they often become increasingly ill-conditioned as the penalty parameter grows, which can severely slow down first-order solvers for complicated constraint sets. Primal–dual techniques are a standard alternative, particularly for linear constraints (Kojima, 1989; Mehrotra, 1992; Chambolle & Pock, 2011), although second-order primal–dual variants can be difficult to scale due to their high per-iteration cost. Furthermore, the state-of-the-art work (Lin et al., 2022) improves penalty-based methods by adding a proximal term and achieves the fastest convergence rate $\mathcal{O}(\epsilon^{-3})$ for non-convex optimization problems with weakly convex constraints in the existing literature, under some regularization assumption.

First-order primal–dual methods. First-order primal–dual algorithms update primal and dual variables using (possibly inexact) gradient/proximal steps (Hong et al., 2017; Chambolle & Pock, 2011). Establishing global convergence and rates is especially challenging

Table 2: Summary of existing augmented-Lagrangian-based approaches. We consider constrained optimization problems of the form $\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x})$, s.t. $\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{0}$, where f is the objective function, g is a proper, closed, convex function (e.g., the indicator function of a convex set defined by inequality constraints), and $\boldsymbol{\varphi}$ represents the equality constraints.

Problem Settings	References	Regularity	Methods	Subproblem Solver	Total Complexity
convex f, linear $\boldsymbol{\varphi}$	(Lu & Zhou, 2023)	A(I,II)	ALM	NAG	$\mathcal{O}(\epsilon^{-7/4})$
	(Patrascu et al., 2017)	A(II)		NAG	$\mathcal{O}(\epsilon^{-1})$
	(Liu et al., 2019)	A(I,II)		FISTA or FW	$\mathcal{O}(\epsilon^{-2})$
	(Li & Qu, 2021)	A(II) A(II), SC g	proximal ALM	NAG-based	$\mathcal{O}(\epsilon^{-1})$ $\mathcal{O}(\epsilon^{-1/2})$
	(Xu, 2017)	-	linearized ALM	explicit	$\mathcal{O}(\epsilon^{-1})$
NC f, linear $\boldsymbol{\varphi}$	(Kong et al., 2023a)	A(I,V,VI)	proximal ALM	ACG	$\mathcal{O}(\epsilon^{-3})$
	(Hajinezhad & Hong, 2019)	A(I,IV)	proximal linearized ALM	PPM	$\mathcal{O}(\epsilon^{-3})$
	(Melo et al., 2024)	A(V,VI)	proximal ALM	ACG	$\mathcal{O}(\epsilon^{-2.5})$
	(Li et al., 2021)	A(I,IV)	(proximal) ALM	APG	$\mathcal{O}(\epsilon^{-2.5})$
NC f, NL $\boldsymbol{\varphi}$	(Xie & Wright, 2019)	A(I,III), $g \equiv 0$	proximal ALM	NCG	$\mathcal{O}(\epsilon^{-2})$
	(Sahin et al., 2019)	A(I,IV)	ALM	APG	$\mathcal{O}(\epsilon^{-4})$
	(Li et al., 2021)	A(I,IV)	(proximal) ALM	APG	$\mathcal{O}(\epsilon^{-3})$
	(Kong et al., 2023b)	A(I,V,VI)	proximal ALM	ACG	$\mathcal{O}(\epsilon^{-3})$
	(Lin et al., 2022)	A(I,IV)	PPP	ACG	$\mathcal{O}(\epsilon^{-3})$

[1] **Basis abbreviations**: NC = nonconvex, SC = strongly convex, (N)L = (non-)linear.
[2] **Algorithm abbreviations:** ALM = augmented Lagrangian method, NAG = Nesterov's accelerated gradient, APG = accelerated proximal gradient, PPM = proximal-point method, ACG = accelerated composite gradient, AdapAPG = adaptive APG, FW = Frank-Wolfe, FISTA = fast iterative shrinkage-thresholding algorithm, NCG = Newton conjugate gradient, PPP = proximal-point penalty.
[3] **Assumptions.** A(I): bounded domain, i.e., dom(g) is bounded and $\sup_{\mathbf{x}\in\mathrm{dom}(g)} \|\boldsymbol{\varphi}(\mathbf{x})\|$ is finite. A(II): strong duality. A(III): full rank of $\mathrm{J}_{\boldsymbol{\varphi}}$ on a bounded level set. A(IV): KL regularity. A(V): the function g is Lipschitz continuous restricted to its domain. A(VI): Slater's condition.
[4] **Complexity notion**: iteration complexity for an ϵ-approximate optimum (convex) or an ϵ-stationary point (nonconvex).

in the nonconvex setting (Zhang & Luo, 2022), and many classical results focus on linear constraints (Luo & Tseng, 1993a; Chambolle & Pock, 2011). Error-bound conditions provide a common route to rates in convex problems (Luo & Tseng, 1993b; Pang, 1997; Luo & Tseng, 1992; Hong & Luo, 2017), but they are typically local (requiring iterates to be close to the solution set) and depend on constants that are hard to estimate in practice. More recently, (Zhang & Luo, 2020; 2022; Zhang et al., 2022) proposed a smooth augmented-Lagrangian primal–dual method with optimal worst-case complexity for nonconvex objectives, at the cost of additional *projection* work in each iteration. Further background on augmented Lagrangian methods is deferred to Sec. A.2.

A.2 Augmented Lagrangian Method

Augmented Lagrangian Methods (ALM). The ALM is a classical and widely used approach for solving constrained optimization problems. It addresses equality-constrained problems by dualizing and penalizing the constraints, and seeks saddle points of the resulting augmented Lagrangian function. Specifically, for problem **P**, the update rules of ALM are

$$\mathbf{x}_{k+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{C}} \left\{ f(\mathbf{x}) + (\boldsymbol{\eta}_k)^\top \boldsymbol{\varphi}(\mathbf{x}) + \frac{\rho_k}{2}\|\boldsymbol{\varphi}(\mathbf{x})\|^2 \right\}, \tag{2}$$

$$\boldsymbol{\eta}_{k+1} \leftarrow \boldsymbol{\eta}_k + \rho_k \boldsymbol{\varphi}(\mathbf{x}_{k+1}). \tag{3}$$

In convex settings, the convergence of inexact ALM has been well established, with rigorous iteration complexity analysis provided in works such as (Aybat & Iyengar, 2012; Lan & Monteiro, 2016; Liu et al., 2019; Patrascu et al., 2017; Zhu et al., 2023; Lu & Zhou, 2023). Moreover, ALM has been extended to nonconvex problems, where either the objective or the constraints (or both) are nonconvex. Under suitable regularity conditions, several studies have analyzed the convergence behavior and complexity of ALM in this setting (Kong et al., 2023b;a; Li et al., 2021; Hajinezhad & Hong, 2019; Melo et al., 2024; Sahin et al., 2019; Xie & Wright, 2019). One could refer to Table 2 for a summary.

Proximal ALM. Proximal ALM introduces a proximal term for the ALM subproblem as a regularization to improve practical performance (Rockafellar, 1976; Dhingra et al., 2018; Hermans et al., 2022). In the convex setting, the convergence analysis of proximal ALM is established in (Li & Qu, 2021; Ding et al., 2023). Moreover, the convergence analysis of proximal ALM has also been widely studied in various non-convex settings (Melo et al., 2024; Kong et al., 2023b; Sujanani & Monteiro, 2023; Hermans et al., 2022; Zhang et al., 2022). For non-convex optimization over convex constraints, (Zhang et al., 2022) proposes smoothed proximal ALM that achieves a convergence rate $\mathcal{O}(1/\epsilon^2)$ to find an ϵ-approximate stationary point, which is the fastest among all the ALM-based methods, even though the convergence rate is local. Further, accelerated variants aim to improve theoretical convergence rates through Nesterov-type acceleration or related techniques (Kong et al., 2023a;b; Kong & Monteiro, 2023; Xu, 2017).

Other variants of ALM. To improve the efficiency of solving augmented-Lagrangian (AL) subproblems, many variants of ALM have been developed. A prominent line is *linearized* ALM, which reduces inner-loop cost by replacing the exact subproblem with a first-order (typically proximal) approximation (Yang & Yuan, 2013; Xu, 2021; Lu, 2022; Bourkhissi & Necoara, 2023; Bourkhissi et al., 2025). Recent refinements further enhance this paradigm, including Bregman-type linearizations (Shi et al., 2025a) and momentum-accelerated linearized ALM schemes (Shi et al., 2025b). Collectively, these results demonstrate that ALM can be made both flexible and scalable across convex and nonconvex regimes. Beyond deterministic settings, ALM variants have also been proposed for broader problem classes, such as stochastic ALM for stochastic optimization (Shi et al., 2026; 2025a; Geiersbach et al., 2024; Deng et al., 2024) and ALM-based approaches for manifold-constrained optimization (Deng & Peng, 2023; Deng et al., 2024).

Highly Related Work. Zhang & Luo (2022) considers a linearly constrained nonconvex problem. While it attains an $O(\epsilon^{-2})$ complexity, the algorithm relies on projection onto the linear (inequality) constraint set at each iteration (i.e., solving a QP subproblem per iteration), i.e., it is not projection-free. Moreover, its analysis is tied to the linear structure of the constraints and does not directly extend to general convex inequality constraints. Pu et al. (2024) extends the framework to nonlinear inequality constraints and also achieves $O(\epsilon^{-2})$ complexity. However, this method may not handle equality constraints directly, as they note, handling equality constraints simultaneously would require a more complex primal update whose subproblem enforces the equality constraint, or addressing this via first-order ALM or penalty approaches typically introduces multiple loops that may detract from the desired $O(\epsilon^{-2})$ complexity, which restricts its applicability in problems where both constraint types appear. In contrast, our Hom-PALM handles general convex inequality constraints and non-linear equality constraints while remaining projection-free (without requiring expensive optimization oracles such as proximal/linear optimization at each iteration).

We refer interested readers to a comprehensive overview (Deng et al., 2025) for further reading.

A.3 Accelerated Gradient Methods

Gradient descent achieves a convergence rate of $\mathcal{O}(1/\epsilon)$ for convex L_f-smooth objectives and $\mathcal{O}((L_f/\mu_f)\log(1/\epsilon))$ for μ_f-strongly convex L_f-smooth objectives, see e.g., (Boyd et al., 2004; Beck, 2014; 2017). Nesterov's acceleration (Nesterov, 2004; 2013) improves these rates to $\mathcal{O}(1/\sqrt{\epsilon})$ and $\mathcal{O}(\sqrt{L_f/\mu_f}\log(1/\epsilon))$, respectively, which are optimal for first-order methods. These accelerated techniques have been extended to constrained optimization through various frameworks, including the accelerated proximal gradient (APG) and accelerated composite gradient (ACG) methods (Beck & Teboulle, 2009; Tseng, 2008; Nesterov, 2013; Florea & Vorobyov, 2018), with adaptive variants such as AdapAPG (Lin & Xiao, 2014; Monteiro et al., 2016). For projection-free optimization, accelerated Frank-Wolfe methods (Garber & Hazan, 2015; Wirth et al., 2025) achieve improved rates using only linear minimization oracles. Acceleration has also been incorporated into primal-dual frameworks, including the augmented Lagrangian method (ALM) (He & Yuan, 2010; Kong & Monteiro, 2023; Xu, 2017; Kong et al., 2023a) and the alternating direction method of multipliers (ADMM) (Li & Lin,

2019). Beyond convex optimization, accelerated methods have been developed for nonconvex settings (Carmon et al., 2017; Criscitiello & Boumal, 2023; Ghadimi & Lan, 2016), achieving improved complexity for finding approximate stationary points. We refer readers to (Lin et al., 2020) for a comprehensive treatment of accelerated first-order methods.

A.4 Geometry-changing reformulations

In addition to designing new update rules, another effective strategy for accelerating constrained optimization is to *reshape the problem itself*: by reformulating the constraints or reparameterizing the variables, one can replace expensive feasibility operations on a complicated set with much cheaper primitives (e.g., ball projections) or even obtain an unconstrained surrogate, while keeping the original solution set accessible.

Gauge viewpoints and radial duality. Recent progress leverages gauge functions and radial-duality ideas to obtain alternative representations of constrained problems. For a class of nonnegative optimization models, the radial duality theory developed in (Grimmer, 2024a;b) shows that the constraint system can be encoded through an associated gauge, yielding an equivalent unconstrained formulation. Building on this perspective, (Liu & Grimmer, 2023) derives projection-free algorithms with optimal complexity guarantees, and (Samakhoana & Grimmer, 2024) further relaxes conditions to design more practical line-search mechanisms for the resulting unconstrained objectives.

Reparameterization and domain simplification. Reparameterization methods seek a change of variables (often invertible) that preserves optimal solutions but improves algorithmic tractability—for instance, by smoothing nonsmooth structure or by mapping a difficult feasible region to a simpler reference domain. Such techniques appear in semidefinite programming (Cifuentes, 2021), low-rank matrix problems (Mishra et al., 2014; Ha et al., 2020), and statistical risk minimization (Bah et al., 2022). More recently, structured-domain reparameterizations have been proposed to facilitate cheap constraint handling, including Hadamard-type constructions that enable projection-free optimization over simplices (Li et al., 2023) and polyhedra (Tang & Toh, 2024), as well as smooth over-parameterizations designed to speed up optimization in nonsmooth regimes (Poon & Peyré, 2023). Related threads include continuous relaxations for discrete data in generative modeling (Davis et al., 2024) and analyses of how parameterizations influence nonconvex landscapes (Levin et al., 2024). Closest to our setting, (Liu et al., 2025) maps convex constrained problems to ball-constrained ones and achieves convex-like rates, and (Liu et al., 2026) extends this homeomorphic reparameterization method into optimization over a non-convex ball-homeomorphic constrained set. Different from prior reparameterization-based approaches that treat equalities via elimination, we incorporate general equality constraints through a proximal augmented-Lagrangian framework.

Hidden convexity. The term *hidden convexity* describes situations where an apparently nonconvex formulation becomes tractable once an appropriate transformation exposes an underlying convex structure. This idea has been exploited in rotation-matrix optimization (Ramachandran et al., 2024), nonlinear least squares (Drusvyatskiy & Paquette, 2019), revenue/inventory models (Chen et al., 2022), and special QCQP instances such as problems with Toeplitz–Hermitian quadratic forms (Konar & Sidiropoulos, 2015). In stochastic optimization, when such hidden structure is present, projected-gradient-type methods can recover rates comparable to convex optimization for both strongly convex (Fatkhullin et al., 2023) and merely convex objectives (Chen et al., 2022) under suitable assumptions. Moreover, although QCQP is NP-hard in general, hidden convexity can lead to polynomial-time solvability in certain cases (Konar & Sidiropoulos, 2015).

Ball-constrained optimization. Ball constraints occupy a particularly favorable computational regime: Euclidean projections are explicit, and several important subproblems admit highly efficient solvers. The classic ellipsoid method (Grötschel et al., 1981) can be viewed as maintaining an enclosing ellipsoid that shrinks over iterations; despite its strong theoretical properties, its $\mathcal{O}(n^4)$ scaling makes it unattractive in large dimensions. More recent work

studies acceleration using ball-optimization oracles in specific settings (Carmon et al., 2020; Gruntkowska et al., 2025). In addition, for quadratic objectives with a ball constraint, combining bisection with Newton-type updates yields very fast accuracy dependence, namely $\mathcal{O}(\log\log(1/\epsilon))$ (Ye, 1994; 2001), paralleling the rapid local behavior of Newton's method in unconstrained optimization (Boyd et al., 2004). These observations motivate algorithmic designs that reduce general constraints to ball-constrained subroutines whenever possible.

B Homeomorphism between a Convex set and a Ball via Gauge Mapping

In this section, we provide the omitted details related to gauge mapping in Sec. 2, Sec. 3 and Sec. 4.

B.1 Gauge Mapping and Properties

We first recall the definitions of gauge function/mapping.

Definition B.1 (Gauge/Minkowski function (Blanchini & Miani, 2008)). Let $\mathcal{C} \subset \mathbb{R}^n$ be a compact convex set with a non-empty interior. The gauge/Minkowski function $\gamma_{\mathcal{C}} : \mathbb{R}^n \times \text{int}(\mathcal{C}) \to \mathbb{R}_{\geq 0}$ is defined as

$$\gamma_{\mathcal{C}}(\mathbf{x}, \mathbf{x}^\circ) = \inf\{\lambda \geq 0 \mid \mathbf{x} \in \lambda(\mathcal{C} - \mathbf{x}^\circ)\},$$

where $\mathbf{x}^\circ \in \text{int}(\mathcal{C})$.

Building upon this foundation, we define the gauge mapping between two compact convex sets:

Definition B.2 (Gauge mapping (Tabas & Zhang, 2022a)). Let $\mathcal{Z}, \mathcal{C} \subset \mathbb{R}^n$ be compact convex sets with interior points $\mathbf{z}^\circ \in \text{int}(\mathcal{Z})$ and $\mathbf{x}^\circ \in \text{int}(\mathcal{C})$, respectively. Then

1) the gauge mapping $\Psi : \mathcal{Z} \to \mathcal{C}$ is defined as:

$$\Psi(\mathbf{z}) = \frac{\gamma_{\mathcal{Z}}(\mathbf{z} - \mathbf{z}^\circ, \mathbf{z}^\circ)}{\gamma_{\mathcal{C}}(\mathbf{z} - \mathbf{z}^\circ, \mathbf{x}^\circ)}(\mathbf{z} - \mathbf{z}^\circ) + \mathbf{x}^\circ,\ \mathbf{z} \in \mathcal{Z},$$

2) and the inverse gauge mapping $\Psi^{-1} : \mathcal{C} \to \mathcal{Z}$ is given by:

$$\Psi^{-1}(\mathbf{x}) = \frac{\gamma_{\mathcal{C}}(\mathbf{x} - \mathbf{x}^\circ, \mathbf{x}^\circ)}{\gamma_{\mathcal{Z}}(\mathbf{x} - \mathbf{x}^\circ, \mathbf{z}^\circ)}(\mathbf{x} - \mathbf{x}^\circ) + \mathbf{z}^\circ,\ \mathbf{x} \in \mathcal{C}.$$

We have the following remarks based on the definition.

- In essence, the gauge mapping scales the boundary of a convex set from an interior point to another convex set and with translation to its interior point.
- When $\mathcal{Z}$ is a unit ball, the gauge mapping in Def. B.2 is simplified as Def. 3.2:

$$\Psi(\mathbf{z}) = \frac{\|\mathbf{z}\|}{\gamma_{\mathcal{C}}(\mathbf{z}, \mathbf{x}^\circ)}\mathbf{z} + \mathbf{x}^\circ,\ \forall \mathbf{z} \in \mathcal{B}, \qquad \Psi^{-1}(\mathbf{x}) = \frac{\gamma_{\mathcal{C}}(\mathbf{x} - \mathbf{x}^\circ, \mathbf{x}^\circ)}{\|\mathbf{x} - \mathbf{x}^\circ\|}(\mathbf{x} - \mathbf{x}^\circ),\ \forall \mathbf{x} \in \mathcal{C}.$$

- At the center points, we define $\Psi(\mathbf{0}) = \mathbf{x}^\circ$ and $\Psi^{-1}(\mathbf{x}^\circ) = \mathbf{0}$ explicitly; the displayed formulas above are for the nonzero arguments.

First, it should be noted that for linear equality constraints, while $\mathbf{x}^\circ$ satisfies $\mathbf{A}\mathbf{x}^\circ = \mathbf{b}$, the transformed equality constraints $\boldsymbol{\phi}(\mathbf{z}) = 0$ are still linear, where the formal statement is shown below.

Proposition B.3. *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a compact convex set with interior point $\mathbf{x}^\circ \in \text{int}(\mathcal{C})$, and let $\Psi : \mathcal{B} \to \mathcal{C}$ be the gauge mapping. Consider problem $(\mathbf{P})$ with linear equality constraints $\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{p \times n}$ and $\mathbf{b} \in \mathbb{R}^p$. If the interior point satisfies $\mathbf{A}\mathbf{x}^\circ = \mathbf{b}$, then the transformed equality constraint $\boldsymbol{\phi}(\mathbf{z}) = \boldsymbol{\varphi}(\Psi(\mathbf{z}))$ in problem $(\mathbf{H})$ reduces to a linear constraint:*

$$\{\boldsymbol{\phi}(\mathbf{z}) = 0\} \Leftrightarrow \{\mathbf{A}\mathbf{z} = 0\}. \tag{4}$$

Proof. Recall that the gauge mapping $\Psi : \mathcal{B} \to \mathcal{C}$ is defined as

$$\Psi(\mathbf{z}) = \frac{\|\mathbf{z}\|}{\gamma_{\mathcal{C}}(\mathbf{z}, \mathbf{x}^\circ)}\mathbf{z} + \mathbf{x}^\circ, \quad \forall \mathbf{z} \in \mathcal{B}. \tag{5}$$

For the transformed equality constraint, we compute

$$\boldsymbol{\phi}(\mathbf{z}) = \boldsymbol{\varphi}(\Psi(\mathbf{z})) = \mathbf{A}\Psi(\mathbf{z}) - \mathbf{b} = \mathbf{A}\left(\frac{\|\mathbf{z}\|}{\gamma_{\mathcal{C}}(\mathbf{z}, \mathbf{x}^\circ)}\mathbf{z} + \mathbf{x}^\circ\right) - \mathbf{b} \tag{6}$$

$$= \frac{\|\mathbf{z}\|}{\gamma_{\mathcal{C}}(\mathbf{z}, \mathbf{x}^\circ)}\mathbf{A}\mathbf{z} + \mathbf{A}\mathbf{x}^\circ - \mathbf{b}. \tag{7}$$

Since the interior point satisfies $\mathbf{A}\mathbf{x}^\circ = \mathbf{b}$ by assumption, we have

$$\boldsymbol{\phi}(\mathbf{z}) = \frac{\|\mathbf{z}\|}{\gamma_{\mathcal{C}}(\mathbf{z}, \mathbf{x}^\circ)}\mathbf{A}\mathbf{z}. \tag{8}$$

Now, the feasibility condition $\boldsymbol{\phi}(\mathbf{z}) = \mathbf{0}$ requires

$$\frac{\|\mathbf{z}\|}{\gamma_{\mathcal{C}}(\mathbf{z}, \mathbf{x}^\circ)}\mathbf{A}\mathbf{z} = \mathbf{0}. \tag{9}$$

For $\mathbf{z} \neq \mathbf{0}$, since $\gamma_{\mathcal{C}}(\mathbf{z}, \mathbf{x}^\circ) > 0$ by the properties of the gauge function (Proposition B.4), the scalar coefficient $\frac{\|\mathbf{z}\|}{\gamma_{\mathcal{C}}(\mathbf{z}, \mathbf{x}^\circ)} > 0$. Therefore, $\boldsymbol{\phi}(\mathbf{z}) = \mathbf{0}$ holds if and only if $\mathbf{A}\mathbf{z} = \mathbf{0}$.

For $\mathbf{z} = \mathbf{0}$, we have $\Psi(\mathbf{0}) = \mathbf{x}^\circ$ and thus $\boldsymbol{\phi}(\mathbf{0}) = \mathbf{A}\mathbf{x}^\circ - \mathbf{b} = \mathbf{0}$, which is consistent with $\mathbf{A}\mathbf{0} = \mathbf{0}$.

Combining both cases, the transformed equality constraint is equivalent to the linear constraint $\mathbf{A}\mathbf{z} = \mathbf{0}$, which is independent of the nonlinear gauge function $\gamma_{\mathcal{C}}(\cdot, \mathbf{x}^\circ)$. □

Furthermore, the gauge function satisfies the following properties, where the proof can be found in Appendix B of (Liu et al., 2025).

Proposition B.4 (Basic properties of gauge function)**.** *Let $\mathcal{C}$ be a compact and convex set. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\alpha \geq 0$, gauge function $\gamma_{\mathcal{C}}(\cdot, \mathbf{x}^\circ)$ given an interior point $\mathbf{x}^\circ \in \operatorname{int}(\mathcal{C})$ satisfies:*

- *Non-negativity:* $\gamma_{\mathcal{C}}(\mathbf{x}, \mathbf{x}^\circ) \geq 0$.
- *Positive homogeneity:* $\gamma_{\mathcal{C}}(\alpha\mathbf{x}, \mathbf{x}^\circ) = \alpha\gamma_{\mathcal{C}}(\mathbf{x}, \mathbf{x}^\circ)$.
- *Subadditivity:* $\gamma_{\mathcal{C}}(\mathbf{x} + \mathbf{y}, \mathbf{x}^\circ) \leq \gamma_{\mathcal{C}}(\mathbf{x}, \mathbf{x}^\circ) + \gamma_{\mathcal{C}}(\mathbf{y}, \mathbf{x}^\circ)$.
- *Convexity. As a direct corollary, the gauge mapping Ψ defined in Def. B.2 is twice differentiable almost everywhere.*
- *Differentiability: Gauge function is twice differentiable almost everywhere.*
- *Upper/lower bounds:* $\gamma_{\mathcal{C}}(\mathbf{x}, \mathbf{x}^\circ) \in [\|\mathbf{x}\|/r_{\mathrm{o}}, \|\mathbf{x}\|/r_{\mathrm{i}}]$.
- *Lipschitz continuous:* $\|\gamma_{\mathcal{C}}(\mathbf{x}, \mathbf{x}^\circ) - \gamma_{\mathcal{C}}(\mathbf{y}, \mathbf{x}^\circ)\| \leq \frac{1}{r_{\mathrm{i}}}\|\mathbf{x} - \mathbf{y}\|$.

Based on the non-negativity, positive homogeneity, and subadditivity, the gauge function generalizes the concept of a norm. For a set $\mathcal{C}$ that is symmetric about the origin, the gauge function $\gamma_{\mathcal{C}}(\mathbf{x}, \mathbf{0})$ defines a norm. In particular, when $\mathcal{C} = \mathcal{B}_p = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_p \leq 1\}$ is the unit ball of the p-norm, we have $\gamma_{\mathcal{B}_p}(\mathbf{x}, \mathbf{0}) = \|\mathbf{x}\|_p$.

Remark. Note that the gauge mapping is twice differentiable almost everywhere. *However, the presence of a zero-measure set of non-differentiable points introduces a gap between our theoretical assumptions and the current formulation.* To address this, we adopt a smoothing approximation technique that renders the gauge mapping smooth everywhere, at the cost of a mild approximation error. See Appendix B in (Liu et al., 2025) for details.

Proposition B.5 (Bi-Lipschitz constants of the gauge mapping, (Liu et al., 2025))**.** *Let $\mathcal{C} \subset \mathbb{R}^n$ be a compact convex set and let $\mathbf{x}^\circ \in \operatorname{int}(\mathcal{C})$ be an interior point. Define the inner and outer radii with respect to $\mathbf{x}^\circ$ as*

$$r_{\mathrm{i}} := \sup\{r \geq 0 : \mathcal{B}(\mathbf{x}^\circ, r) \subseteq \mathcal{C}\}, \tag{10}$$

$$r_{\mathrm{o}} := \inf\{r \geq 0 : \mathcal{C} \subseteq \mathcal{B}(\mathbf{x}^\circ, r)\}, \tag{11}$$

such that $\mathcal{B}(\mathbf{x}^\circ, r_{\mathrm{i}}) \subseteq \mathcal{C} \subseteq \mathcal{B}(\mathbf{x}^\circ, r_{\mathrm{o}})$. *Then the Lipschitz constant (denoted as* $\mathrm{L}(\cdot)$*) of gauge mapping* Ψ *associated with* $\mathcal{C}$ *satisfies the following bounds:*

$$\textit{Forward Lipschitz:} \quad \kappa_2 := \mathrm{L}(\Psi) \leq 2\, r_{\mathrm{o}} + r_{\mathrm{o}}^2 / r_{\mathrm{i}}, \tag{12}$$

$$\textit{Inverse Lipschitz:} \quad \frac{1}{\kappa_1} := \mathrm{L}(\Psi^{-1}) \leq 2/r_{\mathrm{i}}. \tag{13}$$

As shown in Sec. 4, selecting a central interior point with a larger inner radius r_{i} or smaller outer radius r_{o} can reduce the Lipschitz constant of the gauge mapping Ψ, thereby improving the convergence speed of Hom-PALM.

B.2 Computation of Gauge Mapping

For computational purposes, we introduce a point-to-boundary distance function and its inverse.

Definition B.6 (Point-to-boundary distance (Tordesillas et al., 2023))**.** Let $\mathcal{C} \subset \mathbb{R}^n$ be a compact convex set and $\mathbf{x}^\circ \in \mathrm{int}(\mathcal{C})$. For any unit vector $\mathbf{v} \in \mathbb{S}^{n-1} = \{\mathbf{u} \in \mathbb{R}^n \mid \|\mathbf{u}\| = 1\}$, we define the interior-point-to-boundary distance function $d_{\mathcal{C}} : \mathrm{int}(\mathcal{C}) \times \mathbb{S}^{n-1} \to \mathbb{R}_{\geq 0}$ along direction $\mathbf{v}$ as

$$d_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{v}) = \sup\{\lambda \geq 0 \mid \mathbf{x}^\circ + \lambda \mathbf{v} \in \mathcal{C}\}.$$

The inverse distance function $\kappa_{\mathcal{C}} : \mathrm{int}(\mathcal{C}) \times \mathbb{S}^{n-1} \to \mathbb{R}_{\geq 0}$ is defined as $\kappa_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{v}) := 1/d_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{v})$.

We then have the following rules for computing the gauge mapping with the help of the point-to-boundary distance function.

- This distance function relates to the gauge function as:

$$\gamma_{\mathcal{C}}(\mathbf{x}, \mathbf{x}^\circ) = \kappa_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{x}/\|\mathbf{x}\|) \cdot \|\mathbf{x}\| = \frac{\|\mathbf{x}\|}{d_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{x}/\|\mathbf{x}\|)}.$$

- Further, considering the positive homogeneity, we have:

$$\gamma_{\mathcal{C}}(\mathbf{x}/\|\mathbf{x}\|, \mathbf{x}^\circ) = \kappa_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{x}/\|\mathbf{x}\|) = \frac{1}{d_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{x}/\|\mathbf{x}\|)}.$$

- The gauge mapping between $\mathcal{B}$ and $\mathcal{C}$ can be simplified as:

$$\Psi(\mathbf{z}) = d_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{z}/\|\mathbf{z}\|) \cdot \mathbf{z} + \mathbf{x}^\circ, \ \forall \mathbf{z} \in \mathcal{B}$$

$$\Psi^{-1}(\mathbf{x}) = \frac{\mathbf{x} - \mathbf{x}^\circ}{d_{\mathcal{C}}(\mathbf{x}^\circ, (\mathbf{x} - \mathbf{x}^\circ)/\|\mathbf{x} - \mathbf{x}^\circ\|)}, \ \forall \mathbf{x} \in \mathcal{C}.$$

Therefore, we can compute the inverse distance function to obtain the gauge mapping. In practice, we can compute the inverse distance function $\kappa_{\mathcal{C}}(\mathbf{x}^\circ, \cdot)$ as follows.

- *Closed-form for common constrained set.* For various common constraint types, Table 3 provides closed-form expressions of the inverse distance function. Most matrix calculations can be computed and stored offline before being applied online given $\mathbf{v}$.
- *Bisection-based algorithm for general constrained set* . When the inverse distance function lacks a closed-form expression, we employ an efficient bisection algorithm detailed in Alg. 3. This algorithm supports batch processing, enabling efficient parallel computation for multiple inputs simultaneously.

B.3 Complexity Analysis

In this section, we provide the omitted details in Sec. 4.3.

Oracles. We list specific oracles used besides general ones (e.g., zeroth-order oracle for explicit function evaluation).

Table 3: Closed-form Expressions for Inverse Distance Functions (Tordesillas et al., 2023)

Constrained Set $\mathcal{C}$	Formulation	Inverse Distance Function $\kappa_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{v})$
Linear	$\{\mathbf{x} : \mathbf{a}^\top \mathbf{x} \leq b\}$	$(\frac{\mathbf{a}^\top \mathbf{v}}{b - \mathbf{a}^\top \mathbf{x}^\circ})^+$
Quadratic	$\{\mathbf{x} : \mathbf{x}^\top \mathbf{Q} \mathbf{x} + \mathbf{a}^\top \mathbf{x} \leq b\}$	$(1/\text{root}\,(A_{\mathbf{Q}}, B_{\mathbf{Q}}, C_{\mathbf{Q}}))^+$
Second Order Cone	$\{\mathbf{x} : \|\mathbf{A}^\top \mathbf{x} + \mathbf{p}\|_2 \leq \mathbf{a}^\top \mathbf{x} + b\}$	$(1/\text{root}(A_{\text{S}}, B_{\text{S}}, C_{\text{S}}))^+$
Linear Matrix Inequality	$\{\mathbf{x} : \sum_{i=1}^n x_i \cdot \mathbf{F}_i + \mathbf{F}_0 \succeq 0\}$	$\left(\text{eig}\left(\mathbf{L}^\top (-\mathbf{S})\, \mathbf{L}\right)\right)^+$
Intersections	$\mathcal{C} = \bigcap_{i=1}^m \mathcal{C}_i$	$\max_{1 \leq i \leq m} \{\kappa_{\mathcal{C}_i}(\mathbf{x}^\circ, \mathbf{v})\}$

[1] Notation: $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$, $b \in \mathbb{R}$, $\mathbf{Q} \in \mathbb{S}_+^n$, $\mathbf{A} \in \mathbb{R}^{n \times m}$, $\mathbf{p} \in \mathbb{R}^m$, $\mathbf{F}_0, \cdots, \mathbf{F}_n \in \mathbb{R}^{m \times m}$, $\mathbf{X} \in \mathbb{R}^{n \times n}$, $(\cdot)^+ = \max(\cdot, 0)$.

[2] $\text{root}(x_1, x_2, x_3) = \frac{-x_2 \pm \sqrt{x_2^2 - 4x_1 x_3}}{2x_1}$ denotes the roots for quadratic equation.

[3] $A_{\mathbf{Q}} = \mathbf{v}^\top \mathbf{Q} \mathbf{v}$, $B_{\mathbf{Q}} = 2\mathbf{x}^{\circ\top} \mathbf{Q} \mathbf{v} + \mathbf{a}^\top \mathbf{v}$, $C_{\mathbf{Q}} = \mathbf{x}^{\circ\top} \mathbf{Q} \mathbf{x}^\circ + \mathbf{a}^\top \mathbf{x}^\circ - b$.

[4] $A_{\text{S}} = (\mathbf{A}^\top \mathbf{v})^\top (\mathbf{A}^\top \mathbf{v}) - (\mathbf{a}^\top \mathbf{v})^2$, $B_{\text{S}} = 2(\mathbf{A}^\top \mathbf{x}^\circ + \mathbf{p})^\top (\mathbf{A}^\top \mathbf{v}) - 2(\mathbf{a}^\top \mathbf{x}^\circ + b)(\mathbf{a}^\top \mathbf{v})$, $C_{\text{S}} = (\mathbf{A}^\top \mathbf{x}^\circ + \mathbf{p})^\top (\mathbf{A}^\top \mathbf{x}^\circ + \mathbf{p}) - (\mathbf{a}^\top \mathbf{x}^\circ + b)^2$.

[5] $\mathbf{H} = \mathbf{F}_0 + \sum_{i=1}^n \mathbf{x}_i^\circ \mathbf{F}_i$, $\mathbf{H}^{-1} = \mathbf{L}^\top \mathbf{L}$, $\mathbf{S} = \sum_{i=1}^n v_i \mathbf{F}_i$. $\text{eig}(\mathbf{X}) = \lambda_1, \cdots, \lambda_n$ denotes the eigenvalues satisfying $\det(\mathbf{X} - \lambda \mathbf{I}) = 0$. Note that only the maximum eigenvalue is needed. Thus, power iteration methods can be applied to compute it efficiently.

[6] Note that all $\mathbf{v}$-independent terms can be computed only once and stored for use.

Algorithm 3 Bisection Algorithm for Inverse Point-to-Boundary Distance

Input: A compact convex set $\mathcal{C}$, an interior point $\mathbf{x}^\circ \in \mathcal{C}$, a unit vector $\mathbf{v}$, and a tolerance $\epsilon > 0$.

$\alpha_l = 0$, and $\alpha_u = 1$
while $\mathbf{x}^\circ + \alpha_u \cdot \mathbf{v} \in \mathcal{C}$ **do**
  increase lower bound: $\alpha_l \leftarrow \alpha_u$
  double upper bound: $\alpha_u \leftarrow 2 \cdot \alpha_u$
end while
while $|\alpha_l - \alpha_u| \geq \epsilon$ **do**
  bisection: $\alpha_m = (\alpha_l + \alpha_u)/2$
  if $\mathbf{x}^\circ + \alpha_m \cdot \mathbf{v} \in \mathcal{C}$ **then**
    increase lower bound: $\alpha_l \leftarrow \alpha_m$
  else
    decrease upper bound: $\alpha_u \leftarrow \alpha_m$
  end if
end while

Output: inverse distance: $\kappa_{\mathcal{C}}(\mathbf{x}^\circ, \mathbf{v}) \approx 1/\alpha_m$

- *Membership oracle*: Given $\mathbf{x} \in \mathbb{R}^n$, the oracle $\mathcal{M}_{\mathcal{C}}(\mathbf{x}) := \mathbb{I}(\mathbf{x} \in \mathcal{C}) : \mathbb{R}^n \to \{0, 1\}$ returns 1 if and only if $\mathbf{x} \in \mathcal{C}$. It checks feasibility without solving optimization subproblems. For common convex sets, such as polyhedra and second-order cones, it can be implemented with complexity at most $\mathcal{O}(n^2)$ (Mhammedi, 2022), and is typically much cheaper than LMO or projection in practice.

- *Interior point oracle*: This oracle returns an interior point of $\mathcal{C}$ by solving a convex feasibility problem, a common requirement in advanced projection-free optimization methods (Mhammedi, 2022; Grimmer, 2024a;b). It can be addressed using first-order methods with $\tilde{\mathcal{O}}(n^2)$ complexity or interior-point methods with $\tilde{\mathcal{O}}(n^{3.5})$. Notably, the oracle is called only **once** throughout the entire algorithm.

Computation of basic operations in Hom-PALM. We now summarize the practical methods used to compute the key operations in Hom-PALM, including the gauge function/mapping and the gradient of $\star \circ \Psi$ where '$\star$' represents f or the AL function $\hat{f}$. The theorem statements in the main text treat these derivatives as available; the numerical procedures below describe the practical implementation fallback.

- *Computing the gauge function* $\gamma_{\mathcal{C}}(\cdot, \mathbf{x}^\circ)$: Compute the inverse distance function $\kappa_{\mathcal{C}}(\mathbf{x}^\circ, \cdot)$ via Algorithm 3 or using the closed form when available. The bisected algorithm 3 will cost $\mathcal{O}(\log 1/\epsilon)$ membership oracles to an accuracy ϵ.
- *Computing the gauge mapping* Ψ: Derived from Definition B.2, given $\gamma_{\mathcal{C}}(\cdot, \mathbf{x}^\circ)$. Thus, for computation of gauge mapping, it involves evaluating the gauge function $\gamma_{\mathcal{C}}(\cdot, \mathbf{x}^\circ)$, plus $\mathcal{O}(n)$ operations for the scalar-vector product (see Def. 3.2), leading to a total cost of $\tilde{\mathcal{O}}(n)$.
- *Computing the gradient of* $h = f \circ \Psi$: Approximated using numerical differentiation methods. If h has a closed-form, the gradient can be computed using automatic differentiation methods (see e.g., (Burden & Faires, 1997; Baydin et al., 2018; Lu et al., 2023)). For general cases, we adopt the finite difference method for computing the gradient of the gauge function in (Lu et al., 2023), i.e., for $i = 1, 2, \ldots, n$:

  $$\nabla_i h(\mathbf{x}) = \frac{h(\mathbf{x} + \lambda \mathbf{e}_i) - h(\mathbf{x})}{\lambda}$$

  given a proper small number $\lambda > 0$ where $\mathbf{e}_i$ denotes the standard basis over $\mathbb{R}^n$. Each component $\nabla_i h$ requires $\mathcal{O}(1)$ zeroth-order oracle calls of f and $\tilde{\mathcal{O}}(n)$ cost for evaluating Ψ, resulting in total complexity $\tilde{\mathcal{O}}(n^2)$.

  Similarly, computing the gradient of $\hat{h} = \hat{f} \circ \Psi$ needs $\tilde{\mathcal{O}}(n^2)$ time.

C Preliminaries for Technical Proof

In this section, we summarize the related basic concepts, notations, assumptions, and fundamental propositions and lemmas.

C.1 Basic Concepts

We list the basic concepts used in this paper below.

- Indicator function. For a set $\mathcal{X} \subset \mathbb{R}^n$, the indicator function $\mathbb{I}_{\mathcal{X}} : \mathbb{R}^n \to \mathbb{R}$ is defined as

  $$\mathbb{I}_{\mathcal{X}}(\mathbf{x}) := \begin{cases} 0 & \text{if } \mathbf{x} \in \mathcal{X}, \\ \infty & \text{if } \mathbf{x} \notin \mathcal{X}. \end{cases}$$

- Distance between a point and a set. For a closed set $\mathcal{X} \in \mathbb{R}^n$ and any $\mathbf{x} \in \mathbb{R}^n$, the distance between $\mathbf{x}$ and $\mathcal{X}$ is defined as $\operatorname{dist}(\mathbf{x}, \mathcal{X}) = \inf_{\mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$.
- Orthogonal projection. For a convex and closed set $\mathcal{X}$, the orthogonal projection of a point $\mathbf{x} \in \mathbb{R}^n$ onto $\mathcal{X}$ is defined as $\Pi_{\mathcal{X}}(\mathbf{x}) = \arg\min_{\mathbf{y} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}\|$.
- Function convexity. For a differentiable function $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$, it is said to be convex if one of the following holds:
  1) Jensen's inequality. For θ with $0 \le \theta \le 1$, we have $f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \le \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.
  2) first-order condition. $f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$.
  3) monotone gradient. $(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^T (\mathbf{x} - \mathbf{y}) \ge 0$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.
- L-Smoothness. A differentiable function $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is said be L-smooth if one of the following holds:
  1) zeroth-order condition. $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \ge \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{L}{2} \lambda (1-\lambda) \|\mathbf{y} - \mathbf{x}\|^2$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}, \lambda \in [0, 1]$.
  2) first-order condition. $f(\mathbf{y}) \le f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.
  3) Lipschitz gradient. $\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| \le L \|\mathbf{y} - \mathbf{x}\|$, for all $\mathbf{x}, \mathbf{y}$.
- μ-Strong convexity. A differentiable function $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ is said be μ-strongly convex holds if one of the following holds:
  1) zeroth-order condition. $f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \le \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) - \frac{\mu}{2} \lambda (1-\lambda) \|\mathbf{y} - \mathbf{x}\|^2$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}, \lambda \in [0, 1]$.

2) first-order condition. $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.
3) strictly monotone gradient. $\langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geq \mu\|\mathbf{y} - \mathbf{x}\|^2$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

- Weak convexity. A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be weakly convex with constant $\ell_f > 0$ if the function $f(\mathbf{x}) + (\ell_f/2)\|\mathbf{x}\|^2$ is convex.
- Jacobian matrix. Suppose $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is a function such that each of its first-order partial derivatives exists on $\mathbb{R}^n$. Then the Jacobian matrix of $\mathbf{f}$, denoted $\mathrm{J}_{\mathbf{f}} \in \mathbb{R}^{n \times m}$, is defined as $\mathrm{J}_{\mathbf{f}} = [\nabla f_1, \cdots, \nabla f_m]$ where $\nabla f_i(\mathbf{x}) \in \mathbb{R}^n$ for fixed x.
- A Hessian of a function $f : \mathbb{R}^n \to \mathbb{R}$ is defined as $\nabla^2 f = (\frac{\partial^2 f}{\partial x_i \partial x_j})_{ij} \in \mathbb{R}^{n \times n}$, if its second-order partial derivatives exist. Moreover, for a mapping $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ with existed second-order partial derivatives of each component f_i $(i = 1, 2, \cdots, m)$. The Hessian of $\mathbf{f}$ is defined as
$$\mathrm{H}(\mathbf{f}) = (\nabla^2 f_1, \cdots, \nabla^2 f_m).$$
- The normal cone $\mathrm{N}_{\mathcal{K}}(\mathbf{x})$ of a closed and convex set $\mathcal{K}$ at $\mathbf{x} \in \mathcal{K}$ is defined as
$$\mathrm{N}_{\mathcal{K}}(\mathbf{x}) = \{\mathbf{y} : \langle \mathbf{y}, \mathbf{z} - \mathbf{x} \rangle \leq 0 \text{ for any } \mathbf{z} \in \mathcal{K}\}.$$

C.2 Basic Assumptions and Notations

In the following, we make assumptions throughout the paper.

- Assumptions on f and constrained functions g_i $(i \in [m])$ and ϕ_j $(j \in [p])$ in problem $\mathbf{P}$:
  1) f is $L_{f,0}$-Lipschitz continuous, i.e., $\|f(\mathbf{x}) - f(\mathbf{y})\| \leq L_{f,0}\|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y}$.
  2) f in problem $\mathbf{P}$ is differentiable and L_f-smooth.
  3) $f^* > -\infty$ where $f^* := \min_{\mathbf{x} \in \mathcal{K}} f(\mathbf{x})$.
  4) Each g_i $(i \in [m])$ is $L_{g_i,0}$-Lipschitz continuous, differentiable, and L_{g_i}-smooth.
  5) Each ϕ_i $(i \in [p])$ is $L_{\phi_i,0}$-Lipschitz continuous, differentiable, and L_{ϕ_i}-smooth.
- Assumptions on the homeomorphic mapping $\Psi : \mathbb{R}^n \to \mathbb{R}^n$:
  1) Ψ is differentiable with non-singular Jacobian $\mathrm{J}_{\Psi}(\cdot)$,
  2) Ψ is (κ_1, κ_2)-bi-Lipschitz continuous for $\kappa_2 \geq \kappa_1 > 0$, i.e.,
  $$\kappa_1\|\mathbf{u} - \mathbf{v}\| \leq \|\Psi(\mathbf{u}) - \Psi(\mathbf{v})\| \leq \kappa_2\|\mathbf{u} - \mathbf{v}\|.$$
  Then the Jacobian matrix, $\mathrm{J}_{\Psi}(\cdot)$ and $\mathrm{J}_{\Psi^{-1}}(\cdot)$ will satisfy
  $$\|\mathrm{J}_{\Psi}(\mathbf{z})\| \leq \kappa_2, \quad \forall \mathbf{z}, \quad \|\mathrm{J}_{\Psi^{-1}}(\mathbf{x})\| \leq \frac{1}{\kappa_1}, \ \forall \mathbf{x}.$$
  3) Ψ has L_{Ψ}-Lipschitz continuous Jacobian matrix, i.e.,
  $$\|\mathrm{J}_{\Psi}(\mathbf{u}) - \mathrm{J}_{\Psi}(\mathbf{v})\| \leq L_{\Psi}\|\mathbf{u} - \mathbf{v}\|, \ \forall \mathbf{u}, \mathbf{v}.$$
  4) Ψ has continuous Hessian, i.e.,
  $$\mathrm{H}_{\Psi}(\mathbf{z}) = (\nabla^2 \Psi_1, \cdots, \nabla^2 \Psi_n)$$
  exists and is continuous.

In addition, we summarize the commonly used notations in this paper in Table 4.

C.3 Basic Facts

In this section, we list the fundamental facts we will use in this paper.

Proposition C.1 (Global optimality condition, see e.g., (Boyd et al., 2004; Beck, 2017)). *For convex constrained optimization* $\{\min_{\mathbf{x}} f(\mathbf{x}), \ s.t. \ \mathbf{x} \in \mathcal{X}\}$, $\mathbf{x}^*$ *is a global optimum if and only if* $\mathbf{x}^*$ *is a stationary point of the problem, i.e.,*
$$\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geq 0, \forall \mathbf{x} \in \mathcal{X}.$$

Table 4: Summary of Notations. The notations shown in the table are for problem **P**, and we use the same type of notations for problem **H**.

Notation	Definition
$L_{f,0}$	Lipschitz constant of f
L_f	L_f-smooth property of f
μ_f	μ_f-strong convexity of f
κ_1, κ_2	bi-Lipschitz constant of Ψ
D	distortion of Ψ, i.e., κ_2/κ_1
L_Ψ	Lipschitz constant of J_Ψ
$\hat{h}(\mathbf{z})$	$\hat{h}_\rho(\mathbf{z};\boldsymbol{\gamma})$
$\hat{f}(\mathbf{x})$	$\hat{f}_\rho(\mathbf{x};\boldsymbol{\eta})$
$\mathrm{int}(\mathcal{K})$,$\partial\mathcal{K}$	the interior,boundary of $\mathcal{K}$

Next, we list the following rules for basic computation:

- Jacobian equivalence: $\mathrm{J}_{\Psi^{-1}}(\mathbf{x}) = \mathrm{J}_\Psi^{-1}(\mathbf{z})$ for $\mathbf{x} = \Psi(\mathbf{z})$.
- Chain rule for computing the gradient of $h = f \circ \Psi$:

$$\nabla h(\mathbf{z}) = \mathrm{J}_\Psi(\mathbf{z})\nabla f(\Psi(\mathbf{z})) = \mathrm{J}_\Psi(\mathbf{z})\nabla f(\mathbf{x}).$$

- Chain rule for computing the gradient of f:

$$\nabla f(\mathbf{x}) = \mathrm{J}_{\Psi^{-1}}(\mathbf{x})\nabla h(\mathbf{z}) = \mathrm{J}_\Psi^{-1}(\mathbf{z})\nabla h(\mathbf{z}).$$

- Chain rule for computing Jacobian of $\boldsymbol{\phi}(\mathbf{z}) = \boldsymbol{\varphi}(\Psi(\mathbf{z}))$:

$$\mathrm{J}_{\boldsymbol{\phi}}(\mathbf{z}) = \mathrm{J}_\Psi(\mathbf{z})\mathrm{J}_{\boldsymbol{\varphi}}(\mathbf{x}).$$

- Chain rule for computing Hessian of $h = f \circ \Psi$:

$$\nabla^2 h(\mathbf{z}) = \mathrm{J}_\Psi(\mathbf{z})\nabla^2 f(\Psi(\mathbf{z}))\mathrm{J}_\Psi(\mathbf{z})^\top + \sum_{i=1}^n \frac{\partial f}{\partial \mathbf{x}_i}(\Psi(\mathbf{z}))\nabla^2 \Psi_i(\mathbf{z}).$$

D Landscape Analysis

In this section, we provide landscape analysis to understand important relationships between problem (**P**) and (**H**).

D.1 Action of Homeomorphism on a Constrained Set

Recall that the constrained set is $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{g}(\mathbf{x}) \leq 0\}$ with $\mathbf{g} = (g_1, g_2, \cdots, g_m)$ where each g_i $(i = 1, 2, \cdots, m)$ is a continuous function. For problem (**H**),

$$\mathcal{B} = \Psi^{-1}(\mathcal{C}) = \{\mathbf{z} \in \mathbb{R}^n \mid \Psi(\mathbf{z}) \in \mathcal{C}\} = \{\mathbf{z} \in \mathbb{R}^n \mid \mathbf{G}(\mathbf{z}) := \mathbf{g}(\Psi(\mathbf{z})) \leq \mathbf{0}\}$$

where G_i is non-convex in general. However, $\mathcal{B}$ is assumed to be convex (actually a ball set) in this paper. One can refer to Fig. 4 for an illustration.

Moreover, we assume there are no redundant inequalities in $\mathcal{C}$, i.e., there is no g_i such that $\mathcal{C} = \{\mathbf{x} \mid \mathbf{g}_{-i}(\mathbf{x}) \leq \mathbf{0}\}$ where $\mathbf{g}_{-i} = (g_1, \cdots, g_{i-1}, g_{i+1}, \cdots, g_m)$. In this case, any feasible point $\mathbf{x}$ satisfying $g_i(\mathbf{x}) = 0$ for some i is on the boundary of the set $\mathcal{C}$. Thus, we have

$$\{\mathbf{x} \in \mathcal{C} \mid g_j(\mathbf{x}) = 0, g_k(\mathbf{x}) \neq 0\} \bigcap \{\mathbf{x} \in \mathcal{C} \mid g_k(\mathbf{x}) = 0, g_j(\mathbf{x}) \neq 0\} = \emptyset$$

for any $k \neq j$. Note $\mathcal{B} = \{\mathbf{z} \mid G_i(\mathbf{z}) \leq 0, i = 1, 2, \cdots, m\} = \{\mathbf{z} \mid \|\mathbf{z}\|^2 \leq 1\}$. Moreover, $\{G_i(\mathbf{z}) \leq 0, i = 1, 2, \cdots, m\}$ also has no redundant constraints by the non-singularity of the Jacobian of Ψ and similarly,

$$\{\mathbf{z} \in \mathcal{B} \mid G_j(\mathbf{z}) = 0, G_k(\mathbf{z}) \neq 0\} \bigcap \{\mathbf{z} \in \mathcal{B} \mid G_k(\mathbf{z}) = 0\} = \emptyset$$

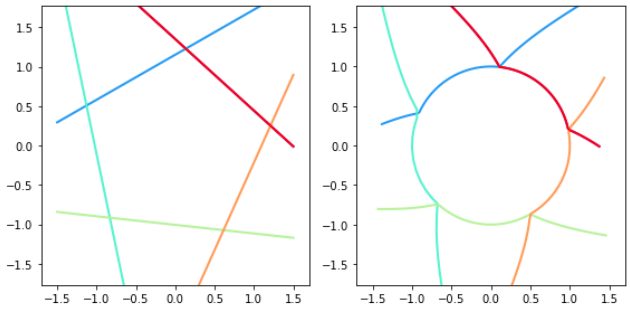


Figure 4: Illustration of the action of homeomorphism on a star-shaped set. The left figure shows the star-shaped constraints of problem (**P**). Each color of line represents the boundary characterized by a constraint inequality $\{\mathbf{a}_i^\top \mathbf{x} \le b_i\}$ for some i. Under a homeomorphic mapping Ψ, the constrained set is transformed to a ball (right figure). Each constraint inequality $\{G_i(\mathbf{z}) \le 0\}$ (colored differently) is non-convex in general.

for any $j \neq k$. Hence if $\mathbf{z} \in \mathcal{B}$ satisfies $G_i(\mathbf{z}) = 0$ for some i, it lies on the boundary of $\mathcal{B}$. Clearly, we have

$$G_i(\mathbf{z}) = \|\mathbf{z}\|^2 - 1 \quad \text{at} \quad \mathbf{z}' \in \partial\mathcal{B}, G_i(\mathbf{z}') = 0, \tag{14}$$

and

$$\nabla G_i(\mathbf{z}) = 2\mathbf{z}, \nabla^2 G_i(\mathbf{z}) = 2\mathbf{I}_n \quad \text{at} \quad \mathbf{z}' \in \partial\mathcal{B}, G_i(\mathbf{z}') = 0. \tag{15}$$

where $\mathbf{I}_n$ is the identity matrix of n by n.

D.2 Properties of AL Functions

We begin with the basic regularity of the augmented-Lagrangian objective under a homeomorphic change of variables. For simplicity, we write $\hat{h}(\mathbf{z}) := \hat{h}_\rho(\mathbf{z}; \boldsymbol{\gamma})$ and $\hat{f}(\mathbf{x}) := \hat{f}_\rho(\mathbf{x}; \boldsymbol{\eta})$, where

$$\hat{h}_\rho(\mathbf{z}; \boldsymbol{\gamma}) = h(\mathbf{z}) + \langle \boldsymbol{\gamma}, \boldsymbol{\phi}(\mathbf{z}) \rangle + \frac{\rho}{2}\|\boldsymbol{\phi}(\mathbf{z})\|^2, \quad \hat{f}_\rho(\mathbf{x}; \boldsymbol{\eta}) = f(\mathbf{x}) + \langle \boldsymbol{\eta}, \boldsymbol{\varphi}(\mathbf{x}) \rangle + \frac{\rho}{2}\|\boldsymbol{\varphi}(\mathbf{x})\|^2 \tag{16}$$

are the standard augmented Lagrangian for the original problem (**H**) and (**P**), respectively.

The composite function $h = f \circ \Psi$ enjoys the following properties, where the proof can be found in (Liu et al., 2025).

Lemma D.1 (Properties of $h = f \circ \Psi$)**.** *Under the general assumptions C.2, $h = f \circ \Psi$ has the following properties.*

(i) h is $L_{h,0} := L_{f,0}\kappa_2$ Lipschitz continuous.

(ii) h is L_h-smooth with $L_h = \kappa_2^2 L_f + L_\Psi L_{f,0}$.

(iii) If f is convex, then h is ℓ_h-weakly convex with $\ell_h = L_{f,0}L_\Psi$.

(iv) If f is μ_f-strongly convex, then h satisfies the quadratic growth condition with $m_h = \frac{\mu_f \kappa_1^2}{2}$ on the ball set $\mathcal{B}$, namely

$$h(\mathbf{z}) - h^* \ge m_h \|\mathbf{z} - \mathbf{z}^*\|^2, \ \forall \mathbf{z} \in \mathcal{B}. \tag{17}$$

Lemma D.2. *Suppose $\boldsymbol{\varphi}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^p$ is linear, i,e., $\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$. Then $\hat{f}(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\eta}^\top \boldsymbol{\varphi}(\mathbf{x}) + \frac{\rho}{2}\|\boldsymbol{\varphi}(\mathbf{x})\|^2$ with a vector $\boldsymbol{\eta} \in \mathbb{R}^p$ has the following properties. (i) If f is L_f-smooth, then $\hat{f}$ is $L_{\hat{f}}$-smooth with*

$$L_{\hat{f}} = L_f + \rho\|\mathbf{A}\|^2$$

where $\|\mathbf{A}\|$ denotes the largest eigenvalue of $\mathbf{A}$.

(ii) $\hat{f}$ is $L_{\hat{f},0}$-Lipschitz continuous with $L_{\hat{f},0} := L_{f,0} + (\|\boldsymbol{\eta}\| + \rho\mathfrak{m}) \cdot \|\mathbf{A}\|$ over any bounded set $\mathcal{M}$ where $\mathfrak{m} := \max_{\mathbf{x}\in\mathcal{M}}\{\mathbf{A}\mathbf{x}\}$.

(iii) If f is μ_f-strongly convex, then $\hat{f}$ is $\mu_f + \rho\lambda_{\min}(\mathbf{A}^\top\mathbf{A})$-strongly convex, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue. Moreover, the proximal AL function $\mathbb{L}(\hat{f}) := \hat{f} + \frac{\alpha}{2}\|\cdot - \mathbf{x}_k\|^2$ is $\big(\mu_f + \rho\lambda_{\min}(\mathbf{A}^\top\mathbf{A}) + \alpha\big)$-strongly convex.

Proof. We show them one by one in the following.

1) We can directly derive

$$\nabla\hat{f}(\mathbf{x}) = \nabla f(\mathbf{x}) + \mathbf{A}^\top\boldsymbol{\eta} + \rho\mathbf{A}^\top\boldsymbol{\varphi}(\mathbf{x}).$$

Then

$$\|\nabla\hat{f}(\mathbf{x}) - \nabla\hat{f}(\mathbf{y})\| \le \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| + \rho\|\mathbf{A}^\top(\boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}(\mathbf{y}))\| \le (L_f + \rho\|\mathbf{A}\|^2)\|\mathbf{x} - \mathbf{y}\|.$$

2) To show $\hat{f}$ is Lipschitz continuous over $\mathcal{M}$, we first note that x^2 is $2l$-Lipschitz continuous over any interval $[-l, l]$. Then over the bounded region $\mathcal{M}$, we have

$$\begin{aligned}|\hat{f}(\mathbf{x}) - \hat{f}(\mathbf{y})| \le& |f(\mathbf{x}) - f(\mathbf{y})| + \|\boldsymbol{\eta}\| \cdot \|\boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}(\mathbf{y})\| + \frac{\rho}{2} \cdot 2\mathfrak{m}\|\boldsymbol{\varphi}(\mathbf{x}) - \boldsymbol{\varphi}(\mathbf{y})\| \\ \le& L_{f,0}\|\mathbf{x} - \mathbf{y}\| + (\|\boldsymbol{\eta}\| + \mathfrak{m}\rho)\|\mathbf{A}\| \cdot \|\mathbf{x} - \mathbf{y}\|.\end{aligned}$$

3) One can directly compute

$$\nabla^2\hat{f}(\mathbf{x}) = \nabla^2 f(\mathbf{x}) + \rho\mathbf{A}^\top\mathbf{A} \succeq \big(\mu_f + \rho\lambda_{\min}(\mathbf{A}^\top\mathbf{A})\big)\,\mathbf{I}.$$

Moreover, Similar computation holds for $\mathbb{L}(\hat{f})$.

□

Lemma D.3. *If f is L_f-smooth and $\boldsymbol{\varphi}(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^p$ is linear, i,e., $\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, then $\hat{h}(\mathbf{z}) = h(\mathbf{z}) + \boldsymbol{\gamma}^\top\boldsymbol{\phi}(\mathbf{z}) + \frac{\rho}{2}\|\boldsymbol{\phi}(\mathbf{z})\|^2$ with a vector $\boldsymbol{\eta} \in \mathbb{R}^p$ is $L_{\hat{h}}$-smooth over any bounded region $\mathcal{M}$ with*

$$L_{\hat{h}} := L_{\hat{f}}\kappa_2^2 + L_{\hat{f},0}L_\Psi$$

where $L_{\hat{f},0}$ is defined in Lemma D.2.

Proof. We can directly derive

$$\begin{aligned}\nabla\hat{h}(\mathbf{z}) =& \nabla h(\mathbf{z}) + \mathrm{J}_{\boldsymbol{\phi}}(\mathbf{z})\boldsymbol{\gamma} + \rho\mathrm{J}_{\boldsymbol{\phi}}(\mathbf{z})\boldsymbol{\phi}(\mathbf{z}) \\ =& \nabla h(\mathbf{z}) + \mathrm{J}_\Psi(\mathbf{z})\mathrm{J}_{\boldsymbol{\varphi}}(\mathbf{x})\boldsymbol{\gamma} + \rho\mathrm{J}_\Psi(\mathbf{z})\mathrm{J}_{\boldsymbol{\varphi}}(\mathbf{x})\boldsymbol{\phi}(\mathbf{z}) \\ =& \mathrm{J}_\Psi(\mathbf{z})\nabla\hat{f}(\mathbf{x})\end{aligned}$$

where $\mathbf{x} = \Psi(\mathbf{z})$. Let $\mathbf{y} = \Psi(\mathbf{u})$. We can derive

$$\begin{aligned}\|\nabla\hat{h}(\mathbf{z}) - \nabla\hat{h}(\mathbf{u})\| =& \Big\|\mathrm{J}_\Psi(\mathbf{z})\nabla\hat{f}(\mathbf{x}) - \mathrm{J}_\Psi(\mathbf{u})\nabla\hat{f}(\mathbf{y})\Big\| \\ =& \Big\|\mathrm{J}_\Psi(\mathbf{z})\nabla\hat{f}(\mathbf{x}) - \mathrm{J}_\Psi(\mathbf{u})\nabla\hat{f}(\mathbf{y}) + \mathrm{J}_\Psi(\mathbf{u})\nabla\hat{f}(\mathbf{y}) - \mathrm{J}_\Psi(\mathbf{u})\nabla\hat{f}(\mathbf{y})\Big\| \\ \le& \|J_\Psi(\mathbf{z})\|\,\|\nabla\hat{f}(\mathbf{x}) - \nabla\hat{f}(\mathbf{y})\| + \|\mathrm{J}_\Psi(\mathbf{z}) - \mathrm{J}_\Psi(\mathbf{u})\|\,\|\nabla\hat{f}(\mathbf{y})\| \\ \le& \kappa_2 L_{\hat{f}}\|\mathbf{x} - \mathbf{y}\| + L_{\hat{f},0}L_\Psi\|\mathbf{z} - \mathbf{u}\| \\ \le& L_{\hat{f}}\kappa_2^2\|\mathbf{z} - \mathbf{u}\| + L_{\hat{f},0}L_\Psi\|\mathbf{z} - \mathbf{u}\|.\end{aligned}$$

□

D.3 Global Optimality Property of Optimization Problem (**H**)

In this section, we will show the global optimality condition of problem (**H**). Before moving on, we introduce some basic definitions.

Definition D.4 (KKT stationary point)**.** A point $\mathbf{x}^* \in \mathbb{R}^n$ is a KKT stationary point of problem (**P**) if there exist multipliers $(\boldsymbol{\lambda}^*, \boldsymbol{\eta}^*) \in \mathbb{R}^m \times \mathbb{R}^p$ such that

$$\begin{aligned} &\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \eta_j^* \nabla \varphi_j(\mathbf{x}^*) = \mathbf{0}, \\ &g_i(\mathbf{x}^*) \le 0, \quad \lambda_i^* \ge 0, \quad \lambda_i^* g_i(\mathbf{x}^*) = 0, \quad \forall\, i \in [m], \\ &\varphi_j(\mathbf{x}^*) = 0, \quad \forall\, j \in [p], \end{aligned} \tag{18}$$

where $\boldsymbol{\lambda}^*$ and $\boldsymbol{\eta}^*$ are the multipliers associated with the inequality and equality constraints, respectively.

Definition D.5 (Constraints qualification)**.** Let $\bar{\mathbf{x}}$ be feasible for a problem with constraint set

$$\{\mathbf{x} \mid g_i(\mathbf{x}) \le 0, \varphi_j(\mathbf{x}) = 0, \text{ for } i = 1, 2, \cdots, m, \text{ and } j = 1, 2, \cdots, p\}$$

and put $\mathcal{A}(\bar{\mathbf{x}}) := \{i \mid g_i(\bar{\mathbf{x}}) = 0\}$. We say that

1) the linear independence constraint qualification (LICQ) holds at $\bar{\mathbf{x}}$ (and write LICQ($\bar{\mathbf{x}}$)) if the gradients

$$\nabla g_i(\bar{\mathbf{x}})(i \in \mathcal{A}(\bar{\mathbf{x}})), \quad \nabla \varphi_j(\bar{\mathbf{x}})(j = 1, \ldots, p)$$

are linearly independent.

2) the Mangasarian-Fromovitz constraint qualification (MFCQ) holds at $\bar{\mathbf{x}}$ (and write MFCQ($\bar{\mathbf{x}}$)) if the gradients

$$\nabla \varphi_j(\bar{\mathbf{x}})(j = 1, \ldots, p)$$

are linearly independent and there exists a vector $\mathbf{d} \in \mathbb{R}^n$ such that

$$\nabla g_i(\bar{\mathbf{x}})^\top \mathbf{d} < 0 (i \in \mathcal{A}(\bar{\mathbf{x}})), \quad \nabla \varphi_j(\bar{\mathbf{x}})^\top \mathbf{d} = 0 (j = 1, \ldots, p).$$

Definition D.6 (Strict complementary slackness)**.** It is said that the strict complementary slackness condition holds for problem (**P**), if for any KKT stationary point $\mathbf{x}^*, \boldsymbol{\lambda}^*$, it satisfies

$$\lambda_i^* > 0 \quad \text{for} \quad g_i(\mathbf{x}^*) = 0, \forall i \in [m]$$

and for problem (**H**), if for any KKT stationary point $\mathbf{z}^*, \nu^*$, it satisfies

$$\nu^* > 0 \quad \text{for} \quad \|\mathbf{z}^*\| = 1.$$

We remark that the transformation $\boldsymbol{\psi}$ preserves the strict complementary slackness condition as stated in the following formal result. *Therefore, we can assume the strict complementary condition holds for both problem* **P** *and* **H** *without loss of generality in later analysis.*

Lemma D.7. *Consider the problem* (**P**) *and* (**H**)*. Suppose* $\mathbf{x}^\star, \boldsymbol{\lambda}^*, \boldsymbol{\eta}^*$ *are the (primal and dual) KKT stationary solutions to the problem* (**P**) *satisfying the strict complementary slackness condition D.6. Then the KKT stationary point* $\mathbf{z}^*, \nu^*, \boldsymbol{\gamma}^*$ *for problem* (**H**) *will also satisfy the strict complementary slackness condition.*

Proof. Suppose $\mathbf{z}^*, \nu^*, \boldsymbol{\gamma}^*$ is the KKT stationary point of problem (**H**). Let $\mathbf{x}^* = \boldsymbol{\psi}(\mathbf{z}^*)$ and denote the active set $\mathcal{A} = \{i : g_i(\mathbf{x}^*) = 0\}$. Moreover, we define $\lambda_i^* = 0$ if $i \notin \mathcal{A}$ and $\lambda_i^* = \nu^*/|\mathcal{A}|$; and $\boldsymbol{\eta}^* = \boldsymbol{\gamma}^*$. Then from the proof of Lemma D.14, $\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\eta}^*$ is a KKT stationary point. By assumption, it satisfies the strict complementary condition, i.e., $0 < \lambda_i^* = \nu^*/|\mathcal{A}|$ for $i \in |\mathcal{A}|$. Actually, this means $\nu^* > 0$ for $\mathbf{z}^* \in \partial\mathcal{B}$, which proves our conclusion. □

Definition D.8 (Slater's condition)**.** It is said that the Slater's condition holds for a problem with constrained set $\mathcal{K} := \{g_i(\mathbf{x}) \le 0, i \in [m]; \varphi_i(\mathbf{x}) = 0, i \in [p]\}$, if there exists $\mathbf{x}_0$ such that for all $i = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, p$

$$g_i(\mathbf{x}_0) < 0, \varphi_j(\mathbf{x}_0) = 0.$$

It is well established that the KKT condition is a sufficient and necessary condition for global optimality of a convex optimization if Slater's condition holds. One could see, e.g., (Boyd et al., 2004) for reference. Below, we present the formal statement.

Proposition D.9. *Suppose the primal problem* (**P**) *is convex with differentiable objective and constraint functions and Slater's condition holds. Then we have the following conclusions.*

(i) If $\mathbf{x}^*$ *is a primal optimal, then there exists* $(\boldsymbol{\lambda}^*, \boldsymbol{\eta}^*)$ *being a dual optimal point and strong duality holds. Moreover,* $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\eta}^*)$ *satisfy KKT conditions* (18)*.*

(ii) If $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\eta}^*)$ *satisfy the KKT condition* (18)*, then* $\mathbf{x}^*$ *is a global optimal point for problem* (**P**)*.*

Next, we are ready to show the global optimality property of problem (**H**).

Proposition D.10. *Suppose problem* (**P**) *is a convex optimization, i.e.,* $\mathcal{C}$ *is convex and* $\boldsymbol{\varphi}$ *is linear. Let Slater's condition and the strict complementary slackness condition hold for both problem* (**P**) *and* (**H**)*. Then the following holds.*

(1) If $\mathbf{x}^*$ *is a stationary point (thus a global optimum) in problem* (**P**)*, then* $\mathbf{z}^* = \Psi^{-1}(\mathbf{x}^*)$ *is a global optimum in problem* (**H**) *and* $\mathbf{z}^*$ *is a KKT stationary point of problem* (**H**)*.*

(2) If $\mathbf{z}^*$ *is a KKT stationary point of problem* (**H**)*, then* $\mathbf{x}^* = \Psi(\mathbf{z}^*)$ *is a global optimum of problem* (**P**)*. Hence,* $\mathbf{z}^*$ *is a global optimum.*

Proof. 1) Due to the global optimality of $\mathbf{x}^*$, we have

$$f(\mathbf{x}^*) \leq f(\mathbf{x}), \forall \mathbf{x} \in \mathcal{K}.$$

Hence with the homeomorphism Ψ and $\mathbf{x} = \Psi(\mathbf{z})$, we have

$$f(\Psi(\mathbf{z}^*)) \leq f(\Psi(\mathbf{z})), \forall \mathbf{z} \in \mathcal{F} := \mathcal{B} \cap \{\boldsymbol{\phi}(\mathbf{z}) = 0\},$$

that is

$$h(\mathbf{z}^*) \leq h(\mathbf{z}), \forall \mathbf{z} \in \mathcal{F}.$$

Therefore, $\mathbf{z}^*$ is a KKT stationary point referring to Prop. D.9.

2) By assumption, it follows from Lemma D.14 that $\mathbf{x}^* = \Psi(\mathbf{z}^*)$ is a KKT stationary point of problem (**P**). Hence by Proposition D.9, $\mathbf{x}^*$ is also a global optimum of problem (**P**). Thus, $\mathbf{z}^*$ is a global optimum of problem (**H**) by the conclusion of item (1). □

D.4 Strong Duality of Problem (**H**)

In this section, we show the strong duality of problem **H**. We first provide a classical result in the following. One could refer to, e.g., (Deng et al., 2025) for the proof.

Proposition D.11. *Suppose problem* (**P**) *is convex with a non-empty solution set and Slater's condition holds. Then the strong duality holds with related to the AL dual formulation (also called saddle point formulation), that is,*

$$\max_{\boldsymbol{\eta} \in \mathbb{R}^p} \min_{\mathbf{x} \in \mathcal{C}} \hat{f}_\rho(\mathbf{x}; \boldsymbol{\eta}) = \min_{\mathbf{x} \in \mathcal{C}} \max_{\boldsymbol{\eta} \in \mathbb{R}^p} \hat{f}_\rho(\mathbf{x}; \boldsymbol{\eta}) = f^* \tag{19}$$

where f^* *is the optimal objective of problem* (**P**)*.*

Next, we are ready to prove the strong duality property.

Proposition D.12. *Suppose problem* (**P**) *is convex with a non-empty solution set and Slater's condition holds. Then the strong duality holds for problem* (**H**)*, related to the AL dual formulation (also called saddle point formulation), that is,*

$$\max_{\boldsymbol{\gamma} \in \mathbb{R}^p} \min_{\mathbf{z} \in \mathcal{B}} \hat{h}_\rho(\mathbf{z}; \boldsymbol{\gamma}) = \min_{\mathbf{z} \in \mathcal{B}} \max_{\boldsymbol{\gamma} \in \mathbb{R}^p} \hat{h}_\rho(\mathbf{z}; \boldsymbol{\gamma}) = h^* = f^* \tag{20}$$

where h^* *and* f^* *are the optimal objectives of problem* (**H**) *and* (**P**)*, respectively.*

Proof. First, it is easy to derive

$$\max_{\boldsymbol{\gamma} \in \mathbb{R}^p} \hat{h}_\rho(\mathbf{z}; \boldsymbol{\gamma}) = \begin{cases} \infty, & \text{if } \boldsymbol{\phi}(\mathbf{z}) \neq 0. \\ h(\mathbf{z}), & \text{if } \boldsymbol{\phi}(\mathbf{z}) = 0. \end{cases} \tag{21}$$

where the first term is from $\hat{h}_\rho(\mathbf{z};\boldsymbol{\gamma})$ is a linear function of $\boldsymbol{\gamma}$ if $\boldsymbol{\phi}(\mathbf{z}) \neq 0$. Thus, we have

$$\min_{\mathbf{z}\in\mathcal{B}} \max_{\boldsymbol{\gamma}\in\mathbb{R}^p} \hat{h}_\rho(\mathbf{z};\boldsymbol{\gamma}) = h^* = f^* = \min_{\mathbf{x}\in\mathcal{C}} \max_{\boldsymbol{\eta}\in\mathbb{R}^p} \hat{f}_\rho(\mathbf{x};\boldsymbol{\eta}). \tag{22}$$

Besides, it follows from the homeomorphic transformation $\mathbf{x} = \Psi(\mathbf{z})$ that

$$\max_{\boldsymbol{\gamma}\in\mathbb{R}^p} \min_{\mathbf{z}\in\mathcal{B}} \hat{h}_\rho(\mathbf{z};\boldsymbol{\gamma}) = \max_{\boldsymbol{\eta}\in\mathbb{R}^p} \min_{\mathbf{x}\in\mathcal{C}} \hat{f}_\rho(\mathbf{x};\boldsymbol{\eta}). \tag{23}$$

Therefore, we get the conclusion from Prop. D.11. □

D.5 KKT Conditions of Problem (P) and (H)

First, we recall some basic definitions.

The Lagrangian function of problem (**P**) is defined as

$$\mathcal{L}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\eta}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \eta_i \varphi_i(\mathbf{x})$$

where λ_i (or η_i) represents the Lagrange multiplier associated with each inequality (equality, resp.) constraint.

To define the second-order KKT condition for the optimization problems, we recall that the critical cone in the following.

Definition D.13 (Critical cone)**.** Denote the feasible region of problem (**P**) as $\mathcal{K}$. Then the critical cone $\mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)$ at a feasible point $\mathbf{x}^*$ of problem (**P**) is defined as (Nocedal & Wright, 1999)

$$\mathbf{w} \in \mathrm{C}_{\mathcal{K}}(\mathbf{x}^*) \Leftrightarrow \begin{cases} \nabla\varphi_i(\mathbf{x}^*)^\top \mathbf{w} = 0, & \text{for all } i \in [p], \\ \nabla g_i(\mathbf{x}^*)^\top \mathbf{w} = 0, & \text{for all } i \in \mathcal{A}(\mathbf{x}^*) \text{ with } \lambda_i^* > 0, \\ \nabla g_i(\mathbf{x}^*)^\top \mathbf{w} \geq 0, & \text{for all } i \in \mathcal{A}(\mathbf{x}^*) \text{ with } \lambda_i^* = 0. \end{cases}$$

Here $\boldsymbol{\lambda}^*$ is the Lagrangian multiplier of inequality constraints g_i and $\mathcal{A}(\mathbf{x}^*)$ is the index of active constraints, i.e., $\mathcal{A}(\mathbf{x}^*) = \{i \mid g_i(\mathbf{x}^*) = 0\}$.

Moreover, if *strict complementary slackness* holds, the critical cone is simplified as

$$\mathrm{C}_{\mathcal{K}}(\mathbf{x}^*) = \{\mathbf{w} \in \mathbb{R}^n \mid \nabla\varphi_j(\mathbf{x}^*)^\top \mathbf{w} = 0, \nabla g_i(\mathbf{x}^*)^\top \mathbf{w} = 0, \text{ for all } j \in [p], i \in \mathcal{A}(\mathbf{x}^*)\}.$$

Suppose *strict complementary slackness* holds for problem (**P**) and (**H**). Then, we can write KKT conditions for problem (**P**) and (**H**) in the following.

First-order KKT conditions on $\mathbf{x}^*$. The Lagrangian of (**P**) is

$$\mathcal{L}_{\mathrm{P}}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\eta}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p} \eta_i \varphi_i(\mathbf{x}).$$

The first-order KKT conditions of (**P**) are: there exists $\boldsymbol{\lambda}^*, \boldsymbol{\eta}^*$ such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \eta_i^* \nabla\varphi_i(\mathbf{x}^*) = \mathbf{0}, \tag{24a}$$

$$\varphi_j(\mathbf{x}^*) = 0,\ g_i(\mathbf{x}^*) \leq 0, \quad \forall j \in [p], i \in [m]; \tag{24b}$$

$$\boldsymbol{\lambda}^* \geq \mathbf{0},\ \lambda_i^* g_i(\mathbf{x}^*) = 0, \quad \forall i \in [m]. \tag{24c}$$

Second-order KKT conditions on $\mathbf{x}^*$. It adds the following condition

$$\mathbf{w}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}_{\mathrm{P}}(\mathbf{x}^*, \boldsymbol{\lambda}^*)\mathbf{w} \geq 0 \tag{25}$$

for any $\mathbf{w} \in \mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)$.

First-order KKT conditions on $\mathbf{z}^*$. The Lagrangian of (**H**) is

$$\mathcal{L}_{\mathrm{H}}(\mathbf{z}, \nu, \boldsymbol{\gamma}) = h(\mathbf{z}) + \nu(\|\mathbf{z}\|^2 - 1) + \sum_{i=1}^{p} \gamma_i \phi_i(\mathbf{z}).$$

The first-order KKT conditions of (**H**) are: there exists ν^* and $\boldsymbol{\gamma}^*$ such that

$$\nabla h(\mathbf{z}^*) + 2\nu^* \mathbf{z}^* + \sum_{i=1}^{p} \gamma_i^* \nabla \phi_i(\mathbf{z}^*) = \mathbf{0}, \tag{26a}$$

$$\phi_j(\mathbf{z}^*) = 0,\ \|\mathbf{z}^*\|^2 \le 1, \quad \forall j \in [p]. \tag{26b}$$

$$\nu^* \ge 0,\ \nu^*(\|\mathbf{z}^*\|^2 - 1) = 0. \tag{26c}$$

Second-order KKT condition on $\mathbf{z}^*$. It will add the following condition.

$$\mathbf{d}^\top \nabla^2_{\mathbf{z}} \mathcal{L}_{\mathrm{H}}(\mathbf{z}^*, \nu^*) \mathbf{d} \ge 0 \tag{27}$$

for any $\mathbf{d} \in \mathrm{C}_{\mathcal{F}}(\mathbf{z}^*) = \mathrm{C}_{\mathcal{B}}(\mathbf{z}^*) \bigcap \{\mathbf{d}^* \mid \nabla \phi_j(\mathbf{z}^*)^\top \mathbf{d}^* = 0\}$ where $\mathcal{F}$ denotes the feasible region of problem (**H**). Here recall that

$$\mathrm{C}_{\mathcal{B}}(\mathbf{z}^*) = \begin{cases} \mathbb{R}^n, & \text{if } \mathbf{z}^* \in \mathrm{int}(\mathcal{B}), \\ \{\mathbf{d} : \mathbf{d}^\top \mathbf{z}^* = 0\}, & \text{if } \mathbf{z}^* \in \partial\mathcal{B}. \end{cases}$$

D.6 Relationships of KKT Stationary Points between Problem (**P**) and (**H**)

Lemma D.14. *Suppose strict complementary slackness holds for both problem* (**P**) *and* (**H**). *We have that* $\mathbf{x}^*$ *is a KKT stationary point of* (**P**) *if and only if* $\mathbf{z}^*$ *is also a KKT stationary point of* (**H**) *where* $\mathbf{x}^* = \Psi(\mathbf{z}^*)$.

Proof. 1) First, we assume that $\mathbf{x}^*$ is a KKT stationary point of (**P**). By assumption, there exists $\boldsymbol{\lambda}^*$ such that the KKT condition holds (24) holds. Then we have

$$\mathrm{J}_\Psi(\mathbf{z}^*) \nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \mathrm{J}_\Psi(\mathbf{z}^*) \nabla g_i(\mathbf{x}^*) + \sum_{i=1}^{p} \eta_i^* \mathrm{J}_\Psi(\mathbf{z}^*) \nabla \varphi_i(\mathbf{x}^*) = \mathbf{0},$$
$$\varphi_j(\Psi(\mathbf{z}^*)) = 0,\ g_i(\Psi(\mathbf{z}^*)) \le 0, \quad \forall i \in [m], j \in [p];$$
$$\boldsymbol{\lambda}^* \ge \mathbf{0},\ \lambda_i^* g_i(\Psi(\mathbf{z}^*)) = 0, \quad \forall i \in [m].$$

This is equivalent to

$$\nabla h(\mathbf{z}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla G_i(\mathbf{z}^*) + \sum_{i=1}^{p} \eta_i^* \nabla \phi_i(\mathbf{z}^*) = \mathbf{0}, \tag{28a}$$

$$\phi_j(\mathbf{z}^*) = 0,\ G_i(\mathbf{z}^*) \le 0, \quad \forall i \in [m], j \in [p]; \tag{28b}$$

$$\boldsymbol{\lambda}^* \ge \mathbf{0},\ \lambda_i^* G_i(\mathbf{z}^*) = 0, \quad \forall i \in [m]. \tag{28c}$$

Let $\nu^* = \sum_{i=1}^m \lambda_i^*$ and $\gamma_j^* = \eta_j^*$ for all $j \in [p]$. According to the eq. (14,15), eq. (28a) is actually

$$\nabla h(\mathbf{z}^*) + 2\nu^* \mathbf{z}^* + \sum_{i=1}^{p} \gamma_i^* \nabla \phi_i(\mathbf{z}^*) = \mathbf{0}.$$

By assumption, eq. (28b) is equivalent to

$$\phi_j(\mathbf{z}^*) = 0,\ \|\mathbf{z}^*\|^2 \le 1, \quad \forall j \in [p].$$

Note that if $G_i(\mathbf{z}^*) < 0$ for all i, then $\boldsymbol{\lambda}^* = 0$ and thus $\nu^* = 0$. In this case, $\nu^*(\|\mathbf{z}^*\|^2 - 1) = 0$. If $\mathbf{z}^*$ makes at least one $G_i(\mathbf{z}^*) = 0$, then we have $\|\mathbf{z}^*\|^2 = 1$. In this case, we also have $\nu^*(\|\mathbf{z}^*\|^2 - 1) = 0$. Hence, eq. (28c) implies

$$\nu^* \ge 0, \nu^*(\|\mathbf{z}^*\|^2 - 1) = 0.$$

In conclusion, there exists $\mathbf{z}^*, \nu^*, \boldsymbol{\gamma}^*$ such the KKT condition holds.

2) Now, we assume $\mathbf{z}^*, \nu^*, \boldsymbol{\gamma}^*$ satisfy KKT condition for problem (**H**), i.e.,

$$\nabla h(\mathbf{z}^*) + 2\nu^*\mathbf{z}^* + \sum_{i=1}^{p} \gamma_i^* \nabla \phi_i(\mathbf{z}^*) = \mathbf{0},$$
$$\phi_j(\mathbf{z}^*) = 0,\ \|\mathbf{z}^*\|^2 \le 1, \quad \forall j \in [p];$$
$$\nu^* \ge 0,\ \nu^*(\|\mathbf{z}^*\|^2 - 1) = 0.$$

If $\mathbf{z}^* \in \mathrm{int}(\mathcal{B})$, then $G_i(\mathbf{z}^*) < 0$ for all i and $\nu^* = 0$. In this case, there exists $\mathbf{x}^* = \Psi(\mathbf{z}^*), \boldsymbol{\lambda}^* = \mathbf{0}, \boldsymbol{\eta}^* = \boldsymbol{\gamma}^*$ such that the KKT condition with eq. (24) of problem (**P**) holds.

If $\mathbf{z}^* \in \partial\mathcal{B}$, then there exists at least one $i \in \{1, 2, \cdots, m\}$ such that $G_i(\mathbf{z}^*) = 0$ and $\nu^* > 0$ from strict complementary slackness. Denote $\mathcal{A} = \{i : G_i(\mathbf{z}^*) = 0\}$. Note we define $\lambda_i^* = 0$ if $i \notin \mathcal{A}$ and $\lambda_i^* = \nu^*/|\mathcal{A}|$. Then we have $\mathbf{z}^*, \boldsymbol{\lambda}^*, \eta^* = \boldsymbol{\gamma}^*$ such that the KKT condition (24) of problem (**P**) hold.

□

Lemma D.15. *Suppose strict complementary slackness condition holds for both problem* (**P**) *and* (**H**). *Then* $\mathbf{x}^*$ *is a second-order KKT stationary point of* (**P**) *if and only if* $\mathbf{z}^* = \Psi^{-1}(\mathbf{x}^*)$ *is also a second-order KKT stationary point of* (**H**).

Proof. From Lemma D.14, there exists $\boldsymbol{\lambda}^*$,ν^* and $\boldsymbol{\gamma}^* = \boldsymbol{\eta}^*$ such that $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\eta}^*)$ holds for first-order KKT condition of (**P**) if and only if $(\mathbf{z}^*, \nu^*, \boldsymbol{\gamma}^*)$ holds for first-order KKT condition of (**H**). Hence, it suffices to show the equivalence of condition (27) and (25).

1) Let's first suppose $\mathbf{x}^*$ is a second-order KKT stationary point, i.e., eq. (25) holds. Note

$$\nabla^2_{\mathbf{z}} \mathcal{L}_{\mathrm{H}}(\mathbf{z}^*, \nu^*) = \nabla^2 h(\mathbf{z}^*) + 2\nu^* \mathbf{I}_n + \sum_{i=1}^{p} \gamma_i^* \nabla^2 \phi_i(\mathbf{z}^*),$$

where $\mathbf{I}_n$ is identity matrix of size $n \times n$. We just need to show $\mathbf{d}^\top \nabla \mathcal{L}_{\mathrm{H}}(\mathbf{z}^*, \nu^*)\mathbf{d} \ge 0$ for any $\mathbf{d} \in \mathrm{C}_{\mathcal{F}}(\mathbf{z}^*)$. Recall that

$$\nabla^2 h(\mathbf{z}^*) = \mathrm{J}_\Psi(\mathbf{z}^*) \nabla^2 f(\Psi(\mathbf{z}^*)) \mathrm{J}_\Psi(\mathbf{z}^*)^\top + \sum_{i=1}^{n} \frac{\partial f}{\partial \mathbf{x}_i}(\Psi(\mathbf{z}^*)) \nabla^2 \Psi_i(\mathbf{z}^*),$$

$$\nabla^2 G_i(\mathbf{z}^*) = \mathrm{J}_\Psi(\mathbf{z}^*) \nabla^2 g_i(\Psi(\mathbf{z}^*)) \mathrm{J}_\Psi(\mathbf{z}^*)^\top + \sum_{k=1}^{n} \frac{\partial g_i}{\partial \mathbf{x}_k}(\Psi(\mathbf{z}^*)) \nabla^2 \Psi_k(\mathbf{z}^*), \quad i = 1, 2, \cdots, m,$$

and

$$\nabla^2 \phi_i(\mathbf{z}^*) = \mathrm{J}_\Psi(\mathbf{z}^*) \nabla^2 \varphi_i(\Psi(\mathbf{z}^*)) \mathrm{J}_\Psi(\mathbf{z}^*)^\top + \sum_{k=1}^{n} \frac{\partial \varphi_i}{\partial \mathbf{x}_k}(\Psi(\mathbf{z}^*)) \nabla^2 \Psi_k(\mathbf{z}^*), \quad i = 1, 2, \cdots, p.$$

From eq. (14), note that

$$\nabla^2 G_k(\mathbf{z}^*) = 2\mathbf{I}_n, \forall k \in \mathcal{A}(\mathbf{x}^*) \cap \{k : G_k(\mathbf{z}^*) = 0\}.$$

From Lemma D.14, $\nu^* = \sum_i \lambda_i^*$ and $\boldsymbol{\eta}^* = \boldsymbol{\gamma}^*$. Then we have

$$\nabla^2 \mathcal{L}_{\mathrm{H}}(\mathbf{z}^*, \nu^*) = \nabla^2 h(\mathbf{z}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 G_i(\mathbf{z}^*) + \sum_{i=1}^{p} \gamma_i^* \nabla^2 \phi_i(\mathbf{z}^*) \tag{29a}$$

$$= \mathrm{J}_\Psi(\mathbf{z}^*) \nabla^2 f(\Psi(\mathbf{z}^*)) \mathrm{J}_\Psi(\mathbf{z}^*)^\top + \sum_{i=1}^{m} \mathrm{J}_\Psi(\mathbf{z}^*) \lambda_i^* \nabla^2 g_i(\Psi(\mathbf{z}^*)) \mathrm{J}_\Psi(\mathbf{z}^*)^\top + \sum_{i=1}^{p} \mathrm{J}_\Psi(\mathbf{z}^*) \nabla^2 \varphi_i(\Psi(\mathbf{z}^*)) \mathrm{J}_\Psi(\mathbf{z}^*)^\top \tag{29b}$$

$$+\sum_{k=1}^{n}\frac{\partial f}{\partial \mathbf{x}_k}(\Psi(\mathbf{z}^*))\nabla^2\Psi_k(\mathbf{z}^*)+\sum_{k=1}^{n}\sum_{i=1}^{m}\lambda_i^*\frac{\partial g_i}{\partial \mathbf{x}_k}(\Psi(\mathbf{z}^*))\nabla^2\Psi_k(\mathbf{z}^*)+\sum_{k=1}^{n}\sum_{i=1}^{p}\eta_i^*\frac{\partial \varphi_i}{\partial \mathbf{x}_k}(\Psi(\mathbf{z}^*))\nabla^2\Psi_k(\mathbf{z}^*). \tag{29c}$$

From first-order KKT stationarity of (**P**), i.e.,

$$\nabla f(\mathbf{x}^*)+\sum_{i=1}^{m}\lambda_i^*\nabla g_i(\mathbf{x}^*)+\sum_{i=1}^{m}\eta_i^*\nabla\varphi_i(\mathbf{x}^*)=\mathbf{0},$$

We have

$$\frac{\partial f}{\partial \mathbf{x}_k}(\mathbf{x}^*)+\sum_{i=1}^{m}\lambda_i^*\frac{\partial g_i}{\partial \mathbf{x}_k}(\mathbf{x}^*)+\sum_{i=1}^{p}\eta_i^*\frac{\partial \varphi_i}{\partial \mathbf{x}_k}(\mathbf{x}^*)=0.$$

Hence, we have the second term (29c) is equal to 0.

Now we note it's trivial that $\mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)=\mathrm{C}_{\mathcal{F}}(\mathbf{z}^*)$ if $\mathbf{z}^*\in\mathrm{int}(\mathcal{C})$ where $\mathbf{x}^*=\Psi(\mathbf{z}^*)$. In this case if $\mathbf{d}\in\mathrm{C}_{\mathcal{F}}(\mathbf{z}^*)$, we will have $\mathrm{J}_\Psi(\mathbf{z}^*)^\top\mathbf{d}\in\mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)$

If $\mathbf{x}^*\in\partial\mathcal{C}$, then $\mathcal{A}(\mathbf{x}^*)\neq\emptyset$. For any $\mathbf{d}\in\mathrm{C}_{\mathcal{F}}(\mathbf{z}^*)$, we have

$$(\mathrm{J}_\Psi(\mathbf{z}^*)^\top\mathbf{d})^\top\nabla g_i(\mathbf{x}^*)=\mathbf{d}^\top\mathrm{J}_\Psi(\mathbf{z}^*)^\top\nabla g_i(\mathbf{x}^*)=\mathbf{d}^\top\nabla G_i(\mathbf{z}^*)=2\mathbf{d}^\top\mathbf{z}^*=0,\quad \text{for } i\in\mathcal{A}(\mathbf{x}^*),$$

and

$$(\mathrm{J}_\Psi(\mathbf{z}^*)^\top\mathbf{d})^\top\nabla\varphi_i(\mathbf{x}^*)=\mathbf{d}^\top\mathrm{J}_\Psi(\mathbf{z}^*)^\top\nabla\varphi_i(\mathbf{x}^*)=\mathbf{d}^\top\phi_i(\mathbf{z}^*)=0,\quad \text{for } i\in[p].$$

Thus, we have $\mathrm{J}_\Psi(\mathbf{z}^*)^\top\mathbf{d}\in\mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)$.

So for $\mathbf{d}\in\mathrm{C}_{\mathcal{F}}(\mathbf{z}^*)$, we have the following holds about the first term of $\nabla^2\mathcal{L}_{\mathrm{H}}(\mathbf{z}^*,\nu^*)$.

$$\begin{aligned}
&\left(\mathrm{J}_\Psi\left(\mathbf{z}^*\right)^\top\mathbf{d}\right)^\top\nabla^2 f\left(\Psi\left(\mathbf{z}^*\right)\right)\mathrm{J}_\Psi\left(\mathbf{z}^*\right)^\top\mathbf{d}\\
&+\left(\mathrm{J}_\Psi\left(\mathbf{z}^*\right)^\top\mathbf{d}\right)^\top\left(\sum_{i=1}^{m}\lambda_i^*\nabla^2 g_i\left(\Psi\left(\mathbf{z}^*\right)\right)\right)\mathrm{J}_\Psi\left(\mathbf{z}^*\right)^\top\mathbf{d}\\
&+\left(\mathrm{J}_\Psi\left(\mathbf{z}^*\right)^\top\mathbf{d}\right)^\top\left(\sum_{i=1}^{p}\eta_i^*\nabla^2\varphi_i\left(\Psi\left(\mathbf{z}^*\right)\right)\right)\mathrm{J}_\Psi\left(\mathbf{z}^*\right)^\top\mathbf{d}\\
&\geq 0
\end{aligned}$$

where the last '$\geq$' is from the assumption that $\mathbf{x}^*$ is a second-order KKT stationary point of (**P**). Hence, we have $\mathbf{d}^\top\nabla^2\mathcal{L}_{\mathrm{H}}(\mathbf{z}^*,\nu^*)\mathbf{d}\geq 0$ for any $\mathbf{d}\in\mathrm{C}_{\mathcal{F}}(\mathbf{z}^*)$, i.e., $\mathbf{z}^*=\Psi^{-1}(\mathbf{x}^*)$ is also a second-order KKT stationary point.

2) Let's suppose $\mathbf{z}^*$ is a second-order KKT stationary point of (**H**) and show that $\mathbf{x}^*$ is a second-order KKT stationary point of (**P**).

If $\mathbf{z}^*\in\mathrm{int}(\mathcal{B})$, the proof is trivial because $\nu^*=0$ according to the similar analysis. So we assume $\mathbf{z}^*\in\partial\mathcal{B}$. Define $\mathcal{A}(\mathbf{z}^*)=\{i:G_i(\mathbf{z}^*)=0\}$, and $\lambda_i^*=0$ for $i\notin\mathcal{A}(\mathbf{z}^*)$, $\lambda_i^*=\nu^*/|\mathcal{A}(\mathbf{z}^*)|$ for $i\in\mathcal{A}(\mathbf{z}^*)$.

Note for any $\mathbf{w}\in\mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)$, we have

$$\mathbf{0}=\mathbf{w}^\top\nabla g_i(\mathbf{x}^*)=\mathbf{w}^\top\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)\nabla G_i(\mathbf{z}^*)=(\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top\mathbf{w})^\top\mathbf{z}^*,\quad \text{for}\ \ i\in\mathcal{A}(\mathbf{x}^*)=\mathcal{A}(\mathbf{z}^*).$$

and

$$\mathbf{0}=\mathbf{w}^\top\nabla\varphi_i(\mathbf{x}^*)=\mathbf{w}^\top\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)\nabla\phi_i(\mathbf{z}^*)=(\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top\mathbf{w})^\top\nabla\phi_i(\mathbf{z}^*),\quad \text{for}\ \ i\in[p].$$

Hence $\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w} \in \mathrm{C}_{\mathcal{F}}(\mathbf{z}^*)$. Then for any $\mathbf{w} \in \mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)$,

$$
\begin{aligned}
&\mathbf{w}^\top \nabla^2_{\mathbf{x}} \mathcal{L}_{\mathrm{P}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{w} \\
&= \mathbf{w}^\top \nabla^2 f(\mathbf{x}^*) \mathbf{w} + \mathbf{w}^\top \sum_{i=1}^m \lambda_i^* \nabla^2 g_i(\mathbf{x}^*) \mathbf{w} + \mathbf{w}^\top \sum_{i=1}^p \eta_i^* \nabla^2 \varphi_i(\mathbf{x}^*) \mathbf{w} \\
&= (\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w})^\top \mathrm{J}_\Psi(\mathbf{z}^*) \nabla^2 f(\mathbf{x}^*) \mathrm{J}_\Psi(\mathbf{z}^*)^\top \mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w} \\
&+ \left(\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w}\right)^\top \mathrm{J}_\Psi(\mathbf{z}^*) \left(\sum_{i=1}^m \lambda_i^* \nabla^2 g_i(\mathbf{x}^*) \right) \mathrm{J}_\Psi(\mathbf{z}^*)^\top \mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w} \\
&+ \left(\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w}\right)^\top \mathrm{J}_\Psi(\mathbf{z}^*) \left(\sum_{i=1}^p \eta_i^* \nabla^2 \varphi_i(\mathbf{x}^*) \right) \mathrm{J}_\Psi(\mathbf{z}^*)^\top \mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w} \\
&+ (\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w})^\top (\star) \mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w} \\
&= (\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w})^\top \mathcal{L}_{\mathrm{H}}(\mathbf{z}^*, \nu^*) \mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w} \geq 0,
\end{aligned}
$$

where

$$
(\star) := \sum_{k=1}^n \frac{\partial f}{\partial \mathbf{x}_k}(\Psi(\mathbf{z}^*)) \nabla^2 \Psi_k(\mathbf{z}^*) + \sum_{k=1}^n \sum_{i=1}^m \lambda_i^* \frac{\partial g_i}{\partial \mathbf{x}_k}(\Psi(\mathbf{z}^*)) \nabla^2 \Psi_k(\mathbf{z}^*) + \sum_{k=1}^n \sum_{i=1}^p \eta_i^* \frac{\partial \varphi_i}{\partial \mathbf{x}_k}(\Psi(\mathbf{z}^*)) \nabla^2 \Psi_k(\mathbf{z}^*)
$$

and the term with $(\star)$ of the second '=' is exactly 0 and the last '$\geq$' is from the assumption that $\mathbf{z}^*$ is a second-order KKT stationary point of $(\mathbf{H})$.

□

Definition D.16 (Non-degenerate KKT stationary point)**.** A second-order KKT point $\mathbf{x}^*$ of $(\mathbf{P})$ is said to be non-degenerate if there exists $\boldsymbol{\lambda}^*, \boldsymbol{\eta}^*$ such that

$$
\mathbf{d}^\top \nabla^2 \mathcal{L}_{\mathrm{P}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{d} > 0
$$

for all $0 \neq \mathbf{d} \in \mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)$.

Lemma D.17. *Suppose strict complementary slackness holds for problem* $(\mathbf{P})$ *and* $(\mathbf{H})$*. Then* $\mathbf{x}^\star$ *is a non-degenerate KKT point of optimization* $(\mathbf{P})$ *if and only if* $\mathbf{z}^\star$ *is also a non-degenerate KKT point of problem* $(\mathbf{H})$ *where* $\mathbf{x}^\star = \Psi(\mathbf{z}^*)$*.*

Proof. 1) Suppose $\mathbf{x}^*$ is a non-degenerate KKT stationary point. Note that for $\mathbf{d} \in \mathrm{C}_{\mathcal{F}}(\mathbf{z}^*)$, we have $\mathrm{J}_\Psi(\mathbf{z}^*)^\top \mathbf{d} \in \mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)$ from the proof of Lemma D.15. Moreover, from $\mathrm{J}_\Psi(\mathbf{z}^*) \neq 0$ we have $\mathrm{J}_\Psi(\mathbf{z}^*)^\top \mathbf{d} \neq 0$ if and only if $\mathbf{d} \neq 0$. Then the conclusion is trivial from eq. (29) in the proof of Lemma D.15.

2) Now, we suppose $\mathbf{z}^*$ is a non-degenerate KKT stationary point. It follows from the proof of Lemma D.15 that for any $\mathbf{w} \in \mathrm{C}_{\mathcal{K}}(\mathbf{x}^*)$, we have $\mathrm{J}_\Psi^{-1}(\mathbf{z}^*)^\top \mathbf{w} \in \mathrm{C}_{\mathcal{F}}(\mathbf{z}^*)$. Hence, the conclusion is also trivial from the proof of item (2) of Lemma D.15.

□

E Convergence Analysis

In this section, we will provide the omitted proof in Sec. 4.2.

E.1 Hom-APG

For reference, the Hom-PGD routine used to solve the transformed regularized models is reproduced below.

We consider the general convex optimization problem

$$
\min_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}), \tag{30}
$$

Algorithm 4 Hom-PGD (Liu et al., 2025)

Input: initial point $\mathbf{z}_0$, problem $(\mathbf{H}_s)$ with homeomorphism Ψ
for $k = 0$ **to** K **do**
  Compute stepsize α_k
  Update: $\mathbf{z}_{k+1} = \Pi_{\mathcal{B}}(\mathbf{z}_k - \alpha_k \nabla h(\mathbf{z}_k))$
end for
Output: $\mathbf{x}_K = \Psi(\mathbf{z}_K)$

Algorithm 5 Hom-APG

Input: Initial point $\mathbf{z}_0 \in \mathcal{B}$, regularization parameter $\lambda_{\text{reg}} > 0$
$\mathbf{x}_0 \leftarrow \Psi(\mathbf{z}_0)$, $\mathbf{y}_0 \leftarrow \mathbf{x}_0$, $q \leftarrow \mu_f/(\mu_f + \lambda_{\text{reg}})$, $\tau_0 \leftarrow \sqrt{q}$
for $k = 1, 2, \ldots$ **do**
  approximately solve $\min_{\mathbf{z}\in\mathcal{B}} \widetilde{Q}_k(\mathbf{z})$ and obtain $\mathbf{z}_k \in \mathcal{B}$
  $\mathbf{x}_k \leftarrow \Psi(\mathbf{z}_k)$
  choose $\tau_k \in (0,1)$ satisfying $\tau_k^2 = (1-\tau_k)\tau_{k-1}^2 + q\tau_k$
  $\iota_k \leftarrow \tau_{k-1}(1-\tau_{k-1})/(\tau_{k-1}^2 + \tau_k)$
  $\mathbf{y}_k \leftarrow \mathbf{x}_k + \iota_k(\mathbf{x}_k - \mathbf{x}_{k-1})$
end for

where f is L_f-smooth and μ_f-strongly convex, and g is a proper closed convex function.

As background, when $g \equiv 0$, a standard approach is gradient descent,

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{1}{L_f}\nabla f(\mathbf{x}_k),$$

which, in the strongly convex case, enjoys the linear convergence rate

$$f(\mathbf{x}_k) - f^\star \le \left(1 - \frac{\mu_f}{L_f}\right)^k \big(f(\mathbf{x}_0) - f^\star\big).$$

Nesterov's accelerated gradient (NAG) (Nesterov, 2004; 2013; Nesterov et al., 2018) further improves this dependence on the condition number and achieves

$$f(\mathbf{x}_k) - f^\star \le C\left(1 - \sqrt{\frac{\mu_f}{L_f}}\right)^k,$$

for some constant C depending on the initialization. Thus, in the strongly convex regime, acceleration improves the convergence factor from $1 - \mu_f/L_f$ to $1 - \sqrt{\mu_f/L_f}$. Catalyst can be viewed as a general acceleration framework that seeks to transfer this NAG-type improvement to broader composite settings.

At outer step k, define

$$Q_k(\mathbf{x}) := F(\mathbf{x}) + \frac{\lambda_{\text{reg}}}{2}\|\mathbf{x} - \mathbf{y}_{k-1}\|^2. \tag{31}$$

Motivated by the catalyst framework of (Lin et al., 2018), we propose **Hom-APG**, a combination of homeomorphic transformation and APG, shown in Algorithm 5, for the transformed problem

$$\min_{\mathbf{z}\in\mathcal{B}} h(\mathbf{z}) := f(\Psi(\mathbf{z})). \tag{32}$$

Then the transformed Catalyst subproblem becomes

$$\widetilde{Q}_k(\mathbf{z}) := h(\mathbf{z}) + \frac{\lambda_{\text{reg}}}{2}\|\Psi(\mathbf{z}) - \mathbf{y}_{k-1}\|^2, \quad \mathbf{z} \in \mathcal{B}. \tag{33}$$

The following lemma shows that the transformed Catalyst subproblem is exactly the inexact solve of the original regularized model Q_k.

Lemma E.1 (Exact model transfer)**.** *For every $k \geq 1$ and $\mathbf{z} \in \mathcal{B}$, we have*

$$Q_k(\Psi(\mathbf{z})) = \widetilde{Q}_k(\mathbf{z}).$$

Consequently,

$$\min_{\mathbf{x}} Q_k(\mathbf{x}) = \min_{\mathbf{z}\in\mathcal{B}} \widetilde{Q}_k(\mathbf{z}),$$

and any approximate minimizer $\mathbf{z}_k$ of (33) induces an approximate minimizer $\mathbf{x}_k = \Psi(\mathbf{z}_k)$ of (31) with the same objective gap.

Proof. Since $g = \mathbb{I}_{\mathcal{C}}$ and $\Psi(\mathcal{B}) = \mathcal{C}$, for any $\mathbf{z} \in \mathcal{B}$ we have

$$F(\Psi(\mathbf{z})) = f(\Psi(\mathbf{z})) = h(\mathbf{z}).$$

Substituting into (31) yields

$$Q_k(\Psi(\mathbf{z})) = h(\mathbf{z}) + \frac{\lambda_{\text{reg}}}{2}\|\Psi(\mathbf{z}) - \mathbf{y}_{k-1}\|^2 = \widetilde{Q}_k(\mathbf{z}).$$

The rest follows from bijectivity of Ψ between $\mathcal{B}$ and $\mathcal{C}$. □

One should note that, although the transferred model is equivalent to the original one, the proposed Hom-APG avoids the expensive projection cost in each iteration, since the constraint set in the hidden space is simply the ball $\mathcal{B}$.

As a direct corollary of the strongly-convex Catalyst analysis, we obtain the following outer-iteration complexity bound for Hom-APG.

Proposition E.2 (Outer iteration complexity of Hom-APG)**.** *Let $\mu_f > 0$ and $\lambda_{\text{reg}} > 0$, and define*

$$q := \frac{\mu_f}{\mu_f + \lambda_{\text{reg}}}.$$

Consider Algorithm 5 with $\tau_0 = \sqrt{q}$, and suppose each transformed subproblem (33) is solved with accuracy

$$\widetilde{Q}_k(\mathbf{z}_k) - \widetilde{Q}_k(\mathbf{z}_k^\star) \leq \varepsilon_k, \qquad \varepsilon_k = \frac{2}{9}\big(f(\mathbf{x}_0) - f^\star\big)(1-\varpi)^k,$$

for some $\varpi < \sqrt{q}$. Then the mapped iterates $\mathbf{x}_k = \Psi(\mathbf{z}_k)$ satisfy

$$f(\mathbf{x}_k) - f^\star \leq \frac{8}{(\sqrt{q} - \varpi)^2}(1-\varpi)^{k+1}\big(f(\mathbf{x}_0) - f^\star\big).$$

Proof. By Lemma E.1, Algorithm 5 is exactly a hidden-space implementation of the inexact Catalyst scheme applied to the constrained problem (30). Therefore, Proposition 5 of (Lin et al., 2018) for the μ-strongly convex case with criterion (C1) applies directly, which yields the stated estimate. □

To derive the global complexity of Hom-APG, we just need a linearly convergent inner solver, which is guaranteed by the hidden-convex strong-convexity result (Liu et al., 2025).

Proposition E.3 (Linearly-convergent inner solver, (Liu et al., 2025))**.** *Consider problem (30). Then the updating sequence $\{\mathbf{z}_k\}$ by Hom-PGD algorithm 4 with a constant step size $\alpha \in (0, \frac{2}{L_h}]$, converges to a global optimum point $\mathbf{z}^\star$ linearly, i.e.,*

$$h(\mathbf{z}_t) - h(\mathbf{z}^*) \leq (1-\sigma)^t \left(h(\mathbf{z}_0) - h(\mathbf{z}^*)\right),$$

where $\sigma = \frac{\mu_f \kappa_1}{L_h} \in (0, 1)$.

Combined with the above proposition, we can obtain the following theorem showing the convergence rate of Hom-APG in the strongly convex setting.

Theorem 3 (Convergence of Hom-APG)**.** *Let $\mu_f > 0$, and suppose that f is L_f-smooth and μ_f-strongly convex. Consider Hom-APG (Algorithm 5) with Hom-PGD (Algorithm 4) as the inner solver. At the $(k+1)$-st outer iteration, initialize the inner solver from the warm start*

$$\mathbf{z}_{k,0} = \Psi^{-1}\left(\mathbf{x}_k + \frac{\lambda_{\text{reg}}}{\lambda_{\text{reg}} + \mu_f}\left(\mathbf{y}_k - \mathbf{y}_{k-1}\right)\right),$$

and run Hom-PGD for T iterations to solve the transformed regularized subproblem (33). Then, for a suitable choice of the regularization parameter λ_{reg}, the mapped iterates satisfy the total-inner-iteration complexity bound

$$f(\mathbf{x}_s) - f^* \le \mathcal{O}\left(\exp\left(-c\sqrt{\frac{\mu_f}{L_{\text{eff}}}}\, s\right)\right), \tag{34}$$

where $s = kT$ is the total number of Hom-PGD iterations, c is a constant independent of f, and L_{eff} is an effective smoothness constant determined by L_f, the geometry of Ψ, and the compact domain $\mathcal{M}$ containing the iterates; explicitly,

$$L_{\text{eff}} := \frac{L_\Psi L_{f,0} + \kappa_2^2 L_f}{\kappa_2^2 + L_\Psi \operatorname{diam}(\mathcal{M})}.$$

Proof. First, note that the regularized objective

$$Q_k(\mathbf{x}) = F(\mathbf{x}) + \frac{\lambda_{\text{reg}}}{2}\left\|\mathbf{x} - \mathbf{y}_{k-1}\right\|^2$$

is $(\mu_f + \lambda_{\text{reg}})$-strongly convex. Moreover, on the compact domain $\mathcal{M}$ containing the iterates (which may be chosen as the ball $\mathcal{B}$ in this work), the transformed model $\widetilde{Q}_k$ is $L_{\widetilde{\mathcal{Q}}}$-smooth with

$$L_{\widetilde{\mathcal{Q}}} = \kappa_2^2\left(L_f + \lambda_{\text{reg}}\right) + L_\Psi\left(L_{f,0} + \lambda_{\text{reg}} \operatorname{diam}(\mathcal{M})\right),$$

where the expression follows from Lemma D.1. By Proposition E.3, Hom-PGD solves the subproblem (33) to accuracy ϵ within

$$T = \mathcal{O}\left(\frac{1}{\sigma_{\widetilde{\mathcal{Q}}}} \log\left(\frac{1}{\epsilon}\right)\right)$$

inner iterations, where

$$\begin{aligned}\sigma_{\widetilde{\mathcal{Q}}} &:= \frac{\left(\mu_f + \lambda_{\text{reg}}\right)\kappa_1}{L_{\widetilde{\mathcal{Q}}}} \\ &= \frac{c_\Psi\left(\mu_f + \lambda_{\text{reg}}\right)}{L_{\text{eff}} + \lambda_{\text{reg}}}, \qquad c_\Psi := \frac{\kappa_1}{\kappa_2^2 + L_\Psi \operatorname{diam}(\mathcal{M})}.\end{aligned}$$

Further, according to Proposition 12 in (Lin et al., 2018), in the $(k+1)$-st outer loop, the warm start

$$\mathbf{z}_{k,0} = \Psi^{-1}\left(\mathbf{x}_k + \frac{\lambda_{\text{reg}}}{\lambda_{\text{reg}} + \mu_f}\left(\mathbf{y}_k - \mathbf{y}_{k-1}\right)\right)$$

ensures that the number of inner iterations required to satisfy

$$\widetilde{Q}_k(\mathbf{z}_k) - \widetilde{Q}_k(\mathbf{z}_k^\star) \le \varepsilon_k$$

is bounded by $T = \widetilde{\mathcal{O}}\left(1/\sigma_{\widetilde{\mathcal{Q}}}\right)$, independently of the outer-loop index k. The same conclusion applies after the homeomorphic transformation, with constants changed only by the bi-Lipschitz constants of Ψ.

By Lemma E.1, the transformed subproblem $\widetilde{Q}_k(\mathbf{z})$ is equivalent to the original regularized model $Q_k(\mathbf{x})$ under the map $\mathbf{x} = \Psi(\mathbf{z})$. Hence,

$$Q_k(\mathbf{x}_k) - Q_k(\mathbf{x}_k^\star) \le \varepsilon_k \quad \Longleftrightarrow \quad \widetilde{Q}_k(\mathbf{z}_k) - \widetilde{Q}_k(\mathbf{z}_k^\star) \le \varepsilon_k.$$

Applying Proposition E.2 with $\varpi = \sqrt{q}/2$ gives, in terms of the total number $s = kT$ of inner Hom-PGD iterations,

$$f(\mathbf{x}_s) - f^* \leq C\left(1 - \frac{\sqrt{q}}{2}\right)^{s/T} \leq C\exp\left(-\frac{\sqrt{q}\,s}{2T}\right), \tag{35}$$

where the last inequality follows from $1 + x \leq e^x$, and C depends on the initialization.

Since $T = \widetilde{\mathcal{O}}\left(1/\sigma_{\widetilde{\mathcal{Q}}}\right)$, we obtain

$$f(\mathbf{x}_s) - f^* \leq C\exp\left(-\frac{\sqrt{q}\,s}{2T}\right) \leq C'\exp\left(-\frac{\sqrt{q}\,\sigma_{\widetilde{\mathcal{Q}}}}{2}s\right) \leq \mathcal{O}\left(\exp\left(-\sqrt{\frac{\mu_f}{L_{\text{eff}}}}\,s\right)\right). \tag{36}$$

The last inequality follows by optimizing the product $\sqrt{q}\,\sigma_{\widetilde{\mathcal{Q}}}$ over λ_{reg}. In particular, choosing $\lambda_{\text{reg}} = L_{\text{eff}} - 2\mu_f$ when $L_{\text{eff}} > 2\mu_f$ yields

$$\sqrt{q}\,\sigma_{\widetilde{\mathcal{Q}}} = \Theta\left(\sqrt{\frac{\mu_f}{L_{\text{eff}} - \mu_f}}\right) \geq \Theta\left(\sqrt{\frac{\mu_f}{L_{\text{eff}}}}\right),$$

which proves the claimed rate.

□

E.2 Convex Case

Algorithm 6 PALM(θ)

Input: $\mathbf{x}_1 \in \mathcal{C}$, $\boldsymbol{\eta}_1 \in \mathbb{R}^p$, and $\rho_k > 0$.
for $k \geq 1$ **do**

$$\mathbf{x}_{k+1} \leftarrow \arg\min_{\mathbf{x}\in\mathcal{C}} \left\{\mathbb{P}^{\theta}_{\rho_k,\alpha_k}(\mathbf{x};\boldsymbol{\eta}_k) := f(\mathbf{x}) + (1-\theta)\,\boldsymbol{\eta}_k^{\top}\boldsymbol{\varphi}(\mathbf{x}) + \frac{\rho_k}{2}\left\|\boldsymbol{\varphi}(\mathbf{x})\right\|^2 + \frac{\alpha_k}{2}\left\|\mathbf{x} - \mathbf{x}_k\right\|^2\right\}, \tag{37}$$

$$\boldsymbol{\eta}_{k+1} \leftarrow (1-\theta)\,\boldsymbol{\eta}_k + \rho_k\boldsymbol{\varphi}(\mathbf{x}_{k+1}). \tag{38}$$

end for

In this subsection, we prove the convergence result stated in Theorem 1. We begin by recalling several definitions and pieces of notation. The augmented Lagrangian associated with (**P**) is

$$\hat{f}_\rho(\mathbf{x};\boldsymbol{\eta}) = f(\mathbf{x}) + \boldsymbol{\eta}^{\top}\boldsymbol{\varphi}(\mathbf{x}) + \frac{\rho}{2}\left\|\boldsymbol{\varphi}(\mathbf{x})\right\|^2. \tag{39}$$

The corresponding dual problem is

$$\max_{\boldsymbol{\eta}\in\mathbb{R}^p} d(\boldsymbol{\eta}), \tag{40}$$

where the dual function is defined by

$$d(\boldsymbol{\eta}) := \min_{\mathbf{x}\in\mathcal{C}} \hat{f}_\rho(\mathbf{x};\boldsymbol{\eta}). \tag{41}$$

It is well known that $d(\boldsymbol{\eta})$ is differentiable when Ψ is smooth, with gradient $\nabla d(\boldsymbol{\eta}) = \boldsymbol{\phi}(\mathbf{x}(\boldsymbol{\eta}))$, where $\mathbf{x}(\boldsymbol{\eta})$ denotes the minimizer in (41); see, e.g., (Bertsekas, 1997). The standard PALM(θ) method is summarized in Algorithm 6. Throughout this subsection, we focus on the case $\theta = 0$.

We assume that the strong duality holds for problem (**P**), which is formally stated as follows.

Assumption 1 (Strong duality)**.** There exists a Lagrange multiplier $\boldsymbol{\eta}^* < \infty$ such that the optimal value of problem (**P**) is equal to $d(\boldsymbol{\eta}^*)$.

This assumption imposes strong duality between the augmented-Lagrangian dual problem and the primal problem. In the convex setting, it holds under mild conditions. For example, if problem (**P**) is convex and Slater's condition is satisfied, then strong duality holds for the augmented-Lagrangian dual formulation; see, e.g., (Deng et al., 2025).

The following key lemma shows that, under a suitable parameter choice, the trajectories of PALM and Hom-PALM coincide. This lemma is central to our convergence analysis of Hom-PALM.

Lemma E.4. *Suppose* $\Psi(\mathbf{z}_1) = \mathbf{x}_1$ *and* $\boldsymbol{\gamma}_1 = \boldsymbol{\eta}_1$. *Let* $\{(\mathbf{z}_k, \boldsymbol{\gamma}_k)\}_{k\geq 1}$ *be any sequence generated by Hom-PALM*($\theta = 0$) *in Algorithm 1 with proximal operator* $\frac{1}{2}\|\Psi(\mathbf{z}) - \Psi(\mathbf{z}_k)\|^2$. *Define* $\mathbf{x}_k = \Psi(\mathbf{z}_k)$, $\alpha_k = \beta_k$, *and* $\boldsymbol{\eta}_k = \boldsymbol{\gamma}_k$. *Then* $\{(\mathbf{x}_k, \boldsymbol{\eta}_k)\}_{k\geq 1}$ *is a trajectory of PALM*($\theta = 0$) *in Algorithm 6, namely,*

$$\mathbf{x}_{k+1} \in \arg\min_{\mathbf{x}\in\mathcal{C}} \left\{ f(\mathbf{x}) + \boldsymbol{\eta}_k^\top \boldsymbol{\varphi}(\mathbf{x}) + \frac{\rho_k}{2}\|\boldsymbol{\varphi}(\mathbf{x})\|^2 + \frac{\alpha_k}{2}\|\mathbf{x} - \mathbf{x}_k\|^2 \right\}, \tag{42}$$

$$\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_k + \rho_k \boldsymbol{\varphi}(\mathbf{x}_{k+1}). \tag{43}$$

Proof. It follows from Lemma D.10 that $\mathbf{x}_{k+1} = \Psi(\mathbf{z}_{k+1})$ is a minimizer of $\hat{f}_\rho(\mathbf{x}; \boldsymbol{\eta}_k)$ for each $k \in \mathbb{N}$. The dual update $\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_k + \rho_k \boldsymbol{\varphi}(\mathbf{x}_{k+1})$ then follows directly by induction. □

Notation. In this subsection, we consider Hom-PALM with $\theta = 0$ and proximal operator $\frac{1}{2}\|\Psi(\mathbf{z}) - \Psi(\mathbf{z}_k)\|^2$. Let $\mathbb{P}_k(\mathbf{x})$ denote the proximal augmented-Lagrangian function at the k-th iteration of PALM in Algorithm 6, i.e., $\mathbb{P}_k(\mathbf{x}) := \mathbb{P}^{\theta=0}_{\rho_k,\alpha_k}(\mathbf{x}; \boldsymbol{\eta}_k)$ for problem (**P**), and let $\mathbb{P}_k^* := \min_{\mathbf{x}\in\mathcal{C}} \mathbb{P}_k(\mathbf{x})$. Similarly, let $\mathbb{H}_k(\mathbf{z})$ denote the proximal augmented-Lagrangian function at the k-th iteration of Hom-PALM in Algorithm 1. Finally, let $\mathcal{A}_t(\mathbf{x}; \mathbb{P}_k)$ denote the output obtained after running t iterations of an inner algorithm $\mathcal{A}$ on problem (37), initialized at $\mathbf{x}$. We consider inner solvers $\mathcal{A}$ satisfying the following condition.

Assumption 2 (Linearly convergent inner solver)**.** For any outer iteration $k \in \{0, 1, \ldots\}$ of proximal ALM, there exists $M_k \geq 1$ such that, for any $\mathbf{x}$,

$$\mathbb{E}\left[\mathbb{P}_k(\mathcal{A}_t(\mathbf{x}, \mathbb{P}_k)) - \mathbb{P}_k^\star \mid \mathcal{F}_{t-1}\right] \leq 2^{-\lfloor t/M_k \rfloor}\left(\mathbb{P}_k(\mathbf{x}) - \mathbb{P}_k^\star\right), \tag{44}$$

where $\mathcal{F}_t$ denotes the σ-algebra generated by $\{\mathbf{x}_s : s \leq t\} \cup \{\boldsymbol{\eta}_s : s \leq t+1\}$.

Remark. Since each subproblem is strongly convex, using a linearly convergent inner solver is well justified. Indeed, many first-order methods, including projection-based methods, are known to converge linearly for strongly convex problems; see, e.g., (Boyd et al., 2004; Beck, 2017).

Next, we introduce the definition of a stationary point considered in (Li & Qu, 2021).

Definition E.5 (Expected ϵ-optimal solution (Li & Qu, 2021))**.** Let $\mathbf{x}_k$ be the k-th iterate of an algorithm initialized at a randomized point $\mathbf{x}_0$. For any $\epsilon > 0$, we call $\mathbf{x}_k$ an expected ϵ-optimal solution of a convex optimization (**P**) if: $\mathbb{E}\|\boldsymbol{\varphi}(\mathbf{x}_k)\| \leq \epsilon$, $\mathbb{E}\|[\mathbf{g}(\mathbf{x}_k)]_+\| \leq \epsilon$, $\mathbb{E}[f(\mathbf{x}_k) - f(\mathbf{x}^*)] \leq \epsilon$.

The following theorem from (Li & Qu, 2021) establishes the convergence rate of standard PALM in Algorithm 6.

Theorem 4 (Convergence of proximal ALM, (Li & Qu, 2021))**.** *Let Assumption 1 hold. Consider PALM*($\theta = 0$) *in Algorithm 6 with parameters satisfying*

$$\rho_k \alpha_k = 1, \qquad \alpha_{k+1} = c_1 \alpha_k, \qquad \epsilon_{k+1} = c_2 \epsilon_k,$$

for some $c_1 \in (1/2, 1)$ *and* $c_2 \in (0, 1)$, *with initial tolerance* ϵ_0. *Suppose that the number* m_k *of inner iterations satisfies*

$$2\epsilon_k + R_k \leq 2^{\lfloor m_{k+1}/M_{k+1} \rfloor}\epsilon_{k+1}/2,$$

where R_k is a constant depending on α_k and $\boldsymbol{\eta}_k$, defined by

$$\begin{aligned} R_k :=& \alpha_k \left\|\boldsymbol{\eta}_{k+1} - \boldsymbol{\eta}_k\right\|_2 + \frac{\alpha_k - \alpha_{k+1}}{2} \left\|\Lambda\left(\boldsymbol{\varphi}\left(\mathbf{x}_k\right); \boldsymbol{\eta}_{k+1}, \alpha_{k+1}\right) - \boldsymbol{\eta}_{k+1}\right\|^2 \\ &+ \frac{\alpha_k^2}{2\alpha_{k+1} - \alpha_k} \left\|\mathbf{x}_{k-1} - \mathbf{x}_k\right\|_2 + \left\|\boldsymbol{\eta}_{k+1} - \boldsymbol{\eta}_k\right\| \delta_k, \\ \delta_k :=& \left\|\alpha_k \boldsymbol{\eta}_k - \alpha_{k+1}\boldsymbol{\eta}_{k+1}\right\|, \\ \Lambda(\mathbf{u}; \boldsymbol{\eta}, \alpha) :=& \arg\max_{\mathbf{v} \in \mathbb{R}^d} \left\{\langle \mathbf{v}, \mathbf{u}\rangle - \frac{\alpha}{2}\|\mathbf{v} - \boldsymbol{\eta}\|^2\right\} = \boldsymbol{\eta} + \mathbf{u}/\alpha. \end{aligned}$$

If there exist constants $\varsigma \geq 0$, $\omega > 0$, and $\ell > 0$ such that

$$M_k \leq \frac{\omega}{\alpha_k^\ell} + \varsigma, \quad \forall k \geq 1,$$

then, for any $\epsilon \leq \epsilon_0$, an ϵ-optimal solution in the sense of Definition E.5 can be obtained within a total expected number of inner iterations bounded by

$$\widetilde{\mathcal{O}}\left(\frac{\omega + \varsigma}{\epsilon^\ell}\right), \tag{45}$$

where the hidden constants depend on c_1, c_2, α_0, and ϵ_0, but are independent of ϵ.

Remark. The work (Li & Qu, 2021) considers a convex composite optimization problem

$$\min_{\mathbf{x}} f\left(\mathbf{x}\right) + g\left(\mathbf{x}\right) + h_1\left(p_1\left(\mathbf{x}\right)\right) + h_2\left(p_2\left(\mathbf{x}\right)\right). \tag{46}$$

If g is the indicator function of the inequality constraints, $h_1 = 0$, and h_2 is the indicator function of the equality constraints, then this problem reduces to the general formulation (**P**) considered here.

To solve each Hom-PALM subproblem (1), we employ the proposed Hom-APG method. The following corollary follows from Theorem 3.

Corollary 5. *Suppose that Hom-APG is used to solve the subproblem* (1) *of Hom-PALM*(0) *in Algorithm 1, with $\mathcal{P}\left(\mathbf{z}, \mathbf{z}_k\right) = \frac{1}{2}\left\|\Psi\left(\mathbf{z}\right) - \Psi\left(\mathbf{z}_k\right)\right\|^2$. Assume that all iterates $\mathbf{z}_k$ lie in a compact region $\mathcal{M}$ containing the unit ball $\mathcal{B}$; this is reasonable because each iteration includes a projection step. Under the same setting as in Theorem 3, the convergence rate satisfies*

$$\mathbb{P}_k\left(\mathbf{x}_t\right) - \mathbb{P}_k^* \leq \mathcal{O}\left(\exp\left(-c\sqrt{\frac{\mu_{\mathbb{P}}}{L_{\mathbb{P}}}}t\right)\right), \tag{47}$$

where $\mathbf{x}_t = \Psi(\mathbf{z}_t)$, c is a constant independent of $\mathbb{P}$, and $\mu_{\mathbb{P}}$ is the strong-convexity parameter of $\mathbb{P}_k\left(\mathbf{x}\right) = \hat{f}\left(\mathbf{x}\right) + \frac{\alpha_k}{2}\left\|\mathbf{x} - \mathbf{x}_k\right\|^2$; that is, $\mu_{\mathbb{P}} = \mu_f + \alpha_k$. Moreover, $L_{\mathbb{P}}$ is the constant defined as L_{eff} in Theorem 3, with L_f and $L_{f,0}$ replaced by $L_{\mathbb{P}}$ and $L_{\mathbb{P},0}$, respectively, where these two quantities denote the smoothness parameter and the Lipschitz constant of $\mathbb{P}$.

Remark. Here, we do not require $\mu_f > 0$, since the subproblem is naturally hidden strongly convex due to the additional proximal term.

Theorem 1. *Suppose that strong duality holds for problem* (**P**). *Let the hyperparameters, including α_k and ρ_k, be defined as in Theorem 4. Consider Hom-PALM*(0) *in Algorithm 1, equipped with the proximal operator $\mathcal{P}\left(\mathbf{z}, \mathbf{z}_k\right) = \frac{1}{2}\left\|\Psi\left(\mathbf{z}\right) - \Psi\left(\mathbf{z}_k\right)\right\|^2$ and $\beta_k = \alpha_k$. Then Hom-PALM obtains an ϵ-optimal solution (Def. E.5) of problem* (**P**) *within $\tilde{\mathcal{O}}\left(1/\sqrt{\epsilon}\right)$ iterations when f is strongly convex, and within $\tilde{\mathcal{O}}\left(1/\epsilon\right)$ iterations when f is convex.*

Proof. We consider the trajectory $\mathbf{x}_k = \Psi\left(\mathbf{z}_k\right)$, where $\mathbf{z}_k$ is generated by Hom-PALM within a compact domain $\mathcal{M}$. It follows from Corollary 5 that the constant M_k in Assumption 2 satisfies

$$\begin{aligned} M_k &\leq c\sqrt{\frac{L_{\mathbb{P}}}{\mu_{\mathbb{P}}}} = c\sqrt{\frac{L_\Psi L_{\mathbb{P},0} + \kappa_2^2 L_{\mathbb{P}}}{\left(\kappa_2^2 + L_\Psi \operatorname{diam}\left(\mathcal{M}\right)\right)\left(\mu_f + \alpha_k\right)}} \\ &= c'\sqrt{\frac{\left(L_f + \rho_k\left\|\mathbf{A}\right\|^2\right)\kappa_2^2 + L_{\hat{f},0} L_\Psi + \alpha_k\left(\kappa_2^2 + L_\Psi^2 \operatorname{diam}\left(\mathcal{M}\right)\right)}{\mu_f + \alpha_k}}, \end{aligned} \tag{48}$$

where c and c' are constants independent of α_k, and $L_{\hat{f},0}$ is the Lipschitz constant of $\hat{f}$ defined in Lemma D.2.

Next, the hyperparameter definitions in Theorem 4 imply $\rho_k \alpha_k = 1$. Therefore, M_k is of order $1/\sqrt{\alpha_k}$ when $\mu_f > 0$ and of order $1/\alpha_k$ when $\mu_f = 0$, which yields the final convergence-rate bounds by Theorem 4. □

E.3 Non-Convex Case: Non-Convex Objective with Linear Equality

In this section, *we consider Hom-PALM(θ) with $\theta \in (0,1)$ and the quadratic proximal term* $\mathcal{P}(\mathbf{z}, \mathbf{z}_k) = \frac{1}{2}\|\mathbf{z} - \mathbf{z}_k\|^2$. Moreover, in problem (**P**), the objective f may be nonconvex, while the equality constraints are still assumed to be linear. We first make the following additional assumptions.

Assumption 3. We suppose that

- for problem (**H**), h is ℓ_h-weakly convex; and
- Slater's condition holds: there exists $\mathbf{x}^\circ \in \operatorname{int}(\mathcal{C})$ such that $\mathbf{A}\mathbf{x}^\circ = \mathbf{b}$.

Next, we can provide the standard convergence rate of PALM(θ) from (Melo et al., 2024).

Definition E.6 (Approximate stationary point)**.** A point $\mathbf{x}_\epsilon$ is an ϵ -stationary point of problem (**P**) if there are $\boldsymbol{\xi} \in \partial \mathbb{I}_{\mathcal{C}}(\mathbf{x}_\epsilon), \boldsymbol{\lambda} \in \mathbb{R}^m_+$, and $\boldsymbol{\eta}_\epsilon \in \mathbb{R}^p$ such that

$$\|\nabla f(\mathbf{x}_\epsilon) + \mathrm{J}_{\boldsymbol{\varphi}}(\mathbf{x}_\epsilon)\boldsymbol{\eta}_\epsilon + \boldsymbol{\xi}\| \le \epsilon, \|\boldsymbol{\varphi}(\mathbf{x}_\epsilon)\| \le \epsilon, \|[\mathbf{g}(\mathbf{x}_\epsilon)]_+\| \le \epsilon \tag{49}$$

We recall the composite subproblem form used by ACG and AdapAPG:

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := u(\mathbf{x}) + r(\mathbf{x}), \tag{50}$$

where $u : \mathbb{R}^d \to \mathbb{R}$ is μ_u-strongly convex and L_u-smooth, and $r : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lower-semicontinuous convex function. In the PALM subproblem, r is the indicator function of the inequality-constrained set and u is the smooth proximal augmented-Lagrangian model. The ACG routine used in the following theorem is given in Algorithm 7.

Algorithm 7 ACG

Input: functions u and r with $(\mu_u, L_u) \in \mathbb{R}^2_+$, initial point $\mathbf{y}_0 \in \operatorname{dom} r$
Initialize: $\mathbf{x}_0 \leftarrow \mathbf{y}_0$, $A_0 \leftarrow 0$, $\tau_0 \leftarrow 1$, $\zeta \leftarrow 1/(L_u - \mu_u)$, $j \leftarrow 0$
repeat
  Compute iterates
  $a_j \leftarrow \frac{\zeta\tau_j + \sqrt{(\zeta\tau_j)^2 + 4\tau_j A_j}}{2}$
  $A_{j+1} \leftarrow A_j + a_j$
  $\tilde{\mathbf{x}}_j \leftarrow \frac{A_j \mathbf{y}_j + a_j \mathbf{x}_j}{A_{j+1}}$
  $\tau_{j+1} \leftarrow \tau_j + \mu_u a_j$
  $\mathbf{y}_{j+1} \leftarrow \arg\min_{\mathbf{y}\in\mathbb{R}^n} \left\{ u(\tilde{\mathbf{x}}_j) + \langle \nabla u(\tilde{\mathbf{x}}_j), \mathbf{y} - \tilde{\mathbf{x}}_j \rangle + r(\mathbf{y}) + \frac{L_u}{2}\|\mathbf{y} - \tilde{\mathbf{x}}_j\|^2 \right\}$
  $\mathbf{x}_{j+1} \leftarrow \frac{1}{\tau_{j+1}} \left[\frac{a_j}{\zeta}(\mathbf{y}_{j+1} - \tilde{\mathbf{x}}_j) + \mu_u a_j \mathbf{y}_{j+1} + \tau_j \mathbf{x}_j \right]$
  Compute stopping criteria quantities
  $\mathbf{u}_{j+1} \leftarrow \mu_u(\mathbf{y}_{j+1} - \mathbf{x}_{j+1}) + \frac{\mathbf{x}_0 - \mathbf{x}_{j+1}}{A_{j+1}}$
  $\eta_{j+1} \leftarrow \frac{1}{2A_{j+1}} \left(\|\mathbf{x}_0 - \mathbf{y}_{j+1}\|^2 - \tau_{j+1}\|\mathbf{x}_{j+1} - \mathbf{y}_{j+1}\|^2 \right)$
  $j \leftarrow j + 1$
until $\|\mathbf{u}_j\|^2 + 2\eta_j \le \sigma_0^2 \|\mathbf{y}_0 - \mathbf{y}_j + \mathbf{u}_j\|^2$
Output: $(\mathbf{y}, \mathbf{u}, \eta) \leftarrow (\mathbf{y}_j, \mathbf{u}_j, \eta_j)$

Theorem 6 (Convergence of PALM(θ), (Melo et al., 2024))**.** *Let f be ℓ_f-weakly convex, and suppose that Slater's condition holds for problem* (**P**)*. Consider PALM(θ) with $\theta \in (0,1)$ and parameters*

$$\tau_\theta := \min\left\{ \frac{1}{2}, \frac{\theta}{88(1-\theta)} \right\}, \quad \alpha_k := \frac{\tau_\theta}{\ell_f}, \quad \rho_{k+1} = 2\rho_k, \quad \forall k = 1, 2, \cdots. \tag{51}$$

Suppose that the accelerated composite gradient (ACG) method (Florea & Vorobyov, 2018; Kong et al., 2021; Melo et al., 2024) (outlined in Alg. 7) is used as an inner solver for subproblem (37)*. Then PALM(θ) converges to an ϵ-approximate stationary point (Def. E.6) of problem* (**P**) *in $\tilde{\mathcal{O}}\left(\epsilon^{-2.5}\right)$, where the hidden constants depend on $\theta, L_f, \ell_f, \|\mathbf{A}\|$, and ρ_1, but are independent of ϵ.*

Remark. (i) The work (Melo et al., 2024) considers the composite optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}) + r(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b},$$

and defines an ϵ-approximate stationary solution $(\mathbf{x}, \mathbf{v}, \boldsymbol{\eta})$ via the condition

$$\mathbf{v} \in \nabla f(\mathbf{z}) + \partial r(\mathbf{z}) + \mathbf{A}^\top \boldsymbol{\eta}, \quad \|\mathbf{v}\| \le \epsilon, \quad \|\mathbf{A}\mathbf{x} - \mathbf{b}\| \le \epsilon.$$

When r is the indicator function of an inequality-constrained set, e.g., $\mathcal{C}$ in our setting, this definition reduces to Definition E.6. We refer the reader to standard textbooks, e.g., (Boyd et al., 2004; Burden & Faires, 1997; Beck, 2017), for details.

(ii) The work (Melo et al., 2024) assumes that $r(\mathbf{x})$ is a closed proper convex function satisfying

$$\partial r(\mathbf{x}) \subseteq \bar{B}(0, m_r) + N_{\mathcal{R}}(\mathbf{z}), \quad \forall \mathbf{z} \in \mathcal{R} := \operatorname{dom} r,$$

for some $m_r > 0$. This assumption holds naturally when r is the indicator function of a convex inequality-constrained set.

Based on the theorem, one can directly derive the generalized convergence rate of Hom-PALM(θ) as follows.

Theorem 7 (Convergence of Hom-PALM(θ), nonconvex objective)**.** *Let Assumption 3 hold. Let the homeomorphic mapping Ψ be the gauge mapping associated with an interior point $\mathbf{x}^\circ$ satisfying $\mathbf{A}\mathbf{x}^\circ = \mathbf{b}$. In this case, one can set $\boldsymbol{\phi}(\mathbf{z}) = A\mathbf{z}$ for Hom-PALM(θ). Consider Hom-PALM(θ) for some $\theta \in (0,1)$ with parameters*

$$\tau_\theta := \min\left\{\frac{1}{2}, \frac{\theta}{88(1-\theta)}\right\}, \quad \beta_k := \frac{\tau_\theta}{\ell_f}, \quad \rho_{k+1} = 2\rho_k, \quad \forall k = 1, 2, \cdots. \tag{52}$$

Suppose that the accelerated composite gradient (ACG) method (Florea & Vorobyov, 2018; Kong et al., 2021; Melo et al., 2024) (outlined in Alg. 7) is used as an inner solver for subproblem (1)*. Then Hom-PALM(θ) converges to an ϵ-approximate stationary point (Def. E.6) of problem* (**H**) *in $\tilde{\mathcal{O}}\left(\epsilon^{-2.5}\right)$, where the hidden constants depend on $\theta, L_h, \ell_h, \|\mathbf{A}\|$, and ρ_1, but are independent of ϵ.*

Proof. This is a direct corollary of Theorem 6. It suffices to note the following facts. First, for linear constraints of the form $\mathbf{A}\mathbf{x} = \mathbf{b}$, the gauge transformation Ψ preserves linearity when the interior point lies in the kernel $\mathcal{K}$ rather than merely in the feasible set $\mathcal{C}$. Specifically, under the change of variables $\mathbf{x} = \Psi(\mathbf{z})$, the constraint is transformed as

$$\{\boldsymbol{\varphi}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{0}\} \quad \longleftrightarrow \quad \{\boldsymbol{\phi}(\mathbf{z}) = \mathbf{0}\} = \{\mathbf{A}\mathbf{z} = 0\}. \tag{53}$$

See Proposition B.3 for details. Second, since Slater's condition holds for problem (**P**), it also holds for problem (**H**). Therefore, Theorem 6 applies directly. □

Remark. (i) The assumption that h is weakly convex is mild. In particular, if f is weakly convex, then h is weakly convex over any compact region. One can refer to Appendix E.4 for details. (ii) This theorem leverages the properties of the gauge mapping. Specifically, the gauge mapping transforms linear equality constraints into linear equality constraints that admit a special interior point; see Proposition B.3 for details. Moreover, the result extends to general ball-homeomorphic mappings under similar assumptions. Specifically, one can apply Hom-PALM with the proximal term $\frac{1}{2}\|\Psi(\mathbf{z}) - \Psi(\mathbf{z}_k)\|^2$ and, analogous to Lemma E.4, analyze the trajectory in the original space to obtain a comparable convergence rate with Theorem 6. (iii) The ACG algorithm (Melo et al., 2024) for problem (**P**) involves projection onto $\mathcal{C}$ at each iteration. In contrast, when ACG is applied to the Hom-PALM subproblem, it becomes projection-free since the inequality constraint reduces to a unit ball.

E.4 Non-Convex Case: Non-Convex Objective with Non-Linear Equality

Algorithm 8 Proximal-Point Penalty (PPP) Method (Lin et al., 2022)

Input: initial point $\mathbf{x}_0$, problem (**P**), proximal parameters $\{\beta_k\}$, penalty parameters $\{\rho_k\}$, target optimality measures for subproblems $\{\epsilon_k\}$, and the number of iterations K
for $k = 0$ **to** K **do**
Primal update:

$$\mathbf{x}_{k+1} := \arg\min_{\mathbf{x}\in\mathcal{C}} \hat{f}_k(\mathbf{x}) := f(\mathbf{x}) + \frac{\rho_k}{2}\|\boldsymbol{\varphi}(\mathbf{x})\|^2 + \frac{\beta_k}{2}\|\mathbf{x}-\mathbf{x}_k\|^2, \tag{54}$$

such that

$$\min_{\boldsymbol{\xi}\in\partial\mathbb{I}_{\mathcal{C}}(\mathbf{x}_{k+1})} \left\|\nabla\hat{f}_k(\mathbf{x}_{k+1}) + \boldsymbol{\xi}\right\| \le \epsilon_k. \tag{55}$$

Dual update:

$$\boldsymbol{\eta}_{k+1} = \rho_k \boldsymbol{\varphi}(\mathbf{x}_{k+1}). \tag{56}$$

end for
Output: $\mathbf{x}_K$

In this section, we consider problem (**P**) with a nonconvex objective and nonlinear equality constraints $\{\boldsymbol{\varphi}(\mathbf{x}) = 0\}$. We apply Hom-PALM($\theta$) with $\theta = 1$ using the quadratic proximal term $\mathcal{P}(\mathbf{z}, \mathbf{z}_k) = \frac{1}{2}\|\mathbf{z}-\mathbf{z}_k\|^2$. With this choice, Hom-PALM reduces to the proximal-point penalty (PPP) method introduced in (Lin et al., 2022), which is presented in Algorithm 8.

Next, we state the **additional** assumptions from (Lin et al., 2022) used to establish convergence of the PPP method in the nonconvex case.

Assumption 4. We impose the following assumptions.

(1) The proximal operator $\mathrm{prox}_{\mathbb{I}_{\mathcal{C}}} := \Pi_{\mathcal{C}}$ can be computed efficiently, e.g., in closed form.

(2) f and $\boldsymbol{\varphi}_i$ are weakly convex for all $i \in [p]$.

(3) A variant of the KL condition holds: there exists a constant $\nu > 0$ such that, for any $\mathbf{x} \in \mathcal{C}$,

$$\nu\|\boldsymbol{\varphi}(\mathbf{x})\| \le \mathrm{dist}\left(\mathrm{J}_{\boldsymbol{\varphi}}(\mathbf{x})\,\boldsymbol{\varphi}(\mathbf{x}), -\mathrm{N}_{\mathcal{C}}(\mathbf{x})\right). \tag{57}$$

Remark. The other assumptions in (Lin et al., 2022) naturally hold because we consider the indicator function $\mathbb{I}_{\mathcal{C}}$.

For the PPP subproblem, r in (50) is the indicator function of $\mathcal{C}$, and u corresponds to $\hat{f}_k$ in (54). The AdapAPG routine used in the following theorem is given in Algorithm 11; the line-search subroutines are included for completeness. Given $\mathbf{w} \in \mathbb{R}^d$ and $L > 0$, define

$$\psi_L(\mathbf{w};\mathbf{x}) := u(\mathbf{w}) + \nabla u(\mathbf{w})^\top(\mathbf{x}-\mathbf{w}) + \frac{L}{2}\|\mathbf{x}-\mathbf{w}\|^2 + r(\mathbf{x}). \tag{58}$$

The proximal gradient step, proximal gradient mapping, first-order suboptimality measure, and local Lipschitz constant are

$$T_L(\mathbf{w}) := \arg\min_{\mathbf{x}\in\mathbb{R}^d} \psi_L(\mathbf{w};\mathbf{x}) = \mathrm{prox}_{L^{-1}r}\left(\mathbf{w} - L^{-1}\nabla u(\mathbf{w})\right), \tag{59}$$

$$g_L(\mathbf{w}) := L\left(\mathbf{w} - T_L(\mathbf{w})\right), \tag{60}$$

$$\omega(\mathbf{x}) := \min_{\boldsymbol{\xi}\in\partial r(\mathbf{x})} \|\nabla u(\mathbf{x}) + \boldsymbol{\xi}\|, \tag{61}$$

$$S_L(\mathbf{w}) := \frac{\|\nabla u(T_L(\mathbf{w})) - \nabla u(\mathbf{w})\|}{\|T_L(\mathbf{w}) - \mathbf{w}\|} \le L_u. \tag{62}$$

These quantities satisfy

$$\omega(T_L(\mathbf{w})) \leq \left(1 + \frac{S_L(\mathbf{w})}{L}\right) \|g_L(\mathbf{w})\|. \tag{63}$$

Algorithm 9 $\{\mathbf{x}_{t+1}, M_t, \mathbf{p}_t, S_t\} \leftarrow \text{LineSearch}\,(u, r, \mathbf{x}_t, L_t)$

Input: current point $\mathbf{x}_t$, initial step size estimate L_t, functions u and r
Choose: $\gamma_{\text{inc}} > 1$
$L \leftarrow L_t/\gamma_{\text{inc}}$
repeat
  $L \leftarrow L\gamma_{\text{inc}}$
  $\mathbf{x}_{t+1} \leftarrow T_L(\mathbf{x}_t)$
until $F(\mathbf{x}_{t+1}) \leq \psi_L(\mathbf{x}_t; \mathbf{x}_{t+1})$
$M_t \leftarrow L$
$\mathbf{p}_t \leftarrow M_t(\mathbf{x}_t - \mathbf{x}_{t+1})$
$S_t \leftarrow S_L(\mathbf{x}_t)$
Output: $\mathbf{x}_{t+1}, M_t, \mathbf{p}_t, S_t$

Algorithm 10 $\{\mathbf{x}_{t+1}, M_t, \alpha_t, \mathbf{p}_t, S_t\} \leftarrow \text{AccelLineSearch}\,(u, r, \mathbf{x}_t, \mathbf{x}_{t-1}, L_t, \mu, \alpha_{t-1})$

Input: current point $\mathbf{x}_t$, previous point $\mathbf{x}_{t-1}$, initial step size estimate L_t, strong convexity parameter μ, previous momentum parameter α_{t-1}, functions u and r
Choose: $\gamma_{\text{inc}} > 1$
$L \leftarrow L_t/\gamma_{\text{inc}}$
repeat
  $L \leftarrow L\gamma_{\text{inc}}$
  $\alpha_t \leftarrow \sqrt{\frac{\mu}{L}}$
  $\mathbf{w}_t \leftarrow \mathbf{x}_t + \frac{\alpha_t(1-\alpha_{t-1})}{\alpha_{t-1}(1+\alpha_t)}(\mathbf{x}_t - \mathbf{x}_{t-1})$
  $\mathbf{x}_{t+1} \leftarrow T_L(\mathbf{w}_t)$
until $F(\mathbf{x}_{t+1}) \leq \psi_L(\mathbf{w}_t; \mathbf{x}_{t+1})$
$M_t \leftarrow L$
$\mathbf{p}_t \leftarrow M_t(\mathbf{w}_t - \mathbf{x}_{t+1})$
$S_t \leftarrow S_L(\mathbf{w}_t)$
Output: $\mathbf{x}_{t+1}, M_t, \alpha_t, \mathbf{p}_t, S_t$

Algorithm 11 $\{\hat{\mathbf{x}}, \hat{M}, \hat{\mu}\} \leftarrow \text{AdapAPG}\,(u, r, \mathbf{x}_{\text{ini}}, L_{\text{ini}}, \mu_0, \hat{\varepsilon})$

Input: initial point $\mathbf{x}_{\text{ini}}$, initial step size estimate L_{ini}, initial strong convexity estimate μ_0, target accuracy $\hat{\varepsilon}$, functions u and r
Choose: $L_{\min} \in [\mu_0, L_{\text{ini}}]$, $\gamma_{\text{dec}} \geq 1$, $\gamma_{\text{sc}} > 1$, $\theta_{\text{sc}} \in (0,1)$
$\{\mathbf{x}_0, M_{-1}, \mathbf{p}_{-1}, S_{-1}\} \leftarrow \text{LineSearch}(u, r, \mathbf{x}_{\text{ini}}, L_{\text{ini}})$
$\mathbf{x}_{-1} \leftarrow \mathbf{x}_0$, $L_0 \leftarrow M_{-1}$, $\mu \leftarrow \mu_0$, $\alpha_{-1} \leftarrow 1$, $\tau_0 \leftarrow 1$, $t \leftarrow 0$
repeat
  $\{\mathbf{x}_{t+1}, M_t, \alpha_t, \mathbf{p}_t, S_t\} \leftarrow \text{AccelLineSearch}(u, r, \mathbf{x}_t, \mathbf{x}_{t-1}, L_t, \mu, \alpha_{t-1})$
  $\tau_{t+1} \leftarrow \tau_t(1-\alpha_t)$
  if $\|\mathbf{p}_t\| \leq \theta_{\text{sc}}\|\mathbf{p}_{-1}\|$ **then**
    $\mathbf{x}_0 \leftarrow \mathbf{x}_{t+1}$, $\mathbf{x}_{-1} \leftarrow \mathbf{x}_{t+1}$, $L_0 \leftarrow M_t$, $\mathbf{p}_{-1} \leftarrow \mathbf{p}_t$, $M_{-1} \leftarrow M_t$, $S_{-1} \leftarrow S_t$
    $t \leftarrow 0$
  else
    if $2\sqrt{2\tau_t}\frac{M_t}{\mu}\left(1 + \frac{S_{-1}}{M_{-1}}\right) \leq \theta_{\text{sc}}$ **then**
      $\mu \leftarrow \mu/\gamma_{\text{sc}}$
      $t \leftarrow 0$
    else
      $L_{t+1} \leftarrow \max\{L_{\min}, M_t/\gamma_{\text{dec}}\}$
      $t \leftarrow t+1$
    end if
  end if
until $\omega(\mathbf{x}_{t+1}) \leq \hat{\varepsilon}$
Output: $\hat{\mathbf{x}} \leftarrow \mathbf{x}_{t+1}$, $\hat{M} \leftarrow M_t$, $\hat{\mu} \leftarrow \mu$

We next state the convergence result for the PPP method, which follows from Theorem 3 in (Lin et al., 2022).

Theorem 8 (Convergence of PPP)**.** *Suppose that Assumption 4 holds for problem* (**P**)*. Consider PPP in Algorithm 8 with parameters*

$$\beta_k = 2B_k, \quad \rho_k = \rho\,(k+1)^{1/3}, \quad \epsilon_k = \frac{1}{\rho\,(1+k)^{4/3}},$$

where $\rho > 0$ is a constant and $f(\mathbf{x}) + \frac{\rho_k}{2}\|\boldsymbol{\varphi}(\mathbf{x})\|^2$ is B_k-weakly convex. If subproblem (54) *is solved by the adaptive accelerated proximal gradient (AdapAPG) method (Lin & Xiao, 2014; Lin et al., 2022) (outlined in Alg. 11), then PPP returns an ϵ-solution $\mathbf{x}_\epsilon$ of problem* (**P**)*, as defined in Def. E.6, in $\tilde{\mathcal{O}}\left(\epsilon^{-3}\right)$, where the hidden constants depend on $\rho_0, \beta_0, \ell_{\boldsymbol{\varphi}_i}, \ell_f$ and $\max_{\mathbf{x}\in\mathcal{C}}\{\|\nabla f(\mathbf{x})\|, f(\mathbf{x})\}$, $\max_{\mathbf{x}\in\mathcal{C}}\{\|\nabla\boldsymbol{\varphi}(\mathbf{x})\|, \boldsymbol{\varphi}(\mathbf{x})\}$.*

Remark. (i) The stationary-point definition in Definition E.6 is the weak version defined in (Lin et al., 2022). (ii) Item (1) in Assumption 4 is generally difficult to satisfy. However, for problem (**H**), this assumption is easily satisfied because $\Pi_{\mathcal{B}}$ admits a closed form. Thus, to apply the PPP method to problem (**H**), it remains to verify whether the other items in Assumption 4 hold for problem (**H**). (iii) Basic properties, such as smoothness of h, still hold by Lemma D.1. In the following, we show that the remaining requirements are also satisfied under Assumption 4.

Lemma E.7. *Suppose that Assumption 4 holds for problem* (**P**)*. Then the following properties hold for problem* (**H**)*.*

(i) $h = f \circ \Psi$ and $\boldsymbol{\phi} = \boldsymbol{\varphi} \circ \Psi$ are weakly convex over any bounded region $\mathcal{M}$.

(ii) There exists a constant $\mu > 0$ such that, for any $\mathbf{z} \in \mathcal{B}$, the following inequality holds with $\mu = \nu \cdot \kappa_1$:

$$\mu\,\|\boldsymbol{\phi}(\mathbf{z})\| \leq \operatorname{dist}\left(\mathrm{J}_{\boldsymbol{\phi}}(\mathbf{z})\,\boldsymbol{\phi}(\mathbf{z}), -\mathrm{N}_{\mathcal{B}}(\mathbf{z})\right). \tag{64}$$

Remark. Although h and $\boldsymbol{\phi}$ may not be weakly convex globally, the PPP iterates are approximately feasible because the PPP subproblem solver is a proximal-point-based method involving the indicator function. Thus, it is sufficient that f and $\boldsymbol{\phi}$ are weakly convex on a bounded region, e.g., on $\mathcal{C}$.

Proof. We prove the two claims separately.

(1) It suffices to show that h is weakly convex; the proof for $\boldsymbol{\phi}$ is analogous.

Since f is differentiable and ℓ_f-weakly convex, we have:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle - \frac{\ell_f}{2}\|\mathbf{y} - \mathbf{x}\|^2;$$

for $\mathbf{x} = \Psi(\mathbf{z})$ and $\mathbf{y} = \Psi(\mathbf{u})$, that is:

$$f(\Psi(\mathbf{u})) \geq f(\Psi(\mathbf{z})) + \langle \nabla f(\Psi(\mathbf{z})), \Psi(\mathbf{u}) - \Psi(\mathbf{z}) \rangle - \frac{\ell_f}{2}\|\Psi(\mathbf{u}) - \Psi(\mathbf{z})\|^2.$$

Denote

$$\Psi(\mathbf{u}) - \Psi(\mathbf{z}) = \mathrm{J}_\Psi(\mathbf{z})^\top(\mathbf{u} - \mathbf{z}) + \mathbf{r}(\mathbf{z}, \mathbf{u})$$

where $\|\mathbf{r}(\mathbf{z}, \mathbf{u})\| \leq \frac{L_\Psi}{2}\|\mathbf{u} - \mathbf{z}\|^2$.

Using the inequality (Lemma E.8), we have:

$$\langle \nabla f(\Psi(\mathbf{z})), \Psi(\mathbf{u}) - \Psi(\mathbf{z}) \rangle = \langle \mathrm{J}_\Psi(\mathbf{z}) \nabla f(\Psi(\mathbf{z})), \mathbf{u} - \mathbf{z} \rangle + \langle \nabla f(\Psi(\mathbf{z})), \mathbf{r}(\mathbf{z}, \mathbf{u}) \rangle$$

Since $\nabla h(\mathbf{z}) = \mathrm{J}_\Psi(\mathbf{z}) \nabla f(\Psi(\mathbf{z}))$:

$$\langle \nabla f(\Psi(\mathbf{z})), \Psi(\mathbf{u}) - \Psi(\mathbf{z}) \rangle = \langle \nabla h(\mathbf{z}), \mathbf{u} - \mathbf{z} \rangle + \langle \nabla f(\Psi(\mathbf{z})), \mathbf{r}(\mathbf{z}, \mathbf{u}) \rangle$$

Bounding the error term:

$$\langle \nabla f(\Psi(\mathbf{z})), \mathbf{r}(\mathbf{z}, \mathbf{u}) \rangle \geq -\|\nabla f(\Psi(\mathbf{z}))\| \cdot \frac{L_\Psi}{2}\|\mathbf{u} - \mathbf{z}\|^2 \geq -\frac{ML_\Psi}{2}\|\mathbf{u} - \mathbf{z}\|^2$$

where M bounds $\|\nabla f\|$ on the compact set. Also $\|\Psi(\mathbf{u}) - \Psi(\mathbf{z})\|^2 \leq \kappa_2^2 \|\mathbf{u} - \mathbf{z}\|^2$.

Therefore, we have

$$h(\mathbf{u}) \geq h(\mathbf{z}) + \langle \nabla h(\mathbf{z}), \mathbf{u} - \mathbf{z} \rangle - \frac{\ell_f \kappa_2^2 + ML_\Psi}{2}\|\mathbf{u} - \mathbf{z}\|^2.$$

So h is ℓ_h-weakly convex with

$$\ell_h = \ell_f \kappa_2^2 + ML_\Psi.$$

(2) By the chain rule, with $\mathbf{x} = \Psi(\mathbf{z})$, we have

$$\mathrm{J}_{\boldsymbol{\varphi}}(\mathbf{x}) = \mathrm{J}_\Psi^{-1}(\mathbf{z})\,\mathrm{J}_{\boldsymbol{\phi}}(\mathbf{z}), \quad \mathrm{N}_{\mathcal{C}}(\mathbf{x}) = \mathrm{J}_\Psi^{-1}(\mathbf{z})\,\mathrm{N}_{\mathcal{B}}(\mathbf{z}).$$

Thus, we obtain

$$\begin{aligned}
\nu\,\|\boldsymbol{\varphi}(\mathbf{x})\| &= \nu\,\|\boldsymbol{\phi}(\mathbf{z})\| \\
&\leq \operatorname{dist}\left(\mathrm{J}_{\boldsymbol{\varphi}}(\mathbf{x})\,\boldsymbol{\varphi}(\mathbf{x}), -\mathrm{N}_{\mathcal{C}}(\mathbf{x})\right) \\
&= \operatorname{dist}\left(\mathrm{J}_\Psi^{-1}(\mathbf{z})\,\mathrm{J}_{\boldsymbol{\phi}}(\mathbf{z})\,\boldsymbol{\phi}(\mathbf{z}), -\mathrm{J}_\Psi^{-1}(\mathbf{z})\,\mathrm{N}_{\mathcal{B}}(\mathbf{z})\right) \\
&\leq \frac{1}{\kappa_1}\operatorname{dist}\left(\mathrm{J}_{\boldsymbol{\phi}}(\mathbf{z})\,\boldsymbol{\phi}(\mathbf{z}), -\mathrm{N}_{\mathcal{B}}(\mathbf{z})\right).
\end{aligned}$$

Therefore, taking $\mu = \nu \cdot \kappa_1$ yields the desired conclusion. □

Lemma E.8. *Suppose that* J_Ψ *is* L_Ψ*-Lipschitz, i.e.,* $\|\mathrm{J}_\Psi(\mathbf{u}) - \mathrm{J}_\Psi(\mathbf{z})\| \leq L_\Psi\|\mathbf{u} - \mathbf{z}\|$ *for any* $\mathbf{u}$ *and* $\mathbf{z}$*. Then*

$$\left\|\Psi(\mathbf{u}) - \Psi(\mathbf{z}) - \mathrm{J}_\Psi(\mathbf{z})^\top(\mathbf{u} - \mathbf{z})\right\| \leq \frac{L_\Psi\|\mathbf{u} - \mathbf{z}\|^2}{2}, \quad \forall \mathbf{u}, \mathbf{z}.$$

The proof follows from Lemma 1.2.3 in (Nesterov et al., 2018).

Assumption 5. We assume: (1) f, φ_i for $i \in [p]$ are weakly convex; and (2) A variant of the KL condition holds: there exists a constant $\nu > 0$ such that for any $\mathbf{x} \in \mathcal{C}$, the following inequality holds

$$\nu \|\boldsymbol{\varphi}(\mathbf{x})\| \leq \operatorname{dist}\left(\mathrm{J}_{\boldsymbol{\varphi}}(\mathbf{x})\boldsymbol{\varphi}(\mathbf{x}), -\mathrm{N}_{\mathcal{C}}(\mathbf{x})\right). \tag{65}$$

where $\mathrm{N}_{\mathcal{C}}(\cdot)$ is the normal cone of constrained set $\mathcal{C}$ defined in Appendix C.1.

Given the above assumption, we are now ready to apply Hom-PALM(1) to problem (**H**) and provide the convergence analysis. As a direct corollary of Lemma E.7 and Theorem 8, Algorithm 1 with Hom-PALM(1) converges to a stationary point. The formal statement is given below.

Theorem 9. *Consider Hom-PALM($\theta = 1$) in Algorithm 1 with a quadratic proximal operator. Suppose that Assumption 5 holds for problem* (**P**)*. Then, with appropriate parameters β_k and ρ_k, Hom-PALM is intended to return an ϵ-solution $\mathbf{z}_\epsilon$ of problem* (**H**) *in $\tilde{\mathcal{O}}\left(\epsilon^{-3}\right)$, where subproblem* (1) *is solved by AdapAPG (Lin & Xiao, 2014) (outlined in Alg. 11).*

Remark. (i) The hidden constants in the complexity bound are influenced by the smoothness of $\hat{h}_k$ and the regularity condition; see item (2) in Lemma E.7. Consequently, they depend on the bi-Lipschitz constants κ_1 and κ_2 of the mapping Ψ.

F Experiments

In this section, we provide the omitted details of our experiments.

F.1 Reporting Metric

For every method, we report the objective value, wall-clock time, outer iteration count, equality violation, inequality violation, and the combined feasibility residual. Let $g_i(\mathbf{x}) \leq 0$ denote the inequality constraints and let $h_j(\mathbf{x}) = 0$ denote the equality constraints. The reported feasibility quantities are

$$\mathrm{vio}_{\mathrm{ineq}}(\mathbf{x}) := \max_i [g_i(\mathbf{x})]_+, \qquad \mathrm{vio}_{\mathrm{eq}}(\mathbf{x}) := \max_j |\varphi_j(\mathbf{x})|, \tag{66}$$

where $[a]_+ := \max\{a, 0\}$. The full violation is the maximum residual certificate

$$\mathrm{vio}_{\mathrm{full}}(\mathbf{x}) := \max\{\mathrm{vio}_{\mathrm{ineq}}(\mathbf{x}), \mathrm{vio}_{\mathrm{eq}}(\mathbf{x})\}. \tag{67}$$

Thus feasibility is measured by the worst constraint residual, rather than by an averaged residual or an ℓ_2 norm. This convention is used for both the convex QCQP and Stiefel experiments.

For each benchmark instance, we plot objective value and feasibility residuals against both outer iteration and running time. The time-axis plots use the measured cumulative wall-clock time of the corresponding implementation. The reference objective obtained by the problem-specific solver is shown as a black dotted horizontal line and is annotated directly by the solver name, namely MOSEK for convex QCQP instances and IPOPT for Stiefel instances.

F.2 (Baseline) Algorithms

In this section, we list all algorithms that we mentioned in this work, including the baseline algorithms in the empirical study.

- **PGD**: projected gradient descent in the original variable. For the convex QCQP experiments, the projection subproblem is solved by the convex solver used by the implementation.
- **ALM**: ALM transforms the constrained optimization problem into a sequence of unconstrained subproblems by incorporating constraints directly into the objective function, and updates the dual variable based on constraint violations at the current solution. See e.g., (Bertsekas, 1997; Deng et al., 2025).

- **PALM**: proximal ALM shown in Alg. 6, that applies the augmented Lagrangian only to equality constraints, while convex inequalities are handled explicitly through the constrained subproblem. It separates equality handling from inequality projection and is used in the convex QCQP comparison.
- **Hom-PALM**: a combination of homeomorphic reformulation and PALM, shown in Alg. 1.
- **PPP**: proximal-point penalty method, shown in Alg. 8.
- **AdapAPG**: adaptive APG (Lin & Xiao, 2014; Lin et al., 2022), used as the inner solver in the PPP and Hom-PALM(1) convergence proofs; see Algorithm 11 in Appendix E.4.
- **ACG**: accelerated composite gradient (Florea & Vorobyov, 2018), used as the inner solver in the nonconvex PALM and Hom-PALM(θ) convergence proofs; see Algorithm 7 in Appendix E.3.
- **Hom-APG**: a combination of homeomorphic transformation and APG method, as outlined in Alg. 5, which served as the inner solver of Hom-PALM in the convex setting.
- (i) **Retraction-Penalty**: a Stiefel-specific baseline that keeps the orthogonality equality constraint by tangent-space gradient projection followed by QR retraction, and handles convex side constraints by a quadratic penalty. (ii) **Retraction-ALM**: a Stiefel-specific baseline with the same tangent-space and QR-retraction update, but with augmented-Lagrangian updates for the convex side inequalities. See e.g., (Absil & Malick, 2012; Liu & Boumal, 2020; Andreani et al., 2024).
- **MOSEK**: A commercial interior-point optimizer that solves conic optimization problems efficiently using highly optimized primal-dual interior-point methods with predictor-corrector techniques and sparse linear algebra. Note that we use an Academic license for MOSEK.
- **IPOPT**: an interior-point nonlinear programming solver used as the reference solver for the nonconvex Stiefel constrained PCA experiments.

F.3 Convex QCQP with Linear Equalities

The convex benchmark is a quadratic program with convex quadratic inequalities, box constraints, and linear equalities:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{p}^\top \mathbf{x} \\ \text{s.t.} \quad & g_i(\mathbf{x}) \le 0, \qquad i = 1, \dots, m_{\text{quad}}, \\ & \mathbf{A}_{\text{eq}}\mathbf{x} - \mathbf{b}_{\text{eq}} = 0, \\ & L \le \mathbf{x} \le U. \end{aligned} \tag{68}$$

The active quadratic inequality residuals are

$$g_i(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Q}_i\mathbf{x} + \mathbf{p}_i^\top \mathbf{x} - \mathbf{b}_i, \qquad \mathbf{Q}_i \succeq 0. \tag{69}$$

The box constraints are included in the inequality residual vector as $L - \mathbf{x} \le 0$ and $\mathbf{x} - U \le 0$. Hence the total inequality count is $m = m_{\text{quad}} + 2n$. In the reported instances, the objective matrix and all quadratic constraint matrices are diagonal positive semidefinite matrices, so the benchmark remains convex while allowing a large number of inequality residuals.

F.4 Constrained PCA on the Stiefel Manifold

The Stiefel benchmark is a constrained PCA problem. The decision variable is a matrix $\mathbf{X} \in \mathbb{R}^{n\times k}$, flattened internally as a vector in $\mathbb{R}^{nk}$. Given a data covariance matrix $\mathbf{A}$, the objective is

$$\min_{\mathbf{X}} \quad -\operatorname{tr}(\mathbf{X}^\top \mathbf{A}\mathbf{X}). \tag{70}$$

For the data-driven instances, the covariance matrix is standardized and normalized by the sum of its largest k eigenvalues. Under the unconstrained PCA solution, this normalization

makes the best unconstrained objective close to -1; active side constraints generally move the constrained reference objective above this value.

The nonlinear equality constraint is the Stiefel constraint

$$\mathbf{X}^\top \mathbf{B}\mathbf{X} = \mathbf{I}_k. \tag{71}$$

In the reported data-driven settings, $\mathbf{B} = \mathbf{I}_n$, so the equality is $\mathbf{X}^\top \mathbf{X} = \mathbf{I}_k$. We count the equality residual as $p = k^2$, matching the full matrix residual used by the implementation. The convex side constraints consist of elementwise box bounds and one group budget over selected feature rows:

$$L \le X_{ij} \le U, \qquad \|\mathbf{X}_{\mathcal{R},:}\|_F \le \rho_{\mathrm{grp}}. \tag{72}$$

The row set $\mathcal{R}$ selects coordinates with large covariance loadings.

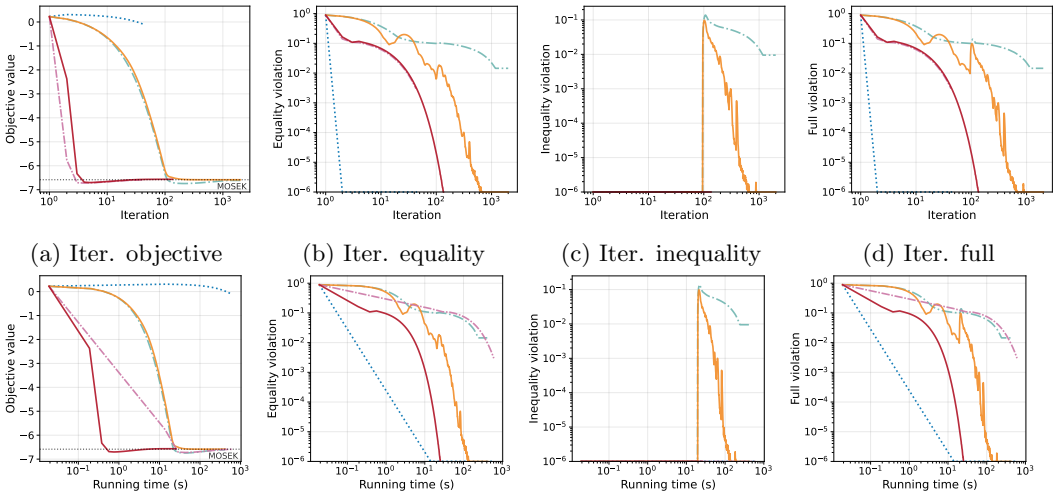
F.5 Implementation Details

Gradient calculation. For simple quadratic objective functions, gradients are calculated using closed-form formulas. Other nontrivial gradients are implemented using PyTorch automatic differentiation. This keeps the implementation uniform across the QCQP and Stiefel benchmarks; replacing all automatic differentiation calls by specialized closed-form gradients could further improve wall-clock time, but would not change the reported residual definitions or objective values.

Step-size policy. The algorithms have different theoretical step-size prescriptions, including rules based on smoothness constants, convexity parameters, or reference objective values. In the numerical comparison, first-order updates use the same practical adaptive protocol whenever possible: an initial learning rate is specified for the experiment family, and the learning rate is decayed when the objective/violation does not decrease. This keeps the comparison focused on the constraint-handling mechanism rather than on method-specific line-search engineering.

F.6 Complete QCQP Convergence Plots

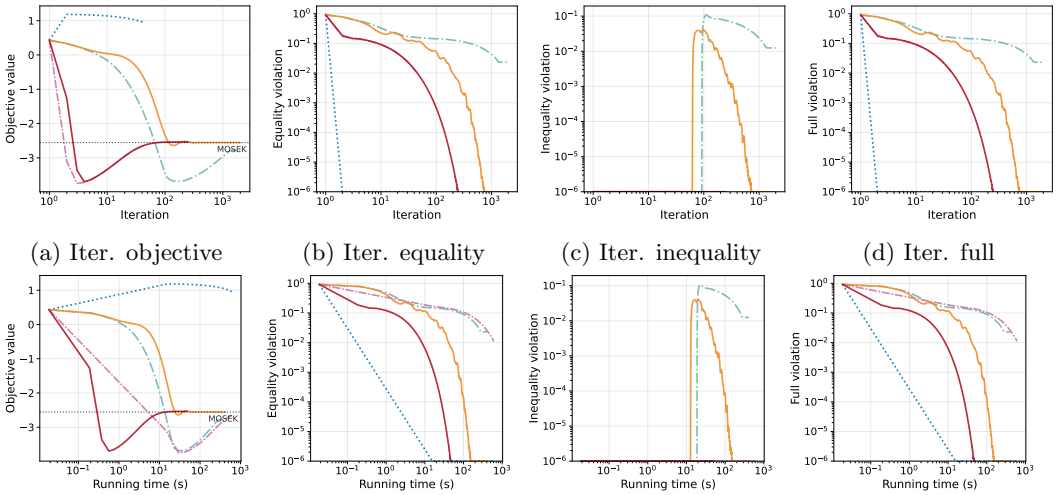
We provide the detailed iteration- and time-based convergence plots. For each instance, we show objective value, equality violation, inequality violation, and full violation using the same maximum residual metric from Appendix F.1.



(a) Iter. objective (b) Iter. equality (c) Iter. inequality (d) Iter. full

(e) Time objective (f) Time equality (g) Time inequality (h) Time full

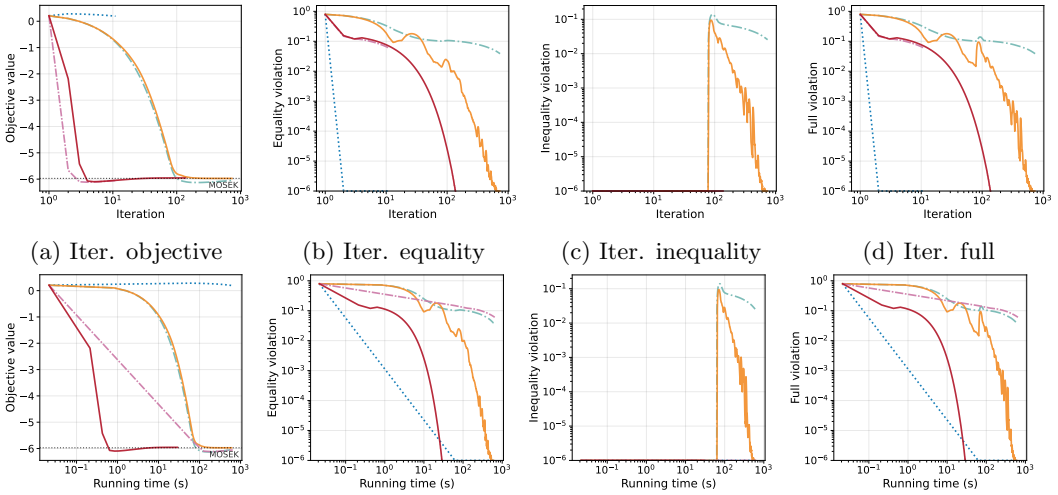
Figure 5: Convex QCQP with $(n, p, m) = (1000, 100, 2100)$.



(a) Iter. objective (b) Iter. equality (c) Iter. inequality (d) Iter. full

(e) Time objective (f) Time equality (g) Time inequality (h) Time full

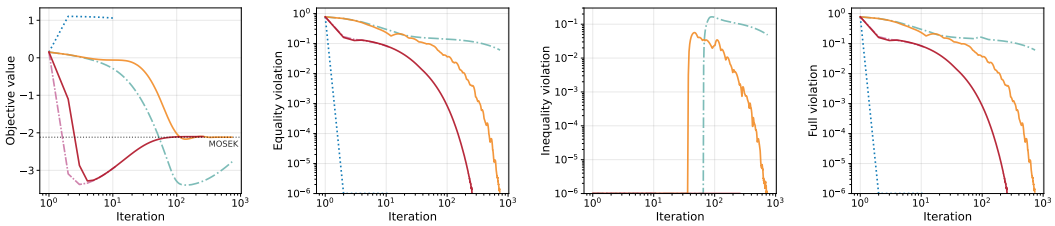
Figure 6: Convex QCQP with $(n, p, m) = (1000, 500, 2100)$.



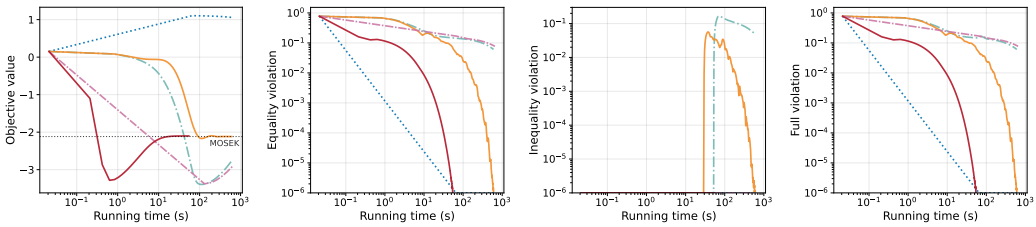
(a) Iter. objective (b) Iter. equality (c) Iter. inequality (d) Iter. full

(e) Time objective (f) Time equality (g) Time inequality (h) Time full

Figure 7: Convex QCQP with $(n, p, m) = (1000, 100, 2500)$.



(a) Iter. objective (b) Iter. equality (c) Iter. inequality (d) Iter. full

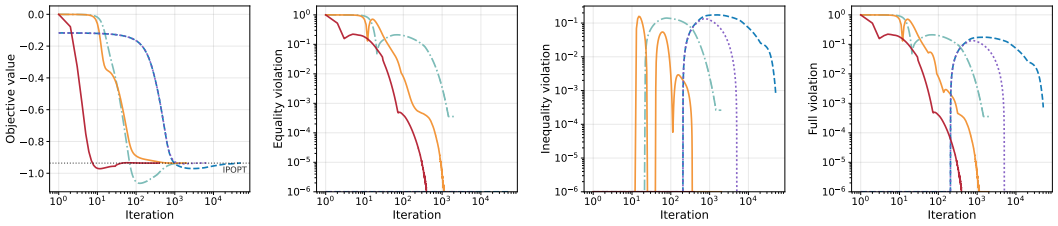


(e) Time objective (f) Time equality (g) Time inequality (h) Time full

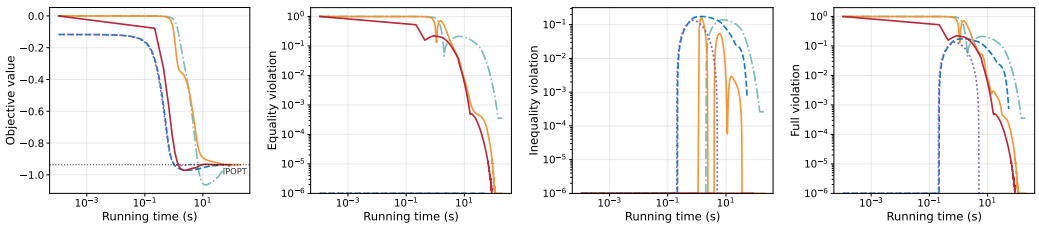
Figure 8: Convex QCQP with $(n, p, m) = (1000, 500, 2500)$.

F.7 Complete Stiefel Convergence Plots

We provide the detailed iteration- and time-based convergence plots. For each instance, we show objective value, equality violation, inequality violation, and full violation using the same maximum residual metric from Appendix F.1.

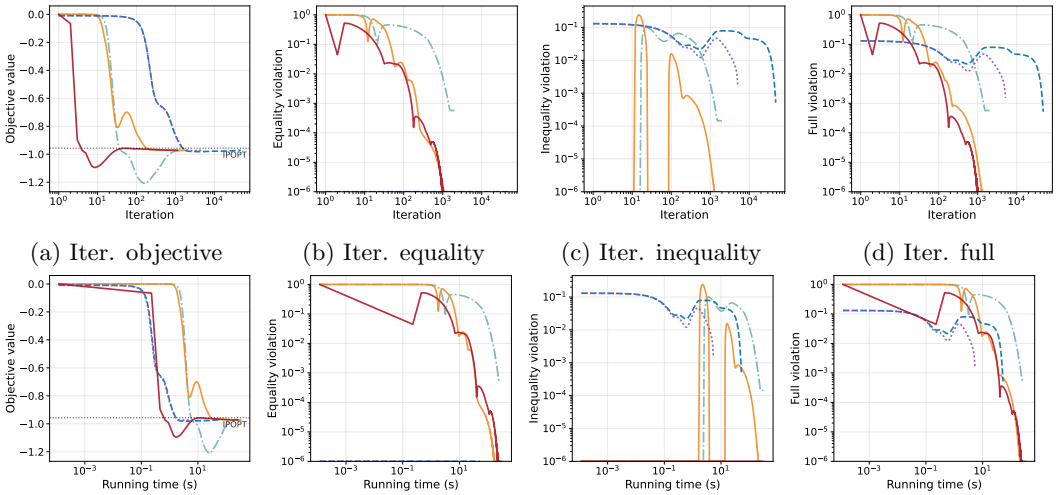


(a) Iter. objective (b) Iter. equality (c) Iter. inequality (d) Iter. full



(e) Time objective (f) Time equality (g) Time inequality (h) Time full

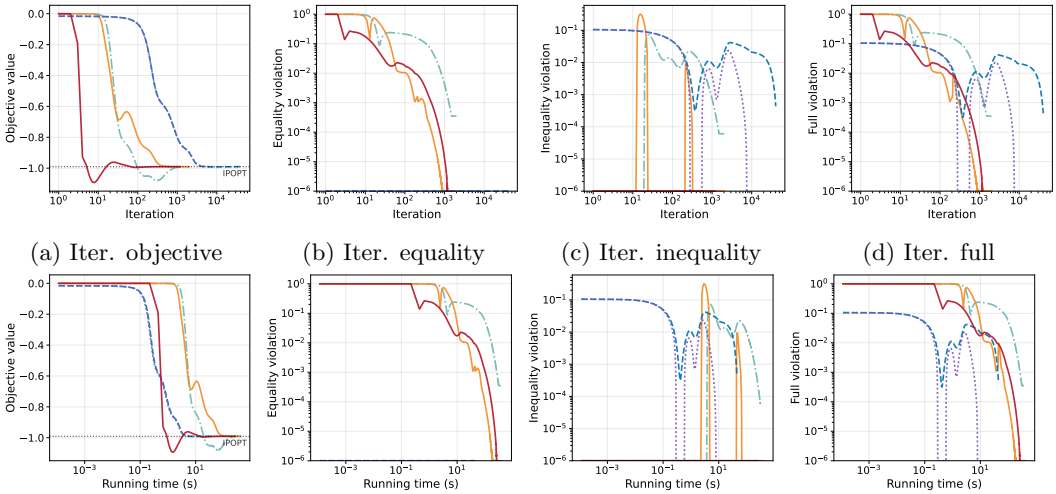
Figure 9: Stiefel constrained PCA on digits with $(n, p, m) = (192, 9, 385)$.



(a) Iter. objective (b) Iter. equality (c) Iter. inequality (d) Iter. full

(e) Time objective (f) Time equality (g) Time inequality (h) Time full

Figure 10: Stiefel constrained PCA on LFW with $(n, p, m) = (2139, 9, 4279)$.



(a) Iter. objective (b) Iter. equality (c) Iter. inequality (d) Iter. full

(e) Time objective (f) Time equality (g) Time inequality (h) Time full

Figure 11: Stiefel constrained PCA on LFW with $(n, p, m) = (3565, 25, 7131)$.