OPTIMALITY-AWARE PROJECTION: ALIGNING PROJECTION GEOMETRY WITH CONSTRAINED OPTIMALITY

Anonymous authors
Paper under double-blind review

ABSTRACT

In learning-based constrained optimization pipelines, feasibility-enforcement layers are widely adopted to guarantee strict constraint satisfaction. However, optimizing an unconstrained loss through such feasibility mappings creates a non-equivalent surrogate formulation for the target constrained optimization problem. Specifically, methods based on orthogonal projection exhibit boundary-collapse behavior, which suppresses gradient components normal to active constraints. Consequently, stationary points of this surrogate formulation misalign with the optimality conditions of the target problem, driving convergence to suboptimal solutions. To address this issue, we propose Optimality-Aware (OA) projection , a novel learning mechanism that endows projection-integrated architectures with genuine optimality-seeking capabilities. Without modifying the orthogonal projection layer itself, OA projection injects a KKT-residual-guided geometric correction into the training gradient flow, making it readily embedded into learning-based optimization pipelines. By recovering informative gradients that propel optimization progress along boundaries, this correction favorably reshapes the parameter trajectory and systematically avoids orthogonal-projection-induced spurious stationary points. Theoretically, we establish the differentiability of OA projection, prove its convergence under stochastic gradient updates, and show that under suitable regularity conditions, the converged iterates satisfy the KKT conditions of the target problem. Experimental results on diverse benchmarks, including real-world tasks, confirm that OA projection outperforms state-of-the-art methods by substantially improving the solution quality of hard-constrained learning models while strictly preserving feasibility.

1 INTRODUCTION

Practical decision-making operates under hard operational, physical, or geometric constraints that delimit the admissible decision space. While learning-based models can efficiently approximate complex context-to-decision mappings, they do not inherently preserve hard constraints, yielding high-performing yet infeasible decisions (Chen et al., 2024; Khaloie et al., 2025). This has motivated a growing body of work on feasibility-enforcement architectures, which integrate constraint-handling mechanisms into learning pipelines (Grontas et al., 2026; Nguyen & Donti, 2025; Donti et al., 2021b). Among them, projection-based approaches are broadly employed, as they strictly preserve feasibility during both training and deployment while offering geometric properties amenable to analysis (Donti et al., 2021a; Liang et al., 2023).

However, projection layers fundamentally alter the optimization problem in a standard end-to-end training pipeline, as composing an unconstrained predictor with a projection operator induces a surrogate objective whose stationarity depends on the local projection geometry. For canonical orthogonal projection (OP), an entire normal region outside the feasible set can collapse onto a lower-dimensional boundary face, preserving tangential variations while eliminating normal cones (Schneider & Kuhn, 2026). The backpropagation consequently inherits this geometric degeneracy, so stationarity of the surrogate does not necessarily imply optimality for the original constrained problem. This raises a fundamental dilemma: how to retain exact feasibility of OP without inducing spurious stationary solutions.

We resolve this dilemma by developing Optimality-Aware (OA) projection, a novel mechanism that utilizes optimality information to refine the geometric degeneracy arising from OP in end-to-end

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.

training. Specifically, OA projection injects a radial geometric offset into the backward gradient flow which is governed by the KKT residual evaluated at the preceding iteration. Detaching this historical residual from the active differentiation path converts it into a closed-loop gradient feedback signal rather than an auxiliary loss term, eliminating circular computational dependencies. By restoring the normal gradient component suppressed by canonical OP, OA projection steers the optimization flow away from spurious boundary traps. As KKT stationarity is approached, the radial offset naturally vanishes and OA projection recovers standard OP. Our theoretical analysis establishes the differentiability and feasibility preservation of OA projection, proves its convergence under stochastic gradient descent (SGD), and shows that non-KKT iterates are unstable under the resulting dynamics. Specifically, any persistent violation of KKT stationarity induces a non-summable normal drift that prevents convergence to spurious stationary points. Under a nondegeneracy condition, we further establish a corresponding KKT convergence guarantee for linear probing. Experiments on synthetic benchmarks and real-world tasks demonstrate that OA projection substantially improves solution quality while maintaining strict feasibility. Our main contributions are summarized as follows:

- We propose OA projection, a novel KKT-residual-guided mechanism that corrects the training trajectory while retaining OP for strict feasibility enforcement, thereby endowing projection-based learning with optimality-seeking capability.
- We characterize how boundary collapse under OP creates non-optimal surrogate stationary points and prove how OA projection inherently restores the suppressed optimization directions. We additionally analyze the boundary saturation associated with static radial-oriented offsets and finally establish that OA projection avoids this limitation through its adaptive, diminishing correction.
- We theoretically establish feasibility, differentiability, and convergence of OA projection, and prove that persistent non-KKT iterates cannot be convergence limits under suitable regularity conditions. We further extend the KKT convergence guarantee to linear probing settings under a nondegeneracy condition.
- Experiments on diverse benchmarks demonstrate that OA projection preserves strict feasibility while substantially improving decision optimality over state-of-the-art constraint-enforcing methods.

2 Related work

Our study intersects several lines of research concerning how learned models participate in optimization, how feasibility is imposed on their outputs, and how constraint-handling operations are differentiated during training. We review these threads below.

Machine learning for optimization. Machine learning has been used to solve families of optimization problems through two broad paradigms. End-to-end approaches learn direct mappings from context to decisions, thereby amortizing solver costs across similar instances (Khaloie et al., 2025; Wu et al., 2024; Hasan et al., 2020; Amos, 2023). They rely on solver supervision or task-driven objectives, with the latter aligning training directly with decision quality (Zhang et al., 2023; Liu et al., 2021). A complementary paradigm accelerates explicit solvers through learned warm starts, active-set prediction, constraint screening, or approximate updates (Sambharya et al., 2024; Banerjee et al., 2025; Natarajan, 2021). Typically, these learning-assisted methods retain the explicit solver as the principal solution mechanism, whereas end-to-end approaches shift most online computation to offline training.

Soft constraint enforcement. To incorporate constraints into model training, soft penalty methods penalizing task objectives with measures of constraint violation (Tanneau & Hentenryck, 2024; Elenter et al., 2024; Pan et al., 2023; Narasimhan et al., 2020). Such formulations encourage rather than enforce feasibility. Finite penalty weights permit residual violations, whereas scaling weights toward infinity induces ill-conditioning and hyperparameter sensitivity. Consequently, low empirical violations do not guarantee feasibility at inference. Lagrangian and primal–dual approaches replace fixed weights with dynamically updated multipliers via augmented-Lagrangian or descent–ascent schemes (Monaco et al., 2026; Nandwani et al., 2019; Park & Van Hentenryck, 2023; Dener et al., 2020), but their theoretical guarantees typically concern expected violations or saddle-point convergence rather than samplewise feasibility.

Hard constraint enforcement. To provide strict feasibility guarantees, hard-constrained methods map raw predictions directly onto the feasible region through explicit forward operators. Iterative methods seek feasibility either by combining equality completion with manifold-based inequality correction or by iteratively minimizing joint constraint violations (Nguyen & Donti, 2025; Ghosh et al., 2024). Although broadly applicable, these methods require repeated correction at inference, making their computational cost sensitive to inner-loop convergence and termination criteria (Zeng et al., 2026; Lee et al., 2019). Projection-based methods instead construct surjective maps from the unconstrained space onto the feasible region through boundary truncation, radial maps, or diffeomorphisms (Tordesillas et al., 2026; Schneider & Kuhn, 2026; Liang et al., 2024; Frerix et al., 2020). These maps admit closed-form evaluation, avoiding iterative feasibility correction at inference. Among these methods, OP serves as a canonical choice due to its favorable mathematical properties and broad applicability, inspiring numerous accelerated and differentiable variants (Agrawal et al., 2019b; Grontas et al., 2026; Zhang et al., 2016; Wang et al., 2024; Min & Azizan, 2025). However, existing projection-based methods primarily concern feasibility, expressivity, or differentiability, while their implications for the correspondence between model-level stationarity and decision-level optimality remain less fully understood (Wang et al., 2023; Zeng et al., 2024).

Differentiable optimization and implicit layers. Constraint-handling modules are frequently formulated as implicit layers defined by optimization problems or fixed-point systems, making differentiating their solution maps essential for end-to-end training (Amos & Kolter, 2017; Pan et al., 2024; Gould et al., 2022). Within this paradigm, problem classes such as quadratic programs, conic programs, nonlinear least-squares, and submodular minimization admit differentiable solution maps under suitable conditions (Agrawal et al., 2019b;a; Blondel et al., 2022; Pineda et al., 2022; Djolonga & Krause, 2017). Without explicit expressions, implicit differentiation derives gradients by applying the implicit function theorem directly to KKT systems or fixed-point conditions (Bolte et al., 2021). In our framework, this technique serves as the foundational tool to backpropagate through the projection operator.

3 Optimality-Aware Projection

3.1 Problem Statement: spurious stationarity induced by projection

In this paper, we consider stochastic constrained problems in which a context $z \sim \mathcal{D}$ is observed before choosing a decision. The goal is to learn a measurable decision rule π that solves

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_{z\sim\mathcal{D}}[\ell(\pi(z), z)] \\ \text{s.t.} \quad & \pi(z) \in \mathcal{C}, \qquad \mathcal{D}\text{-a.s.}, \end{aligned} \tag{1}$$

where $\ell(y, z)$ is the decision cost and $\mathcal{C} = \{y \in \mathbb{R}^d : Ay \leq a,\ By = b\}$ is a nonempty polyhedron. To make this functional problem tractable, learning-based methods address this issue by restricting the decision rule to a parametric model $g_\theta : \mathcal{Z} \to \mathbb{R}^d$. Since the output $u = g_\theta(z)$ does not inherently satisfy the hard constraint $u \in \mathcal{C}$, existing research typically composes g_θ with a constraint-handling operator to enforce feasibility. Here, we focus on the canonical and broadly employed OP operator $\Pi_{\mathcal{C}}$. Composing g_θ with Π_C yields the surrogate learning problem below.

$$\min_{\theta}\ F_{\mathrm{H}}(\theta) := \mathbb{E}_{z\sim\mathcal{D}}\left[\ell\big(\Pi_{\mathcal{C}}(g_\theta(z)), z\big)\right]. \tag{2}$$

While this formulation guarantees feasibility, it does not ensure decision optimality due to the inequivalence between the stationary condition of the surrogate problem and the optimality condition of the original problem. To expose the geometric origin of this mismatch, we next set aside model parameterization and analyze the discrepancy between KKT conditions of the original problem and surrogate stationary conditions directly at the raw-output level.

Fixing a context z , the instance-wise original problem and its projection-composed surrogate can be written as follows,

$$(\mathsf{P}_z) \quad \min_{y\in\mathcal{C}} \ell(y, z) \qquad \text{vs.} \qquad (\mathsf{H}_z) \quad \min_{u\in\mathbb{R}^d} \varphi_z(u) := \ell\big(\Pi_{\mathcal{C}}(u), z\big).$$

For the original problem P_z, the stationary point $y \in \mathcal{C}$ necessarily satisfies the KKT condition below,

$$0 \in \nabla_y \ell(y, z) + N_{\mathcal{C}}(y). \tag{3}$$

where $N_{\mathcal{C}}(y) := \{v \in \mathbb{R}^d : \langle v, \widetilde{y} - y\rangle \le 0, \forall\, \widetilde{y} \in \mathcal{C}\}$ is the normal cone of $\mathcal{C}$ at y. Because $\Pi_{\mathcal{C}}$ is generally nonsmooth, we express surrogate stationarity as $0 \in \mathcal{D}\varphi_z(\bar{u})$ using a conservative gradient. To relate this condition to the local projection geometry, set $\bar{y} := \Pi_{\mathcal{C}}(\bar{u})$ and let F be the unique face satisfying $\bar{y} \in \mathrm{relint}(F)$. Define $T_F := \mathrm{span}(F - F)$ and denote its orthogonal projector by P_{T_F}. For polyhedral projection, P_{T_F} belongs to the conservative Jacobian $\mathcal{D}\varphi_z(\bar{u})$. Applying the conservative chain rule gives the sufficient condition for surrogate stationarity below.

$$P_{T_F}\nabla_y \ell(\bar{y}, z) = 0 \quad \Longrightarrow \quad 0 \in \mathcal{D}\varphi_z(\bar{u}) = \left\{J^\top \nabla_y \ell(\bar{y}, z) : J \in \mathcal{J}_{\Pi_{\mathcal{C}}}(\bar{u})\right\}. \tag{4}$$

Because $P_{T_F}\nabla_y \ell(\bar{y}, z) = 0$ is equivalent to $\nabla_y \ell(\bar{y}, z) \in T_F^{\perp}$, the proper inclusion $N_{\mathcal{C}}(\bar{y}) \subsetneq T_F^{\perp}$, establishes the following one-way implication chain,

$$-\nabla_y \ell(\bar{y}, z) \in N_{\mathcal{C}}(\bar{y}) \quad \Longrightarrow \quad P_{T_F}\nabla_y \ell(\bar{y}, z) = 0 \quad \Longrightarrow \quad 0 \in \mathcal{D}\varphi_z(\bar{u}) \tag{5}$$

The reverse implication fails because P_{T_F} annihilates the entire normal subspace $T_F^{\perp}$ without distinguishing whether $-\nabla_y \ell(\bar{y}, z)$ lies in the normal cone. This normal-gradient collapse can therefore make the surrogate stationary even when the projected decision fails the KKT condition. We formalize such points in the definition below.

Definition 1 (Projection-induced spurious stationary point)**.** *Let $\bar{y} = \Pi_C(\bar{u})$. We call the point $\bar{u}$ a projection-induced spurious stationary point if there exists a face F such that $\bar{y} \in \mathrm{relint}(F)$, $\bar{u} - \bar{y} \in \mathrm{ri}\, N_C(F)$, and*

$$P_{T_F}\nabla_y \ell(\bar{y}, z) = 0, \qquad -\nabla_y \ell(\bar{y}, z) \notin N_C(\bar{y}).$$

Because this pathology appears in raw output space, highly expressive models inherit spurious stationary points when composed with projection layers. It originates from a mismatch between projection derivatives and the normal-cone condition for constrained optimality. This motivates the development of feasibility-enforcing formulations whose stationary conditions remain aligned with the KKT conditions of the original constrained problem.

3.2 Optimality-Aware Projection

To align surrogate stationarity with constrained optimality, we introduce Optimality-Aware (OA) Projection, which restores the normal-gradient information suppressed by OP through a radial correction during training. A lagged KKT diagnostic controls the correction, while a vanishing envelope recovers standard OP as training proceeds. The radial construction is defined without loss of generality in the intrinsic ambient space $\mathcal{A} := \mathrm{aff}(\mathcal{C})$, the affine hull of $\mathcal{C}$, where the equality constraints are automatically satisfied. With $\mathrm{ri}(\mathcal{C})$ denoting the interior relative to $\mathcal{A}$, we fix an anchor $y_0 \in \mathrm{ri}(\mathcal{C})$ and consider a raw input $u \in \mathcal{A}$.

Definition 2 (Soft-Radial (SR) map (Schneider & Kuhn, 2026))**.** *Define the radial endpoint as below.*

$$q_{\mathrm{R}}(u) := y_0 + \lambda^\star(u)(u - y_0), \qquad \textit{where } \lambda^\star(u) := \sup\{\lambda \in [0,1] : y_0 + \lambda(u - y_0) \in \mathcal{C}\}. \tag{6}$$

Let $\rho : [0,\infty) \to [0,1)$ satisfy $\rho(0) > 0$ and $\lim_{t\to\infty} \rho(t) = 1$, and assume that ρ is continuously differentiable and strictly increasing with $\rho' > 0$. The Soft-Radial (SR) map is

$$p_{\mathrm{SR}}(u) := y_0 + \rho\big(\|u - y_0\|_2^2\big)\,\big(q_{\mathrm{R}}(u) - y_0\big). \tag{7}$$

The point $q_{\mathrm{R}}(u)$ is the endpoint of the feasible segment extending from y_0 toward u and it equals u when $u \in \mathcal{C}$. Since the radial factor is strictly below one for every finite u, $p_{\mathrm{SR}}(u)$ always lies in $\mathrm{ri}(\mathcal{C})$ and cannot attain the boundary at any finite raw point. Unlike OP, which suppresses variations orthogonal to an active boundary face, the SR map is an interior parameterization with an almost-everywhere invertible Jacobian, thereby avoiding this normal collapse.

Nevertheless, this advantage comes with a complementary boundary limitation. A boundary point has no finite pre-image under the SR map, so approaching it requires the raw input to diverge in norm. Moreover, as the radial factor saturates toward one, increasingly large variations in raw decisions can produce vanishing changes near the boundary, inducing gradient saturation. Therefore, OA projection employs the SR map solely as a transient training correction to preserve a well-defined raw representation throughout the feasible set. Specifically, we define the correction field below.

$$\Delta_{\mathrm{SR}}(u) := p_{\mathrm{SR}}(u) - \Pi_{\mathcal{C}}(u). \tag{8}$$

Rather than applying a fixed correction strength, OA projection regulates it dynamically across training iterations. At iteration k, the local correction is scaled by $\mathcal{R}_{k-1}$, which denotes the KKT residual of the orthogonally projected decision from iteration $k-1$. Meanwhile, a decaying correction budget progressively controls the overall magnitude. Combining these regulating factors yields the following definition of OA projection.

Definition 3 (OA projection)**.** *Let $\psi : [0,\infty) \to [0,1]$ be continuous and nondecreasing, with $\psi(0) = 0$ and $\psi(r) > 0$ for every $r > 0$. The correction schedule $\{\eta_k\}_{k\geq 0}$ satisfies $0 \leq \eta_k \leq 1$ and $\eta_k \to 0$. Given $u \in \mathcal{A}$ and the pair $(\eta_k, \mathcal{R}_{k-1})$ available at iteration k, the OA projection is*

$$p_{\mathrm{OA},k}(u) := \Pi_{\mathcal{C}}(u) + \alpha_k \Delta_{\mathrm{SR}}(u), \qquad \alpha_k = \eta_k \psi(\mathcal{R}_{k-1}). \tag{9}$$

Definition 3 guarantees strict feasibility throughout training because $p_{\mathrm{OA},k}(u) = (1-\alpha_k)\Pi_{\mathcal{C}}(u) + \alpha_k p_{\mathrm{SR}}(u)$ forms a convex combination of two feasible maps in the convex set $\mathcal{C}$. Evaluating the KKT statistic with a one-iteration lag eliminates circular dependencies between the projection operator and its correction coefficient, preserving an explicit forward pass. Consequently, the preceding state $\mathcal{R}_{k-1}$ dictates current projection geometry, while the resulting decisions update the optimality statistic $\mathcal{R}_k$ for the subsequent step.

In summary, OA projection defines a sequence of feasible and dynamically updated projection maps whose radial corrections are regulated by target optimality and systematically decay over training. This completes the forward formulation. The following subsection develops its differentiability, presenting the end-to-end training algorithm alongside the update rule for the lagged KKT statistic.

3.3 End-to-End Training with OA Projection

End-to-end training with OA projection requires an iteration-wise derivative and an update rule for its lagged optimality signal. We derive both and integrate them into mini-batch stochastic gradient descent (SGD). The following theorem establishes the regularity of the iteration-wise OA map and provides a computable conservative-Jacobian.

Theorem 4 (Regularity and differentiability of OA projection)**.** *Consider the polyhedral feasible set $\mathcal{C} = \{y \in \mathbb{R}^d \mid Ay \leq a, By = b\}$ and its affine hull $\mathcal{A} = \mathrm{aff}(\mathcal{C})$. The SR map is locally Lipschitz and differentiable almost everywhere on $\mathcal{A}$, with its Jacobian denoted by $J_{\mathrm{SR}}(u)$. The orthogonal projector $\Pi_{\mathcal{C}}(\cdot)$ admits a nonempty conservative Jacobian $\mathcal{J}_{\mathrm{H}}(u)$. Because α_k remains fixed during iteration k, $p_{\mathrm{OA},k}$ also admits a nonempty conservative Jacobian. Define the following notations,*

$$\mathcal{I}_{\mathrm{H}}(u) := \{i : A_i p_{\mathrm{H}}(u) = a_i\}, \qquad M_{\mathrm{H}}(u) := \begin{bmatrix} A_{\mathcal{I}_{\mathrm{H}}(u)} \\ B \end{bmatrix}.$$

where A_i and a_i denote the ith row of A and the ith entry of a, respectively, and $A_{\mathcal{I}_{\mathrm{H}}(u)}$ is the submatrix indexed by $\mathcal{I}_{\mathrm{H}}(u)$. Then an efficiently computable element of $\mathcal{J}_{\mathrm{H}}(u)$ is given below.

$$J_{\mathrm{H}}(u) := I_d - M_{\mathrm{H}}(u)^{\mathsf{T}} \big(M_{\mathrm{H}}(u) M_{\mathrm{H}}(u)^{\mathsf{T}}\big)^{\dagger} M_{\mathrm{H}}(u), \tag{10}$$

where $\dagger$ denotes the Moore–Penrose pseudoinverse. Consequently, the selected conservative Jacobian of OA projection can be computed as follows,

$$J_{\mathrm{OA},k}(u) = (1-\alpha_k) J_{\mathrm{H}}(u) + \alpha_k J_{\mathrm{SR}}(u). \tag{11}$$

The proof is provided in Appendix A.3. The selection $J_{\mathrm{H}}(u)$ facilitates efficient gradient computation and agrees with the classical projection Jacobian whenever the active projection face remains unchanged under small perturbations of u (Li et al., 2020; Yang et al., 2026). Geometrically, it projects onto $\ker M_{\mathrm{H}}(u)$ and, therefore, transmits precisely the variations tangent to the active face, while (11) augments this tangential derivative with the radial sensitivity of SR. Detaching $\mathcal{R}_{k-1}$ fixes α_k during iteration k, so backpropagation uses this fixed mixture without differentiating through the KKT-residual component or introducing circular dependence on the current decisions. We next specify the residual used to update $\mathcal{R}_k$.

With the projection derivative established, we now integrate the OA-projection operator into the learning pipeline of the parametric decision rule g_θ. At iteration k, given a mini-batch $\mathcal{B}_k$, parameter θ_k, and learning rate $\gamma_k > 0$, the forward pass maps each input $z \in \mathcal{B}_k$ to a raw candidate $u_k(z) := g_{\theta_k}(z)$ and its feasible counterpart $y_k(z) := p_{\mathrm{OA},k}(u_k(z))$. Denoting the active inequality set as

$\mathcal{I}(y) := \{i : A_i y = a_i\}$, the KKT residual of orthogonal component $y_{\mathrm{H},k}(z) := \Pi_{\mathcal{C}}(u_k(z))$ can be computed as follows,

$$r_k(y) := \min_{\lambda \in \mathbb{R}_+^{|\mathcal{I}(y)|}, \mu \in \mathbb{R}^{m_{\mathrm{eq}}}} \left\| -\nabla_y \ell(y, z) - A_{\mathcal{I}(y)}^\top \lambda - B^\top \mu \right\|_2, \tag{12}$$

where m_{eq} is the number of equality constraints. These residuals are aggregated into the batch optimality statistic $\mathcal{R}_k$, which is stored to regulate the projection at iteration $k+1$. Combining the residual update with the conservative Jacobian computation yields the following training procedure.

Algorithm 1 End-to-end training with OA Projection

Require: model g_θ, feasible set $\mathcal{C}$, anchor y_0, learning rates $\{\gamma_k\}_{k=0}^{K-1}$, response function ψ

Initialize θ_0
for $k = 0, \ldots, K-1$ **do**
  Sample a mini-batch $\mathcal{B}_k$, $\alpha_k \leftarrow \begin{cases} 1, & k = 0, \\ \eta_k \psi(\mathcal{R}_{k-1}), & k \geq 1 \end{cases}$
  for $z \in \mathcal{B}_k$ **do**
    $u_k(z) \leftarrow g_{\theta_k}(z)$, $y_{\mathrm{H},k}(z) \leftarrow p_{\mathrm{H}}(u_k(z))$, $y_{\mathrm{SR},k}(z) \leftarrow p_{\mathrm{SR}}(u_k(z))$
    $y_k(z) \leftarrow (1-\alpha_k) y_{\mathrm{H},k}(z) + \alpha_k y_{\mathrm{SR},k}(z)$
    Form $J_{\mathrm{H}}(u_k(z))$ using (10)
    Evaluate $J_{\mathrm{SR}}(u_k(z))$ using automatic differentiation
    Form $J_{\mathrm{OA},k}(u_k(z))$ using (11)
    Compute $r_k(y_{\mathrm{H},k}(z))$ from (12)
  end for
  $\mathcal{R}_k \leftarrow \left(|\mathcal{B}_k|^{-1} \sum_{z \in \mathcal{B}_k} r_k(z)^2\right)^{1/2}$ ▷ detach and store
  Compute $\widehat{G}_k$ as $\frac{1}{|\mathcal{B}_k|} \sum_{z \in \mathcal{B}_k} J_{g_{\theta_k}}(z)^{\mathsf{T}} J_{\mathrm{OA},k}(u_k(z))^{\mathsf{T}} \nabla_y \ell(y_k(z), z)$,
  $\theta_{k+1} \leftarrow \theta_k - \gamma_k \widehat{G}_k$
end for
return θ_K

4 Theoretical Analysis

This section establishes the theoretical guarantees for the proposed OA-projection framework, tracing its geometric effects from stochastic parameter updates to constrained decision limits. We first prove the convergence of Algorithm 1 under smooth loss functions, and subsequently demonstrate that its limiting decisions satisfy first-order KKT optimality conditions. Together, these results confirm that OA projection systematically improves the quality of the learned decisions.

4.1 Convergence under Stochastic Gradient Descent

We analyze the convergence of Algorithm 1 under common assumptions for SGD. Let $\mathcal{F}_k$ denote the information available before the mini-batch $\mathcal{B}_k$ is sampled, so that θ_k, $\mathcal{R}_{k-1}$, and α_k are $\mathcal{F}_k$-measurable. The regularity conditions are stated below.

Assumption 1 (Step sizes)**.** *The step-size sequence $\{\gamma_k\}_{k \geq 0}$ satisfies the following conditions,*

$$\gamma_k > 0, \qquad \sum_{k=0}^{\infty} \gamma_k = \infty, \qquad \sum_{k=0}^{\infty} \gamma_k^2 < \infty. \tag{13}$$

Assumption 2 (Bounded Iterates)**.** *The iterates $\{\theta_k\}_{k \geq 0}$ generated by Algorithm 1 are bounded and lie within a compact set $\mathcal{K}_\theta \subset \mathbb{R}^{n_\theta}$ almost surely.*

Assumption 3 (Model regularity)**.** *For $\mathcal{D}$-almost every z, the map $\theta \mapsto g_\theta(z)$ is path differentiable with conservative Jacobian $\mathcal{J}_g(\theta, z)$. Moreover, there exist measurable envelopes $B_u(z)$ and $B_g(z)$ such that*

$$\sup_{\theta \in \mathcal{K}_\theta} \|g_\theta(z)\|_2 \leq B_u(z), \quad \sup_{\theta \in \mathcal{K}_\theta} \sup_{J \in \mathcal{J}_g(\theta, z)} \|J\|_2 \leq B_g(z). \tag{14}$$

and B_u and B_J have the finite moments required by the results below.

Suppose the loss function ℓ is differentiable. Since $\mathcal{B}_k$ consists of conditionally independent draws from $\mathcal{D}$, the mini-batch direction in Algorithm 1 satisfies

$$\mathbb{E}[\widehat{G}_k \mid \mathcal{F}_k] = \mathbb{E}_{z\sim\mathcal{D}}\Big[J_{g_{\theta_k}}(z)^{\mathsf{T}} J_{\mathrm{OA}}(g_{\theta_k}(z);\alpha_k)^{\mathsf{T}} \nabla_y \ell\big(p_{\mathrm{OA}}(g_{\theta_k}(z);\alpha_k), z\big)\Big] =: H_{\alpha_k}(\theta_k), \quad (15)$$

where $J_{\mathrm{OA}}(\cdot;\alpha_k)$ is specialized in (11) and $J_{g_{\theta_k}}$ is a measurable selection of $\mathcal{J}_{g_\theta}$. Thus, $H_{\alpha_k}(\theta_k)$ is precisely the mean drift of the parameter update at iteration k. The convergence argument rests on how this drift behaves as the radial correction vanishes. Under the assumptions above, together with the regularity of the OA projection, $H_\alpha(\theta)$ exhibits graph consistency with a conservative field $D_{\mathrm{H}}F_{\mathrm{H}}$ of the orthogonal-projection objective. Hence, as α_k vanishes, the OA updates become asymptotically governed by the orthogonal-projection dynamics. This yields the following convergence result.

Theorem 5 (Convergence guarantee)**.** *Suppose Assumptions 1-3 hold. Suppose further that, for almost every z, $\ell(\cdot, z)$ is continuously differentiable, and that the stochastic direction inside the expectation in (15) has uniformly bounded second moments over $\mathcal{K}_\theta \times [0,1]$. Let $\mathcal{S}_{\mathrm{H}} := \{\theta \in \mathcal{K}_\theta : 0 \in D_{\mathrm{H}}F_{\mathrm{H}}(\theta)\}$ denote the stationary point set of the objective F_{H} in (2) and suppose that $F_{\mathrm{H}}(\mathcal{S}_{\mathrm{H}})$ has empty interior. Then we have*

$$\mathrm{clust}(\theta_k) \subseteq \mathcal{S}_{\mathrm{H}} \qquad \textit{almost surely}. \quad (16)$$

Theorem 5 demonstrates that OA projection alters the optimization trajectory while leaving the stationary set of OP invariant and its proof is given in Appendix B. This establishes the optimality regulated correction as a transient mechanism within the stochastic dynamics, reshaping the training path without altering the learning target or compromising SGD convergence guarantees.

4.2 Convergence to KKT Points

The preceding analysis guarantees the convergence of Algorithm 1 to the stationary set induced by OP. OA projection sharpens this conclusion through its residual-activated radial correction, which prevents persistent non-KKT iterates from becoming attainable limits. We first analyze this mechanism directly in the output space to decouple its behavior from specific model parameterizations. Fixing z and specializing g_θ in Algorithm 1 to the identity model on $\mathcal{A}$ with a singleton batch. With $y_k := p_{\mathrm{OA},k}(u_k)$, the update becomes

$$u_{k+1} = u_k - \gamma_k J_{\mathrm{OA},k}(u_k)^{\mathsf{T}} \nabla_y \ell(y_k, z). \quad (17)$$

This specialization preserves the same projection map, backward rule, and optimality feedback as the end-to-end training algorithm. The key insight behind the KKT guarantee is normal visibility. At a spurious stationary point $(\bar{u}, \bar{y})$, OP transmits only tangent variations of the active face, so a nonzero normal component of the decision gradient can disappear from the backward update. The SR branch restores sensitivity in this suppressed direction. Specifically, for the active face F, the analysis in Appendix C.1 shows that,

$$P_{\mathcal{N}_F} J_{\mathrm{SR}}(\bar{u})^{\mathsf{T}} \nabla_y \ell(\bar{y}, z) \neq 0. \quad (18)$$

This normal visibility implies that whenever the candidate is non-KKT, the OA backward direction retains an escape component that is absent under OP. The KKT residual remains positive in a neighborhood of such a point and therefore keeps the radial correction activated. Under a non-summable cumulative correction, the resulting persistent normal motion is incompatible with convergence to the non-KKT candidate, which leads to the KKT consistency result below.

Theorem 6 (KKT consistency of OA projection)**.** *Let $\{u_k\}$ be generated by (17) and suppose $u_k \to \bar{u}$. Assume $\ell(\cdot, z)$ is continuously differentiable and $\bar{u}$ is stationary for $\ell(\Pi_{\mathcal{C}}(u))$. Set $\bar{y} := \Pi_{\mathcal{C}}(\bar{u})$, and let F be the face satisfying $\bar{y} \in \mathrm{relint}(F)$. Suppose that $\bar{u} - \bar{y} \in \mathrm{ri}\, N_{\mathcal{C}}(F)$ and $N_{\mathcal{C}}(q_{\mathrm{R}}(\bar{u})) \subseteq T_F^{\perp}$. If $\sum_{k=0}^{\infty} \gamma_k \eta_k = \infty$, then $\bar{y}$ satisfies the KKT condition below.*

$$0 \in \nabla_y \ell(\bar{y}, z) + N_{\mathcal{C}}(\bar{y}). \quad (19)$$

Theorem 6 furnishes the optimality guarantee of OA projection by upgrading surrogate stationarity to true first-order KKT optimality and its proof defers to Appendix C. Under the cumulative activation condition $\sum_{k=0}^{\infty} \gamma_k \eta_k = \infty$, OA projection systematically excludes spurious non KKT candidates from the attainable limit set, proving that OA projection restores optimization fidelity in output space.

We next complement the model-independent KKT analysis with a concrete learning parameterization. Consider linear probing, a widely used transfer-learning setting with frozen pretrained features and

a trainable final head (Kumar et al., 2022). Its low adaptation cost and preservation of pretrained representations make it a practically important strategy for downstream transfer. Let $h_0 : \mathcal{Z} \to \mathbb{R}^m$ denote the frozen feature extractor. With a trainable affine head, the predictor is formed below.

$$g_\theta(z) = W h_0(z) + b = \widetilde{W}\widetilde{h}_0(z), \qquad \widetilde{h}_0(z) := \begin{bmatrix} h_0(z)^\top & 1 \end{bmatrix}^\top, \qquad \theta := \mathrm{vec}(\widetilde{W}), \tag{20}$$

where $W \in \mathbb{R}^{d\times m}$ and $b \in \mathbb{R}^d$ constitute the trainable linear head. For the training set $\mathcal{Z}_N = \{z_1, \dots, z_N\}$, let $\widetilde{H} \in \mathbb{R}^{N\times(m+1)}$ collect the augmented frozen features, with the ith row given by $\widetilde{h}_0(z_i)^\mathsf{T}$. We assume $\widetilde{H}\widetilde{H}^\mathsf{T} \succ 0$. This nondegeneracy condition ensures that the trainable head retains first-order control of the training outputs, enabling the output-space escape mechanism to persist in the parameterized model and yielding the following KKT guarantee.

Theorem 7 (Empirical KKT consistency under linear probing)**.** *Given a training dataset $\{z_i\}_{i=1}^N$, assume each $\ell(\cdot, z_i) \in C^1$, $\widetilde{H}\widetilde{H}^\mathsf{T} \succ 0$, and $\sum_{k=0}^\infty \gamma_k\eta_k = \infty$. Let $\theta_k \to \bar\theta$ a.s. under Algorithm 1 with predictor (20), where $\bar\theta$ is a.s. stationary for $F_{\mathrm{H}}^N(\theta) \triangleq \frac{1}{N}\sum_{i=1}^N \ell\big(p_{\mathrm{H}}(g_\theta(z_i)), z_i\big)$. If almost surely for every $i = 1, \dots, N$, the points $\bar u_i := g_{\bar\theta}(z_i)$ and $\bar y_i := \Pi_{\mathcal{C}}(\bar u_i) \in \mathrm{relint}(F_i)$ satisfy $\bar u_i - \bar y_i \in \mathrm{ri}\, N_{\mathcal{C}}(F_i)$ and $N_{\mathcal{C}}(q_{\mathrm{R}}(\bar u_i)) \subseteq T_{F_i}^\perp$, then the following KKT conditions hold a.s.,*

$$0 \in \nabla_y \ell(\bar y_i, z_i) + N_{\mathcal{C}}(\bar y_i). \tag{21}$$

Theorem 7 shows that the KKT guarantee extends directly to linear-probing parameterization, with its proof deferred to Appendix C. Taken together, these guarantees bridge the gap between surrogate stationarity and target optimality. By demonstrating that the OA projection eliminates normal gradient collapse across both output space updates and feature maps, these results confirm that strict constraint enforcement does not necessarily require sacrificing optimization fidelity. In this sense, OA projection retains OP as the feasibility map while endowing the training dynamics with the information needed to recover KKT-consistent solutions.

5 Experiments

This section evaluates whether OA projection translates into improved feasible decisions across direct variable optimization and parameterized linear probing settings. We compare against state-of-the-art constraint-enforcing baselines, including OP, SR (Schneider & Kuhn, 2026), RAYEN (Tordesillas et al., 2026), DC3 (Donti et al., 2021b), FSNet (Nguyen & Donti, 2025), and ΠNet (Grontas et al., 2026). Baseline selections are appropriately adapted to each task scenario for a fair comparison, with complete problem formulations and implementation details deferred to Appendix D.[1] We report constraint feasibility alongside the objective gap $f(y_{\mathrm{learned}}) - f(y_{\mathrm{solver}})$ relative to solver solutions. Across five benchmarks, the empirical results confirm that OA projection systematically delivers superior decision quality while maintaining strict constraint satisfaction.

5.1 Output-Space Optimization

We first directly optimize the raw variable, removing model parameterization to isolate map-induced optimization geometry. We compare OA with OP, SR, RAYEN, and ΠNet. DC3 and FSNet are reserved for linear probing since their iterative corrections would conflate map geometry.

Convex quadratic program. We adopt the parametric quadratic program (QP) generation scheme of DC3 with 100 decision variables, 50 equality constraints and 50 inequality constraints below.

$$\min_y \ \frac{1}{2} y^\mathsf{T} Q y + p^\mathsf{T} y, \qquad \text{s.t.} \quad Ay = x, \ \ Gy \le h.$$

Following the DC3 construction, we generate 10,000 QP instances, using 1,000 for validation and 9,000 for evaluation. Each method is evaluated over 3,000 optimization steps across five raw start seeds. We average the terminal objective gap over the 9,000 evaluation instances and report the mean and standard deviation across seeds.

[1] Source code will be made available upon acceptance of the paper.

As shown in Table 1, OA projection achieves the lowest mean objective gap while maintaining strict feasibility across all instances. It outperforms SR and RAYEN by 28.5% and 32.8% respectively, while exceeding OP and ΠNet by over 94%. These gains demonstrate that leveraging optimality feedback to regulate training geometry systematically improves solution quality.

Table 1: Direct optimization on the DC3-style QP benchmark.

Method	Mean gap ↓	Feas. (%)
OP	1.1933 ± 0.0101	100.0
ΠNet	1.1831 ± 0.0133	100.0
RAYEN	0.0892 ± 0.0002	100.0
SR	0.0837 ± 0.0003	100.0
OA	$\mathbf{0.0599 \pm 0.0013}$	100.0

Nonconvex program. We further stress the same setting by adding oscillatory nonconvex perturbations to the quadratic objective while keeping the constraint structure and initialization protocol unchanged. Across the independent program families, OA projection achieves the lowest mean objective gap of 0.0242, representing reductions of 64.5% relative to SR, the strongest competing baseline. SR and RAYEN obtain mean gaps of 0.0682 and 0.0695, respectively, whereas OP and ΠNet produce substantially larger gaps of 9.7081 and 9.9816. All methods remain feasible across all evaluated runs.

5.2 Linear-Probing Evaluation

We next consider the linear probing parameterization. Each task uses the same frozen feature map across methods and trains only the head layer. This protocol concentrates the comparison on how the constraint layer shapes the gradients received by the trainable head.

Constrained trajectory optimization. A trajectory is represented by eight two-dimensional control points, while the constraints impose boundary conditions, velocity and acceleration limits, as well as containment within convex safe corridors.

Table 2: Trajectory optimization over $10,000$ instances.

Method	Mean gap	Feas. (%)
OP	7.4778	100.00
ΠNet	7.6840	100.00
SR	10.9879	100.00
RAYEN	11.0111	100.00
FSNet	16.0044	96.85
DC3	20.7637	82.66
OA	**6.2863**	100.00

As shown in Table 2, OA projection attains the lowest mean objective gap of 6.2863 while maintaining 100% feasibility, achieving a 15.9% cost reduction over standard OP (7.4778). While ΠNet, SR, and RAYEN also preserve exact feasibility, they yield higher-cost solutions with gaps ranging from 7.68 to 11.01. In contrast, FSNet and DC3 degrade substantially in both solution quality and constraint enforcement, suffering feasibility failures on up to 17.34% of test instances.

DC optimal power flow (OPF). We further evaluate the IEEE 57-bus DC OPF problem under varying load profiles. OA projection reduces the mean relative objective gap from 0.4772% for OP to 0.3538%, a 25.9% reduction, while keeping feasible with tolerance 10^{-9}. SR and RAYEN also remain strictly feasible but produce larger relative gaps than OP. ΠNet attains a gap close to OP, while DC3 and FSNet exhibit larger constraint violations up to 8.55×10^{-3}.

Minimum-cost graph flow. Finally, we consider a capacitated layered logistics network with flow conservation and convex transportation costs. OA projection attains the lowest mean objective gap of 0.0810 alongside 100% feasibility, reducing the OP gap from 0.2260 by 64.1%. Although ΠNet achieves a competitive gap of 0.0934, it incurs constraint violations averaging 2.52×10^{-4}. Consequently, OA projection achieves the best optimization performance among all methods satisfying the physical network constraints.

6 Conclusion

We presented OA Projection, a novel feasibility-preserving mechanism that introduced a residual-regulated geometric correction during training to recover optimization directions suppressed by normal gradient collapse in OP. This construction bridged surrogate stationarity and first-order KKT optimality without altering the standard OP. We established algorithmic convergence and showed, under appropriate geometric conditions, that OA projection yielded KKT-consistent decisions. Experiments across diverse benchmarks confirmed improved decision quality while maintaining exact feasibility. As a transient modification of the training trajectory, OA projection remains compatible with many existing OP acceleration techniques. Extending the KKT guarantees to general parameterized learning models remains an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist in code editing and paper polishing. We have not used generative AI tools for core scientific tasks, including research ideation, experimental design, or data analysis, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to polish the manuscript text to improve readability.

We have thoroughly reviewed all AI-assisted work. Specifically, AI-generated code was manually reviewed, verified, and tested for correctness by the authors, and manuscript edits were carefully checked to ensure they accurately reflect our intended scientific content. We take full responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

A. Agrawal, S. Barratt, S. Boyd, E. Busseti, and W. Moursi. Differentiating through a cone program. *Journal of Applied and Numerical Optimization*, 1(2):107–115, 2019a.

Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b.

Brandon Amos. Tutorial on amortized optimization. *Foundations and Trends in Machine Learning*, 16(5):592–732, 06 2023. doi: 10.1561/2200000102. URL https://doi.org/10.1561/2200000102.

Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 136–145. PMLR, 06–11 Aug 2017.

Somrita Banerjee, Abhishek Cauligi, and Marco Pavone. Deep learning warm starts for trajectory optimization on the international space station. In *2025 International Conference on Space Robotics (iSpaRo)*, pp. 394–401, 2025. doi: 10.1109/iSpaRo66239.2025.11436159.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-Lopez, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5230–5242. Curran Associates, Inc., 2022. doi: 10.52202/068431-0378.

Jérôme Bolte, Tam Le, Edouard Pauwels, and Tony Silveti-Falls. Nonsmooth implicit differentiation for machine-learning and optimization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 13537–13549. Curran Associates, Inc., 2021.

Wenbo Chen, Mathieu Tanneau, and Pascal Van Hentenryck. End-to-end feasible optimization proxies for large-scale economic dispatch. *IEEE Transactions on Power Systems*, 39(2):4723–4734, 2024. doi: 10.1109/TPWRS.2023.3317352.

Alp Dener, Marco Andres Miller, Randy Michael Churchill, Todd Munson, and Choong-Seock Chang. Training neural networks under physical constraints using a stochastic augmented lagrangian approach, 2020. URL https://arxiv.org/abs/2009.07330.

Josip Djolonga and Andreas Krause. Differentiable learning of submodular models. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/192fc044e74dffea144f9ac5dc9f3395-Paper.pdf.

Priya L. Donti, Melrose Roderick, Mahyar Fazlyab, and J Zico Kolter. Enforcing robust control guarantees within neural network policies. In *International Conference on Learning Representations*, 2021a.

Priya L. Donti, David Rolnick, and J Zico Kolter. DC3: A learning method for optimization with hard constraints. In *International Conference on Learning Representations*, 2021b.

Juan Elenter, Luiz F. O. Chamon, and Alejandro Ribeiro. Near-optimal solutions of constrained learning problems. In *The Twelfth International Conference on Learning Representations*, 2024.

Herbert Federer. *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften. Springer, 1969. URL https://cir.nii.ac.jp/crid/1971993809742369702.

Thomas Frerix, Matthias Niessner, and Daniel Cremers. Homogeneous linear inequality constraints for neural network activations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Arnob Ghosh, Xingyu Zhou, and Ness Shroff. Towards achieving sub-linear regret and hard constraint violation in model-free RL. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1054–1062. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/ghosh24a.html.

Stephen Gould, Richard Hartley, and Dylan Campbell. Deep declarative networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):3988–4004, 2022. doi: 10.1109/TPAMI.2021.3059462.

Panagiotis D. Grontas, Antonio Terpin, Efe C. Balta, Raffaello D'Andrea, and John Lygeros. Pinet: Optimizing hard-constrained neural networks with orthogonal projection layers. In *The Fourteenth International Conference on Learning Representations*, 2026.

Fouad Hasan, Amin Kargarian, and Ali Mohammadi. A survey on applications of machine learning for optimal power flow. In *2020 IEEE Texas Power and Energy Conference (TPEC)*, pp. 1–6, 2020. doi: 10.1109/TPEC48276.2020.9042547.

Hooman Khaloie, Mihály Dolányi, Jean-François Toubeau, and François Vallée. Review of machine learning techniques for optimal power flow. *Applied Energy*, 388:125637, 2025. ISSN 0306-2619. doi: https://doi.org/10.1016/j.apenergy.2025.125637. URL https://www.sciencedirect.com/science/article/pii/S0306261925003678.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Jay Yoon Lee, Sanket Vaibhav Mehta, Michael Wick, Jean-Baptiste Tristan, and Jaime Carbonell. Gradient-based inference for networks with output constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4147–4154, 2019.

Xudong Li, Defeng Sun, and Kim-Chuan Toh. On the efficient computation of a generalized jacobian of the projector over the birkhoff polytope. *Mathematical Programming*, 179(1):419–446, Jan 2020. ISSN 1436-4646. doi: 10.1007/s10107-018-1342-9. URL https://doi.org/10.1007/s10107-018-1342-9.

Enming Liang, Minghua Chen, and Steven Low. Low complexity homeomorphic projection to ensure neural-network solution feasibility for optimization over (Non-)Convex set. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20623–20649. PMLR, 23–29 Jul 2023.

Enming Liang, Minghua Chen, and Steven H. Low. Homeomorphic projection to ensure neural-network solution feasibility for constrained optimization. *Journal of Machine Learning Research*, 25(329):1–55, 2024. URL http://jmlr.org/papers/v25/23-1577.html.

Han Liu, Zipeng Fu, Kai Yang, Xinyi Xu, and Mathieu Bauchy. Machine learning for glass science and engineering: A review. *Journal of Non-Crystalline Solids*, 557:119419, 2021. ISSN 0022-3093. doi: https://doi.org/10.1016/j.jnoncrysol.2019.04.039. URL https://www.sciencedirect.com/science/article/pii/S0022309319302662.

Youngjae Min and Navid Azizan. Hardnet: Hard-constrained neural networks with universal approximation guarantees, 2025. URL https://arxiv.org/abs/2410.10807.

Simone Monaco, Flavio Giobergia, Alkis Koudounas, and Daniele Apiletti. Enforcing domain constraints through lagrangian primal-dual learning. *Neural Computing and Applications*, 38(7): 191, Mar 2026. ISSN 1433-3058. doi: 10.1007/s00521-026-11880-z. URL https://doi.org/10.1007/s00521-026-11880-z.

Yatin Nandwani, Abhishek Pathak, Mausam, and Parag Singla. A primal dual formulation for deep learning with constraints. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/cf708fc1decf0337aded484f8f4519ae-Paper.pdf.

Harikrishna Narasimhan, Andrew Cotter, Yichen Zhou, Serena Wang, and Wenshuo Guo. Approximate heavily-constrained learning with lagrange multiplier models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8693–8703. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/62db9e3397c76207a687c360e0243317-Paper.pdf.

Sankaranarayanan Natarajan. Learning initial trajectory using sequence-to-sequence approach to warm start an optimization-based motion planner. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9430–9436, 2021. doi: 10.1109/IROS51168.2021.9636456.

Hoang Nguyen and Priya Donti. Fsnet: Feasibility-seeking neural network for constrained optimization with guarantees. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 39670–39708. Curran Associates, Inc., 2025. doi: 10.52202/085713-1323.

Jianming Pan, Zeqi Ye, Xiao Yang, Xu Yang, Weiqing Liu, Lewen Wang, and Jiang Bian. Bpqp: A differentiable convex optimization framework for efficient end-to-end learning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 77468–77493. Curran Associates, Inc., 2024. doi: 10.52202/079017-2463.

Xiang Pan, Minghua Chen, Tianyu Zhao, and Steven H. Low. Deepopf: A feasibility-optimized deep neural network approach for ac optimal power flow problems. *IEEE Systems Journal*, 17(1): 673–683, 2023. doi: 10.1109/JSYST.2022.3201041.

Seonho Park and Pascal Van Hentenryck. Self-supervised primal-dual learning for constrained optimization. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i4.25520.

Luis Pineda, Taosha Fan, Maurizio Monge, Shobha Venkataraman, Paloma Sodhi, Ricky T. Q. Chen, Joseph Ortiz, Daniel DeTone, Austin Wang, Stuart Anderson, Jing Dong, Brandon Amos, and Mustafa Mukadam. Theseus: a library for differentiable nonlinear optimization. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Rajiv Sambharya, Georgina Hall, Brandon Amos, and Bartolomeo Stellato. Learning to warm-start fixed-point optimization algorithms. *Journal of Machine Learning Research*, 25(166):1–46, 2024. URL http://jmlr.org/papers/v25/23-1174.html.

Philipp J. Schneider and Daniel Kuhn. Soft-radial projection for constrained end-to-end learning, 2026.

Mathieu Tanneau and Pascal Van Hentenryck. Dual lagrangian learning for conic optimization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Jesus Tordesillas, Victor Klemm, Jonathan P. How, and Marco Hutter. Rayen: Imposition of hard convex constraints on neural networks, 2026.

Runzhong Wang, Yunhao Zhang, Ziao Guo, Tianyi Chen, Xiaokang Yang, and Junchi Yan. LinSATNet: The positive linear satisfiability neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36605–36625. PMLR, 23–29 Jul 2023.

Shaoqi Wang, Chunjie Yang, and Siwei Lou. Approximated orthogonal projection unit: Stabilizing regression network training using natural gradient. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 42494–42527. Curran Associates, Inc., 2024. doi: 10.52202/079017-1346. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/4ae7d78ebbe48f772e31c5c3fcc04c43-Paper-Conference.pdf.

Yuwei Wu, Xiatao Sun, Igor Spasojevic, and Vijay Kumar. Deep learning for optimization of trajectories for quadrotors. *IEEE Robotics and Automation Letters*, 9:2479–2486, 03 2024. doi: 10.1109/LRA.2024.3357399.

Zonglin Yang, Zhexuan Gu, and Yancheng Yuan. Efficient and provably convergent end-to-end training of deep neural networks with linear constraints, 2026. URL https://arxiv.org/abs/2605.11526.

Hongtai Zeng, Chao Yang, Yanzhen Zhou, Cheng Yang, and Qinglai Guo. Glinsat: The general linear satisfiability neural network layer by accelerated gradient descent. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 122584–122615. Curran Associates, Inc., 2024. doi: 10.52202/079017-3896.

Lin Zeng, Hsiao-Dong Chiang, and Dong Liang. Toward minimal number of controls for thermal violation mitigation for very large-scale grids. *IEEE Transactions on Power Systems*, 41(2):1498–1509, 2026. doi: 10.1109/TPWRS.2025.3614844.

Shiliang Zhang, Hui Jiang, and Lirong Dai. Hybrid orthogonal projection and estimation (hope): A new framework to learn neural networks. *Journal of Machine Learning Research*, 17(37):1–33, 2016. URL http://jmlr.org/papers/v17/15-335.html.

Shuai Zhang, Oluwaseun T. Ajayi, and Yu Cheng. A self-supervised learning approach for accelerating wireless network optimization. *IEEE Transactions on Vehicular Technology*, 72(6):8074–8087, 2023. doi: 10.1109/TVT.2023.3244043.

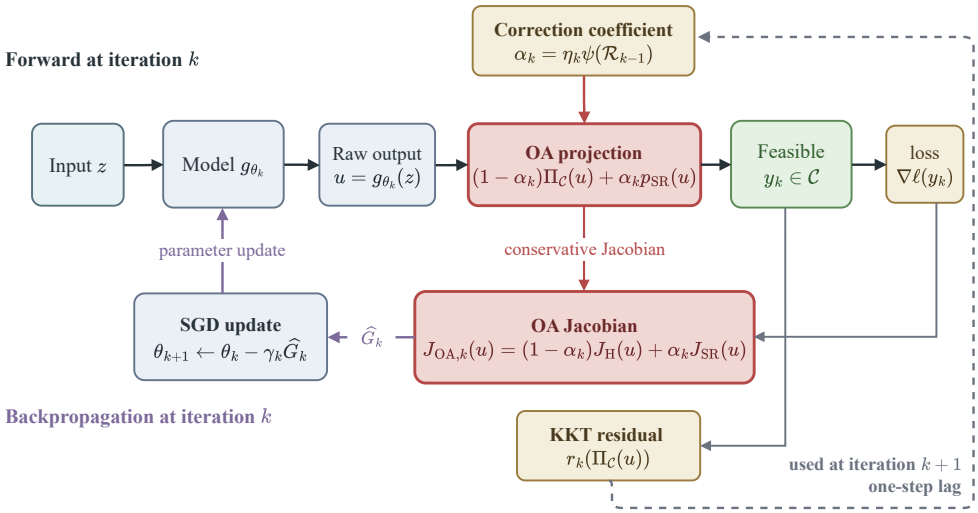
A Analysis of the OA Projection Map

A.1 Mechanistic Analysis of OA Projection

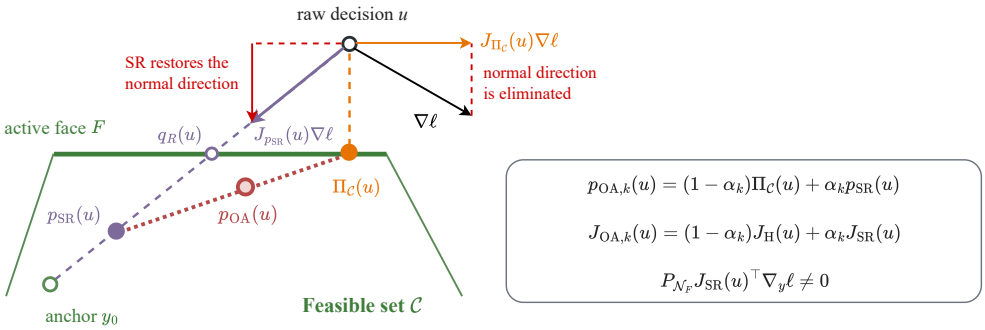
Figure 1 presents both the algorithmic pipeline and local projection geometry of the proposed OA mechanism. Specifically, the training framework dynamically regulates radial corrections via a lagged KKT residual (Figure 1a), while the underlying SR component restores normal gradient sensitivity along active constraint faces (Figure 1b). To elucidate this mechanism, we first analyze optimization trajectories on a two-dimensional quadratic problem before formalizing the mathematical regularity of the OA projection.

With the context fixed, we write the decision loss as $f(y)$ and consider the following program.

$$\begin{aligned} \min_{z\in\mathbb{R}^2} \quad & f(z) := z^\mathsf{T} H z + g^\mathsf{T} z + c, \\ \text{s.t.} \quad & -1 \le z_1 \le 1, \quad -1 \le z_2 \le 1, \end{aligned} \tag{22}$$



(a) Overall OA training framework with adaptive correction and conservative-Jacobian backpropagation.



(b) Geometric interpretation of the OA projection: the SR component restores the normal-direction information eliminated by the standard orthogonal projection.

Figure 1: Overview and geometric interpretation of the proposed OA mechanism. (a) OA projection is integrated into the training loop through adaptive correction, feasible projection, and well-defined conservative Jacobian. (b) Geometrically, the SR component restores normal-direction information that is suppressed by the standard orthogonal projection.

where

$$H = \begin{bmatrix} 1 & -0.3 \\ -0.3 & 1 \end{bmatrix}, \quad g = \begin{bmatrix} 2.3 \\ -0.4 \end{bmatrix}, \quad c = 1.35.$$

The unconstrained minimizer is $z^{\mathrm{unc}} = -\frac{1}{2}H^{-1}g$, which lies outside the feasible box beyond the left face. The unique constrained solution over $\mathcal{C}$ is $z^\star = (-1, -0.1)^\mathsf{T}$, with gradient $\nabla f(z^\star) = (0.36, 0)^\mathsf{T}$. Because $-\nabla f(z^\star) \in N_{\mathcal{C}}(z^\star) = \{-\lambda e_1 : \lambda \geq 0\}$, this boundary point satisfies the KKT optimality conditions.

To isolate the effect of the projection map, we optimize the raw variable directly without a learning model. For each method $M \in \{\mathrm{OP}, \mathrm{SR}, \mathrm{OA}\}$, let $p_{M,k}$ denote its feasible-output map and $J_{M,k}$ denote the corresponding Jacobian used in backpropagation. Starting from the common raw initialization $u_0 = (1.8, -0.4)^\mathsf{T}$, the trajectories of raw states u_k^M and feasible outputs y_k^M are generated by

$$y_k^M = p_{M,k}(u_k^M), \qquad u_{k+1}^M = u_k^M - \gamma J_{M,k}(u_k^M)^\mathsf{T} \nabla f(y_k^M), \tag{23}$$

where $\gamma > 0$ denotes the optimization step size. Figure 2 compares the resulting optimization

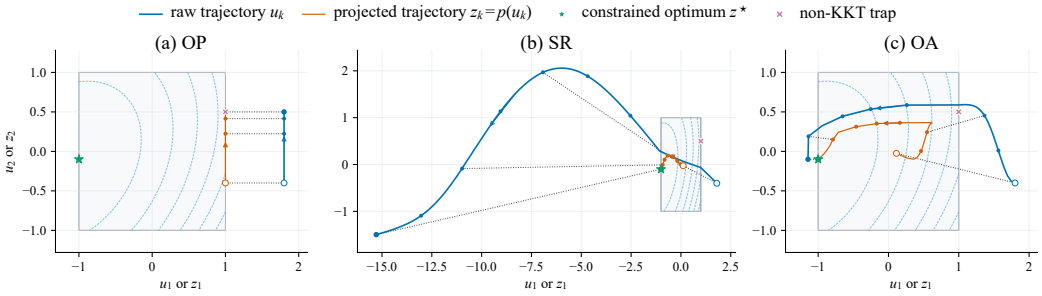


Figure 2: Raw and feasible trajectories for the quadratic problem in (22). (a) OP gets trapped at the non-KKT stationary point $z^{\text{trap}} = (1, 0.5)^\mathsf{T}$ due to normal gradient suppression along the active face. (b) SR projection escapes the trap and approaches $z^\star$, but thousands of iterations and forces raw preimage norm explosion. (c) OA projection uses residual-regulated radial correction to escape the non-KKT face, rapidly reaching the exact solution $z^\star$ in just 93 iterations with a finite raw state. Dotted segments link raw and feasible states at identical iterations.

trajectories generated under update rule (23). Blue curves trace raw iterates u_k^M, orange curves trace the corresponding feasible outputs y_k^M, and dotted segments connect state pairs at identical iterations.

For OP, the initial feasible output is $\Pi_{\mathcal{C}}(u_0) = (1, -0.4)^\mathsf{T}$. On the relative interior of the right face defined below, we obtain the following Jacobian

$$F_+ := \{(1, t) : -1 < t < 1\}, \qquad J_{\text{H}} = P_{T_{F_+}} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

Evaluating the objective gradient yields $\nabla f(1, t) = [4.3 - 0.6t \quad 2t - 1]^\top$. Consequently, back-propagation through J_{H} updates only the vertical coordinate, driving the feasible trajectory directly into the trap $z^{\text{trap}} = (1, 0.5)^\mathsf{T}$. At this point, $J_{\text{H}}^\mathsf{T} \nabla f(z^{\text{trap}}) = 0$ and $\nabla f(z^{\text{trap}}) = (4, 0)^\mathsf{T}$, yet $-\nabla f(z^{\text{trap}}) = (-4, 0)^\mathsf{T} \notin N_{\mathcal{C}}(z^{\text{trap}}) = \{\lambda e_1 : \lambda \geq 0\}$. As shown in Figure 2(a), OP prematurely terminates at a projection-induced non-KKT point because its Jacobian eliminates the horizontal gradient required to pull the raw iterate away from the active boundary.

SR projection retains a radial derivative with a nonzero component along the missing face-normal direction. It consequently escapes the OP trap and advances the feasible trajectory toward $z^\star$, as depicted in Figure 2(b). However, because the SR projection maps any finite raw state into the relative interior of the feasible set, exact boundary attainment requires the raw norm to explode. Thus, while SR projection avoids the boundary collapse of OP, it represents the exact boundary solution only in the limit of an unbounded raw sequence, risking numerical instability.

OA projection addresses both pitfalls by deploying radial correction conditionally through optimality feedback. Near z^{trap}, the non-zero lagged KKT residual maintains an active correction channel, allowing the OA Jacobian to retain horizontal sensitivity and drive the raw state off the right-face normal fiber. As the feasible output approaches the true boundary optimum $z^\star$, the KKT residual vanishes, driving $\alpha_k \to 0$ and seamlessly restoring pure OP. As illustrated in Figure 2(c), OA exhibits rapid convergence. It reaching the exact constrained minimizer $z^\star$ at iteration 93 with a finite raw state $u_{93} \approx (-1.144, -0.1)^\mathsf{T}$, where $p_{\text{OA}}(u_{93}) = \Pi_{\mathcal{C}}(u_{93}) = z^\star$. This behavior instantiates the dynamic control loop in Figure 1a: optimality feedback activates radial sensitivity strictly when required, deactivating it once KKT consistency is re-established.

The comparison isolates the roles of the two component maps. OP directly represents boundary decisions but can suppress the normal information needed to leave a non-KKT face. The SR map preserves this information but approaches boundary decisions through increasingly large raw states. OA projection regulates the radial component through optimality feedback, retaining its escape direction during optimization and recovering the boundary-attaining geometry of OP once the correction is no longer needed.

A.2 REGULARITY OF OA PROJECTION

Let $\mathcal{L} := \mathrm{lin}(\mathcal{A}-\mathcal{A})$ be the direction space of $\mathcal{A}$. All derivatives of maps defined on $\mathcal{A}$ are understood relative to $\mathcal{L}$. Without loss of generality, we assume $By = b$ is a full row rank representation of $\mathcal{A} = \mathrm{aff}(\mathcal{C})$ that absorbs all implicit linear equalities and eliminates redundant constraints. Under this minimal representation, selecting $y_0 \in \mathrm{relint}(\mathcal{C})$ guarantees $\delta_i := a_i - A_i y_0 > 0$ for every remaining inequality constraint. After this removal, the relative-interior condition $y_0 \in \mathrm{ri}(\mathcal{C})$ gives $\delta_i := a_i - A_i y_0 > 0$ for every remaining inequality.

Lemma 8 (Radial representation and regularity)**.** *For $u \in \mathcal{A}$, write $v := u - y_0$ and define $s(v) := \max\left\{1, \max_i \frac{A_i v}{\delta_i}\right\}$. Then we have the explicit representation of the scale and endpoint in (6) as follows,*

$$\lambda^\star(u) = \frac{1}{s(v)}, \qquad q_{\mathrm{R}}(u) = y_0 + \frac{v}{s(v)}. \tag{24}$$

Consequently, q_{R} is locally Lipschitz on $\mathcal{A}$ and differentiable almost everywhere. Furthermore, whenever a single affine term uniquely attains the maximum in $s(u - y_0)$, q_{R} is continuously differentiable in a neighborhood of u.

Proof. Since $u, y_0 \in \mathcal{A}$, the equality constraints are preserved along the segment $y_0 + tv$. For an inequality indexed by i, the condition $A_i(y_0 + tv) \le a_i$ is equivalent to $tA_i v \le \delta_i$. If $A_i v \le 0$, this inequality imposes no upper bound on $t \in [0, 1]$. If $A_i v > 0$, it requires $t \le \delta_i/(A_i v)$. Therefore,

$$\lambda^\star(u) = \min\left\{1, \min_{i:A_i v>0} \frac{\delta_i}{A_i v}\right\} = \left(\max\left\{1, \max_i \frac{A_i v}{\delta_i}\right\}\right)^{-1},$$

which proves (24).

The function s is the maximum of finitely many affine functions and is therefore globally Lipschitz on $\mathcal{L}$. Moreover, $s(v) \ge 1$. On every bounded neighborhood $V \subset \mathcal{L}$, the estimate $|s(v_1)^{-1} - s(v_2)^{-1}| \le |s(v_1) - s(v_2)|$ shows that $1/s$ is Lipschitz. The product $v \mapsto v/s(v)$ is thus Lipschitz on V, proving the local Lipschitz continuity of q_{R}. Rademacher's theorem gives differentiability almost everywhere (Federer, 1969). Finally, suppose that a single affine term uniquely attains the maximum in $s(\bar{v})$. Continuity of the finitely many affine components implies that this same term remains the unique maximizer throughout a neighborhood of $\bar{v}$, so s reduces locally to a single affine function. Because $s \ge 1$, the map $v \mapsto v/s(v)$ is continuously differentiable on this neighborhood, establishing the local smoothness of q_{R}. □

Lemma 9 (Conservative Jacobian of polyhedral projection)**.** *Define the set-valued mapping $\mathcal{J}_{\mathrm{H}}(u)$ as follows,*

$$\mathcal{J}_{\mathrm{H}}(u) := \operatorname*{Limsup}_{v \to u}\{J_{\mathrm{H}}(v)\} = \left\{J \in \mathbb{R}^{d\times d} : \exists\, v_j \to u \text{ such that } J_{\mathrm{H}}(v_j) \to J\right\}. \tag{25}$$

Then $\mathcal{J}_{\mathrm{H}}$ is a nonempty conservative Jacobian of $\Pi_{\mathcal{C}}$, and $J_{\mathrm{H}}(u) \in \mathcal{J}_{\mathrm{H}}(u)$ for every u. Specifically, $J_{\mathrm{H}}(u)$ is the conservative Jacobian selection in (10).

Proof. Let $\mathfrak{F}(\mathcal{C})$ denote the finite collection of nonempty faces of $\mathcal{C}$, including $\mathcal{C}$ itself. For each face F, write $T_F := \mathrm{span}(F - F)$ for its parallel subspace and let $N_{\mathcal{C}}(F)$ denote the normal cone on $\mathrm{relint}(F)$. Define $\mathcal{R}_F := F + N_{\mathcal{C}}(F)$ as the region obtained by extending F along its normal cone. For any $u = z + n \in \mathcal{R}_F$ with $z \in F$ and $n \in N_{\mathcal{C}}(F)$, the minimal distance definition of orthogonal projection implies $\Pi_{\mathcal{C}}(u) = z$. Since $n \perp T_F$, z is also the orthogonal projection of u onto $\mathrm{aff}(F)$. Therefore, fixing $z_F \in F$, the following relationship holds,

$$\Pi_{\mathcal{C}}(u) = \Pi_{\mathrm{aff}(F)}(u) = z_F + P_{T_F}(u - z_F), \qquad u \in \mathcal{R}_F. \tag{26}$$

Thus, on $\mathcal{R}_F$, the projection agrees with an affine map whose linear part is P_{T_F}.

We next relate this facewise representation to the selection J_{H}. For any u, set $y = \Pi_{\mathcal{C}}(u)$ and let F_y be the unique face containing y in its relative interior. The active constraints at y satisfy $\ker M_{\mathrm{H}}(u) = T_{F_y}$. For any matrix M, the following identity holds.

$$M^{\mathsf{T}}(MM^{\mathsf{T}})^{\dagger}M = P_{\mathrm{range}(M^{\mathsf{T}})} = P_{(\ker M)^{\perp}}$$

Consequently, the pseudoinverse representation in (10) yields $J_{\mathrm{H}}(u) = P_{\ker M_{\mathrm{H}}(u)} = P_{T_{F_y}}$. Furthermore, $u - y \in N_{\mathcal{C}}(F_y)$ implies $u \in \mathcal{R}_{F_y}$.

We now establish that the outer limit definition reduces to the finite set of face projections below.

$$\mathcal{J}_{\mathrm{H}}(u) = \{P_{T_F} : F \in \mathfrak{F}(\mathcal{C}),\ u \in \mathcal{R}_F\}. \tag{27}$$

Take any $J \in \mathcal{J}_{\mathrm{H}}(u)$. By (25), there exists a sequence $v_j \to u$ such that $J_{\mathrm{H}}(v_j) \to J$. Finiteness of $\mathfrak{F}(\mathcal{C})$ allows for the extraction of a subsequence whose projections $\Pi_{\mathcal{C}}(v_j)$ share a common minimal active face F. Hence $J_{\mathrm{H}}(v_j) = P_{T_F}$ and $v_j \in \mathcal{R}_F$. Closedness of $\mathcal{R}_F$ ensures $u \in \mathcal{R}_F$ and $J = P_{T_F}$.

Conversely, suppose $u = z + n \in \mathcal{R}_F$, with $z \in F$ and $n \in N_{\mathcal{C}}(F)$. Choose $z_j \in \mathrm{relint}(F)$ such that $z_j \to z$, and set $v_j := z_j + n$. Then $v_j \to u$ and $\Pi_{\mathcal{C}}(v_j) = z_j$, which implies $J_{\mathrm{H}}(v_j) = P_{T_F}$. Taking $v_j \to u$ yields $P_{T_F} \in \mathcal{J}_{\mathrm{H}}(u)$, proving (27).

Characterization (27) shows that $\mathcal{J}_{\mathrm{H}}(u)$ is finite, nonempty, and contains $J_{\mathrm{H}}(u)$ for every u. The graph of $\mathcal{J}_{\mathrm{H}}$, given by $\bigcup_{F \in \mathfrak{F}(\mathcal{C})} \mathcal{R}_F \times \{P_{T_F}\}$, is closed as a finite union of closed sets. Since all elements of $\mathcal{J}_{\mathrm{H}}(u)$ are orthogonal projections with operator norm at most one, the mapping is globally bounded.

Finally, we verify the conservative chain rule. Let $\gamma : [0,1] \to \mathbb{R}^d$ be absolutely continuous and set $y = \Pi_{\mathcal{C}} \circ \gamma$, which is also absolutely continuous by 1-Lipschitzness of $\Pi_{\mathcal{C}}$. For each face $F \in \mathfrak{F}(\mathcal{C})$, define the measurable set $E_F \triangleq \{t \in [0,1] \mid \gamma(t) \in \mathcal{R}_F\}$. By (26), y coincides with the affine map $\Pi_{\mathrm{aff}(F)} \circ \gamma$ on E_F, so their time derivatives agree almost everywhere on E_F with $\dot{y}(t) = P_{T_F}\dot{\gamma}(t)$. Finiteness of $\mathfrak{F}(\mathcal{C})$ guarantees that these identities hold simultaneously for almost every $t \in [0,1]$. Applying (27) confirms

$$\frac{d}{dt}(\Pi_{\mathcal{C}} \circ \gamma)(t) = J\dot{\gamma}(t), \qquad \forall J \in \mathcal{J}_{\mathrm{H}}(\gamma(t)), \quad \text{for almost every } t \in [0,1].$$

Therefore, $\mathcal{J}_{\mathrm{H}}$ constitutes a valid conservative Jacobian of $\Pi_{\mathcal{C}}$. □

A.3 Proof of the OA differentiability theorem

Proof of Theorem 4. By Lemma 8, the radial endpoint q_{R} is locally Lipschitz on $\mathcal{A}$ and continuously differentiable wherever a single affine term uniquely attains the maximum defining s. Since $u \mapsto \rho(\|u - y_0\|_2^2)$ is continuously differentiable, the representation (7) implies that p_{SR} is locally Lipschitz and admits a nonempty conservative Jacobian, denoted by $\mathcal{J}_{\mathrm{SR}}$. Moreover, p_{SR} is differentiable almost everywhere by Rademacher's theorem. Let $J_{\mathrm{SR}}(u) \in \mathcal{J}_{\mathrm{SR}}(u)$ denote a selected element, chosen to agree with the classical Jacobian wherever the latter exists.

By Lemma 9, orthogonal projection $\Pi_{\mathcal{C}}$ admits a nonempty conservative Jacobian $\mathcal{J}_{\mathrm{H}}$, and $J_{\mathrm{H}}(u) \in \mathcal{J}_{\mathrm{H}}(u)$ for every u, where $J_{\mathrm{H}}(u)$ is given by (10).

During iteration k, α_k is fixed with respect to u, so the set-valued mapping $\mathcal{J}_{\mathrm{OA},k}(u) := (1 - \alpha_k)\mathcal{J}_{\mathrm{H}}(u) + \alpha_k \mathcal{J}_{\mathrm{SR}}(u)$ is a nonempty conservative Jacobian of $p_{\mathrm{OA},k}$. Indeed, for any absolutely continuous curve γ, the conservative chain rules of the two component maps give, for almost every t, the following equation holds for every $J_{\mathrm{H}}^{\circ} \in \mathcal{J}_{\mathrm{H}}(\gamma(t))$ and $J_{\mathrm{SR}}^{\circ} \in \mathcal{J}_{\mathrm{SR}}(\gamma(t))$,

$$\frac{d}{dt}(p_{\mathrm{OA},k} \circ \gamma)(t) = \big((1 - \alpha_k)J_{\mathrm{H}}^{\circ} + \alpha_k J_{\mathrm{SR}}^{\circ}\big)\dot{\gamma}(t).$$

Hence, selecting $J_{\mathrm{H}}(u) \in \mathcal{J}_{\mathrm{H}}(u)$ and $J_{\mathrm{SR}}(u) \in \mathcal{J}_{\mathrm{SR}}(u)$ yields (11). □

B Proofs of the Convergence Results

B.1 A stochastic-approximation lemma

We begin with a proof of SGD convergence under a general consistent gradient drift.

Lemma 10 (Stochastic approximation with a graph-consistent drift)**.** *Let $D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a set valued map that is locally bounded with nonempty compact convex values and a closed graph. Let*

$f : \mathbb{R}^p \to \mathbb{R}$ be locally Lipschitz, and suppose that D is a conservative field for f. Define the critical set as $S \triangleq \{x \in \mathbb{R}^p \mid 0 \in D(x)\}$. Consider the update scheme below.

$$x_{k+1} = x_k - \gamma_k(h_k + M_{k+1}), \tag{28}$$

which is adapted to a filtration $\{\mathcal{F}_k\}$, such that x_k and h_k are $\mathcal{F}_k$-measurable. Assume $\{M_{k+1}\}$ is a martingale difference sequence satisfying the following inequality for some $C_M < \infty$.

$$\mathbb{E}[\|M_{k+1}\|_2^2 \mid \mathcal{F}_k] \le C_M \quad \textit{almost surely}. \tag{29}$$

Suppose that the iterates $\{x_k\}$ remain almost surely in a compact set, the step sizes obey (13), and $\{h_k\}$ is almost surely bounded and graph-consistent with D, which means that $h \in D(x)$ whenever a subsequence satisfies $x_{k_j} \to x$ and $h_{k_j} \to h$. If $f(S)$ has empty interior, $f(x_k)$ converges almost surely and we can obtain the following inclusion result,

$$\operatorname{clust}(x_k) \subseteq S \quad \textit{almost surely},$$

In addition, if S is finite, x_k converges almost surely to a single point in S.

Proof. The square-summability of $\{\gamma_k\}$ implies $\gamma_k \to 0$. Combining this with the moment bound in (29) yields the following result,

$$\sum_{k=0}^{\infty} \mathbb{E}\big[\|\gamma_k M_{k+1}\|_2^2 \mid \mathcal{F}_k\big] \le C_M \sum_{k=0}^{\infty} \gamma_k^2 < \infty.$$

By the martingale convergence theorem, the noise series $\sum_{k=0}^{\infty} \gamma_k M_{k+1}$ converges almost surely. For any $T > 0$ and all sufficiently large k, define the terminal index of the corresponding discrete-time window as follows,

$$m(k,T) := \max\left\{ m \ge k : \sum_{i=k}^{m} \gamma_i \le T \right\}.$$

The step-size conditions imply $\gamma_k \to 0$ and $\sum_{k=0}^{\infty} \gamma_k = \infty$, so this index is well defined and finite. Almost-sure convergence of $\sum_{k=0}^{\infty} \gamma_k M_{k+1}$ implies that the supremum of partial sums ending at $m(k,T)$ vanishes asymptotically as $k \to \infty$ below.

$$\sup_{k \le m \le m(k,T)} \left\| \sum_{i=k}^{m} \gamma_i M_{i+1} \right\|_2 \longrightarrow 0 \quad \text{almost surely}. \tag{30}$$

Hence, we fix a sample path on which $\{x_k\}$ and $\{h_k\}$ are bounded, the graph-consistency condition holds, and (30) is satisfied. We first establish that $\operatorname{dist}((x_k, h_k), \operatorname{gph} D) \to 0$. Suppose for the sake of contradiction that this limit does not hold. Then there exist a constant $\varepsilon > 0$ and a subsequence $\{k_j\}$ such that $\operatorname{dist}\big((x_{k_j}, h_{k_j}), \operatorname{gph} D\big) \ge \varepsilon$ for all j. By the uniform boundedness of $\{x_k\}$ and $\{h_k\}$, passing to a further subsequence which is still denoted by $\{k_j\}$ yields $x_{k_j} \to x$ and $h_{k_j} \to h$. Graph consistency of $\{h_k\}$ ensures $h \in D(x)$, which implies $(x, h) \in \operatorname{gph} D$. Continuity of the distance function then gives the identity below.

$$\lim_{j \to \infty} \operatorname{dist}\big((x_{k_j}, h_{k_j}), \operatorname{gph} D\big) = \operatorname{dist}((x,h), \operatorname{gph} D) = 0,$$

This contradicts the preceding lower bound ε. Thus $\operatorname{dist}((x_k, h_k), \operatorname{gph} D) \to 0$ holds.

Set $t_0 = 0$ and $t_k = \sum_{i=0}^{k-1} \gamma_i$. We define the continuous time interpolated processes on $[0, \infty)$ as follows. Let X be the piecewise-linear interpolation satisfying $X(t_k) = x_k$. Define the piecewise-constant interpolation $\bar{h}(t) := h_k$ for $t \in [t_k, t_{k+1})$, and let Z denote the piecewise-linear interpolation of the martingale partial sums as $Z(t_k) := \sum_{i=0}^{k-1} \gamma_i M_{i+1}$. On each interval $[t_k, t_{k+1})$, the update in (28) gives $\dot{X}(t) = -\bar{h}(t) - \dot{Z}(t)$. Therefore, integration over $[s, t]$ delivers the following trajectory equation,

$$X(t) - X(s) = -\int_s^t \bar{h}(r)\, dr - \big(Z(t) - Z(s)\big). \tag{31}$$

By the martingale convergence established previously, $Z(t)$ converges to a finite limit as $t \to \infty$, which implies that Z is uniformly continuous on $[0, \infty)$. Consequently, for any sequence of time

shifts $s_j \to \infty$, the shifted noise term $Z(s_j + \cdot) - Z(s_j)$ vanishes uniformly on compact intervals. Since $\bar{h}$ is uniformly bounded, identity (31) establishes the asymptotic equicontinuity of the shifted trajectory family $\{X(s_j + \cdot)\}_{j\geq 1}$ on every compact time interval. They are also uniformly bounded because X linearly interpolates the bounded sequence $\{x_k\}$. Therefore, for any sequence $s_j \to \infty$, the Arzelà–Ascoli theorem yields a subsequence which is not relabeled, such that $X(s_j + \cdot) \to \bar{X}(\cdot)$ uniformly on compact intervals. The uniform boundedness of $\bar{h}$ further ensures that $\bar{X}$ is locally Lipschitz continuous and consequently absolutely continuous.

We next identify the limiting dynamics of the shifted trajectories. For any fixed horizon $T > 0$, define the shifted interpolations $X_j(t) \triangleq X(s_j + t)$ and $h_j(t) \triangleq \bar{h}(s_j + t)$ for $t \in [0, T]$. Since $\{h_k\}$ is uniformly bounded, the sequence $\{h_j\}$ is bounded in $L^\infty([0, T]; \mathbb{R}^p)$ and consequently bounded in $L^2([0, T]; \mathbb{R}^p)$. Passing to a subsequence, there exists a limit process $\bar{H} \in L^2([0, T]; \mathbb{R}^p)$ such that $h_j \rightharpoonup \bar{H}$ weakly in $L^2([0, T]; \mathbb{R}^p)$.

Taking the limit as $j \to \infty$ in (31) over $[s_j, s_j + t]$, the weak convergence of h_j together with the uniform convergence $Z(s_j + \cdot) - Z(s_j) \to 0$ yields the following dynamics,

$$\bar{X}(t) - \bar{X}(0) = -\int_0^t \bar{H}(r)\, dr, \qquad 0 \leq t \leq T. \tag{32}$$

Applying the Lebesgue differentiation theorem implies $\dot{\bar{X}}(t) = -\bar{H}(t)$ for almost every $t \in [0, T]$.

It remains to verify that $\bar{H}(t) \in D(\bar{X}(t))$ almost everywhere. The set valued map D has nonempty compact convex values and is upper semicontinuous. Since $\operatorname{dist}((x_k, h_k), \operatorname{gph} D) \to 0$, for each k we may choose $\tilde{x}_k$ and $\tilde{h}_k \in D(\tilde{x}_k)$ such that $\|x_k - \tilde{x}_k\|_2 + \|h_k - \tilde{h}_k\|_2 \to 0$. Moreover, convergence of the martingale series and boundedness of $\{h_k\}$ imply $\|x_{k+1} - x_k\|_2 \leq \gamma_k \|h_k\|_2 + \|\gamma_k M_{k+1}\|_2 \to 0$. The graph-distance convergence at the discrete iterates therefore extends to the shifted interpolations as follows,

$$\sup_{0\leq t\leq T} \operatorname{dist}((X_j(t), h_j(t)), \operatorname{gph} D) \to 0 \qquad \text{for every } T > 0. \tag{33}$$

Let $\sigma_K(q) \triangleq \sup_{v\in K} \langle q, v\rangle$ denote the support function of a compact convex set K. Upper semicontinuity of D guarantees that $x \mapsto \sigma_{D(x)}(q)$ is upper semicontinuous for every $q \in \mathbb{R}^p$. For any measurable set $A \subseteq [0, T]$ and vector $q \in \mathbb{R}^p$, weak convergence $h_j \rightharpoonup \bar{h}$ in L^2 ensures the identity below.

$$\int_A \langle q, \bar{H}(t)\rangle\, dt = \lim_{j\to\infty} \int_A \langle q, h_j(t)\rangle\, dt = \lim_{j\to\infty} \int_A \langle q, \tilde{h}_j(t)\rangle\, dt.$$

Because $\tilde{h}_j(t) \in D(\tilde{X}_j(t))$ almost everywhere, we have $\langle q, \tilde{h}_j(t)\rangle \leq \sigma_{D(\tilde{X}_j(t))}(q)$. Taking the limit superior and invoking Fatou's lemma yields the following inequality,

$$\int_A \langle q, \bar{H}(t)\rangle\, dt \leq \int_A \limsup_{j\to\infty} \sigma_{D(\tilde{X}_j(t))}(q)\, dt \leq \int_A \sigma_{D(\bar{X}(t))}(q)\, dt.$$

Since this integral inequality holds for every measurable set $A \subseteq [0, T]$, it follows that $\langle q, \bar{H}(t)\rangle \leq \sigma_{D(\bar{X}(t))}(q)$ almost everywhere for each fixed q. Evaluating this inequality over a countable dense subset of $\mathbb{R}^p$ establishes that $\bar{H}(t) \in D(\bar{X}(t))$ for almost every $t \in [0, T]$. Combining this inclusion with (32) demonstrates that $\dot{\bar{X}}(t) \in -D(\bar{X}(t))$ almost everywhere. Thus every time-shift limit of X is a solution of $\dot{\bar{X}}(t) \in -D(\bar{X}(t))$. The relative compactness of the time-shift family therefore establishes that X is an asymptotic pseudo-trajectory of $\dot{x} \in -D(x)$. Its omega-limit set $L := \bigcap_{\tau\geq 0} \overline{\{X(t) : t \geq \tau\}}$ is consequently nonempty, compact, invariant, and internally chain transitive by the standard limit-set theorem for bounded asymptotic pseudo-trajectories.

We now apply the conservative field property to analyze the energy descent along trajectories. Let $x(\cdot)$ be an arbitrary solution to the differential inclusion $\dot{x}(t) \in -D(x(t))$. By definition of a conservative field, D satisfies the chain rule along $x(\cdot)$, which yields the results below.

$$\frac{d}{dt} f(x(t)) = \langle -\dot{x}(t), \dot{x}(t)\rangle = -\|\dot{x}(t)\|_2^2 \leq 0 \qquad \text{for almost every } t. \tag{34}$$

In particular, f is a strict Lyapunov function for $\dot{\bar{X}}(t) \in -D(\bar{X}(t))$ relative to the critical set $S = \{x : 0 \in D(x)\}$. Indeed, for any $x \notin S$, compactness of $D(x)$ ensures $\operatorname{dist}(0, D(x)) > 0$. By

local boundedness and upper semicontinuity (closed graph property) of D, there exist a neighborhood U of x and a constant $\delta > 0$ such that $\operatorname{dist}(0, D(y)) \geq \delta > 0$ for all $y \in U$. Consequently, any solution segment passing through U obeys $\|\dot{x}(t)\|_2 \geq \delta$, guaranteeing a strict decrease in f as $\frac{d}{dt} f(x(t)) \leq -\delta^2 < 0$ with $x(t) \in U$.

Because L is internally chain transitive and $f(S)$ has empty interior, the Lyapunov principle for internally chain-transitive sets applies. Specifically, no point in L can lie outside S. Otherwise, the strict energy dissipation across regular energy levels would contradict the chain transitivity of L. Thus, we obtain $L \subseteq S$, and f must be constant on L.

Since every cluster point of $\{x_k\}$ is contained in L, the inclusion $L \subseteq S$ directly implies that $\operatorname{clust}(x_k) \subseteq S$ almost surely. Let $c \in \mathbb{R}$ denote the constant value of f on L. Because $\{x_k\}$ is bounded and $\operatorname{dist}(x_k, L) \to 0$ almost surely as $k \to \infty$, the continuity of f guarantees that $f(x_k) \to c$ almost surely.

Finally, suppose that S is finite. On the one hand, $L \subseteq S$ so that L is a finite set. On the other hand, because $L = \bigcap_{T \geq 0} \overline{X([T, \infty))}$ is a nested intersection of non-empty compact connected sets, L is connected. Since any connected subset of a finite set is a singleton, there exists a unique $x^\star \in S$ such that $L = \{x^\star\}$. Boundedness of $\{x_k\}$ along with $\operatorname{clust}(x_k) \subseteq \{x^\star\}$ then confirms that $x_k \to x^\star$ almost surely. □

B.2 Graph Consistency

Proposition 11 (Graph consistency of the OA mean drift)**.** *Suppose Assumptions 2 and 3 hold. Suppose further that, for $\mathcal{D}$-almost every z, $\ell(\cdot, z)$ is continuously differentiable. The sample-wise stochastic directions in (15) can be written as follows,*

$$d_\alpha(\theta, z) := J_g(\theta, z)^\mathsf{T} J_{\mathrm{OA}}(g_\theta(z); \alpha)^\mathsf{T} \nabla_y \ell(p_{\mathrm{OA}}(g_\theta(z); \alpha), z)\,, \tag{35}$$

Assume that the second moments of these directions are uniformly bounded over all admissible measurable conservative Jacobian selections, such that

$$\sup_{\theta \in \mathcal{K}_\theta,\ \alpha \in [0,1]} \mathbb{E}_{z \sim \mathcal{D}} \left[\|d_\alpha(\theta, z)\|_2^2 \right] < \infty. \tag{36}$$

Then $D_{\mathrm{H}} F_{\mathrm{H}}$ is a locally bounded, closed-graph conservative field with nonempty compact convex values. Furthermore, whenever $\theta_j \to \theta$, $\alpha_j \to 0$, and $H_{\alpha_j}(\theta_j) \to h$, the limiting drift satisfies $h \in D_{\mathrm{H}} F_{\mathrm{H}}(\theta)$.

We next prove the regularity of $D_{\mathrm{H}} F_{\mathrm{H}}$. The fields $\mathcal{J}_g$ and $\mathcal{J}_{\mathrm{H}}$ are nonempty and compact valued, so $\Gamma_{\mathrm{H}}(\theta, z)$ is nonempty and compact.

Proof of Proposition 11. For $\theta \in \mathcal{K}_\theta$, $\alpha \in [0, 1]$, and $\mathcal{D}$-almost every z, let $\Gamma_\alpha(\theta, z)$ denote the set of all sample-wise directions of the form in (35) generated by admissible conservative Jacobian selections. When $\alpha = 0$, this set reduces to below.

$$\Gamma_{\mathrm{H}}(\theta, z) := \left\{ J_g^\mathsf{T} J_{\mathrm{H}}^\mathsf{T} \nabla_y \ell(\Pi_{\mathcal{C}}(g_\theta(z)), z) : J_g \in \mathcal{J}_{g_\theta}(z), J_{\mathrm{H}} \in \mathcal{J}_{\Pi_{\mathcal{C}}}(g_\theta(z)) \right\}. \tag{37}$$

The population field $D_{\mathrm{H}} F_{\mathrm{H}}$ of the orthogonal-projection objective can be explicitly expressed as the following closed convex hull,

$$D_{\mathrm{H}} F_{\mathrm{H}}(\theta) := \overline{\mathrm{co}} \left\{ \mathbb{E}_{z \sim \mathcal{D}}[\xi(z)] : \xi(z) \in \Gamma_{\mathrm{H}}(\theta, z) \text{ a.e. } \xi \text{ measurable} \right\}. \tag{38}$$

We first establish the pointwise outer consistency of the instance-wise fields. Fix a context z and consider arbitrary sequences which satisfy $\theta_j \to \theta$ and $\alpha_j \to 0$. Suppose that $d_j \in \Gamma_{\alpha_j}(\theta_j, z)$ and $d_j \to d$. We will prove that the limit $d \in \Gamma_{\mathrm{H}}(\theta, z)$.

Set $u_j := g_{\theta_j}(z)$ and $u := g_\theta(z)$. By the path differentiability in Assumption 3, g_θ is continuous in θ, which implies $u_j \to u$. Recall that $p_{\mathrm{OA}}(u_j; \alpha_j) = \Pi_{\mathcal{C}}(u_j) + \alpha_j \big(p_{\mathrm{SR}}(u_j) - \Pi_{\mathcal{C}}(u_j)\big)$. By Theorem 4, p_{SR} is locally Lipschitz continuous. Since $\Pi_{\mathcal{C}}$ is also continuous and $u_j \to u$, we have $p_{\mathrm{SR}}(u_j) \to p_{\mathrm{SR}}(u)$ and $\Pi_{\mathcal{C}}(u_j) \to \Pi_{\mathcal{C}}(u)$. Therefore, the correction $p_{\mathrm{SR}}(u_j) - \Pi_{\mathcal{C}}(u_j)$ is bounded, and since $\alpha_j \to 0$, the perturbation $\alpha_j \big(p_{\mathrm{SR}}(u_j) - \Pi_{\mathcal{C}}(u_j)\big)$ vanishes. Consequently, we obtain the forward limit behavior below.

$$p_{\mathrm{OA}}(u_j; \alpha_j) \longrightarrow \Pi_{\mathcal{C}}(u) \text{ as } j \to \infty. \tag{39}$$

Let $J_{g,j} \in \mathcal{J}_g(\theta_j, z)$ denote the model Jacobian used in d_j. The regularity bound in Assumption 3 guarantees that the sequence $\{J_{g,j}\}$ is bounded. Passing to a subsequence if necessary, we obtain $J_{g,j} \to J_g$. Since $\mathcal{J}_g$ has a closed graph, the limiting Jacobian satisfies $J_g \in \mathcal{J}_g(\theta, z)$.

Similarly, by (11), the OA Jacobian decomposes as $J_{\mathrm{OA},j} = (1-\alpha_j)J_{\mathrm{H},j} + \alpha_j J_{\mathrm{SR},j}$, where $J_{\mathrm{H},j} \in \mathcal{J}_{\mathrm{H}}(u_j)$ and $J_{\mathrm{SR},j} \in \mathcal{J}_{\mathrm{SR}}(u_j)$. Lemma 9 gives $\|J_{\mathrm{H},j}\|_2 \le 1$. By Theorem 4, p_{SR} is locally Lipschitz, so there exist a neighborhood U of u and a constant $L_U < \infty$ such that p_{SR} is L_U-Lipschitz on U. Hence, at every differentiable point $x \in U$, the bound $\|J_{p_{\mathrm{SR}}}(x)\|_2 \le L_U$ holds. For the conservative Jacobian $\mathcal{J}_{\mathrm{SR}}$ compatible with the classical Jacobians, the same bound holds below.

$$\sup_{x \in U} \sup_{J \in \mathcal{J}_{\mathrm{SR}}(x)} \|J\|_2 \le L_U.$$

Since $u_j \to u$, we have $u_j \in U$ for all sufficiently large j, which yields $\|J_{\mathrm{SR},j}\|_2 \le L_U$. Combined with $\alpha_j \to 0$, this implies $\alpha_j J_{\mathrm{SR},j} \to 0$. Passing to a subsequence if necessary, we have $J_{\mathrm{H},j} \to J_{\mathrm{H}}$, and the closed graph property of $\mathcal{J}_{\mathrm{H}}$ ensures $J_{\mathrm{H}} \in \mathcal{J}_{\mathrm{H}}(u)$. We thus establish the backward limiting behavior below.

$$J_{\mathrm{OA},j} \to J_{\mathrm{H}} \text{ as } j \to \infty. \tag{40}$$

Because $\ell(\cdot, z)$ is continuously differentiable, (39) guarantees $\nabla_y \ell(p_{\mathrm{OA}}(u_j; \alpha_j), z) \longrightarrow \nabla_y \ell(\Pi_{\mathcal{C}}(u), z)$. Combining this limit with $J_g \in \mathcal{J}_g(\theta, z)$ and (40) in (35) confirms that $d \in \Gamma_{\mathrm{H}}(\theta, z)$, establishing the following inclusion result,

$$\mathrm{Limsup}_{j \to \infty} \Gamma_{\alpha_j}(\theta_j, z) \subseteq \Gamma_{\mathrm{H}}(\theta, z) \qquad \text{for } \mathcal{D}\text{-almost every } z. \tag{41}$$

We next establish the regularity of $D_{\mathrm{H}}F_{\mathrm{H}}$. The fields $\mathcal{J}_g$ and $\mathcal{J}_{\mathrm{H}}$ are nonempty and compact valued, so $\Gamma_{\mathrm{H}}(\theta, z)$ is also nonempty and compact. Thus, the measurable Jacobian selection of $\mathcal{J}_g$ and $\mathcal{J}_{\mathrm{H}}$ induce measurable selections of $\Gamma_{\mathrm{H}}(\theta, \cdot)$. Applying Jensen's inequality alongside the second-moment bound in (36, every selection $\xi(z)$ in $\Gamma_{\mathrm{H}}(\theta, z)$ satisfies the following bound,

$$\|\mathbb{E}[\xi(z)]\|_2 \le \sqrt{\sup_{\theta \in \mathcal{K}_\theta,\ \alpha \in [0,1]} \mathbb{E}_{z \sim \mathcal{D}} \left[\|d_\alpha(\theta, z)\|_2^2 \right]}.$$

It then follows from (38) that $D_{\mathrm{H}}F_{\mathrm{H}}(\theta)$ is nonempty, bounded, closed, and convex. Since the parameter space is finite-dimensional, its values are compact. Furthermore, because this bound holds uniformly over $\theta \in \mathcal{K}_\theta$, $D_{\mathrm{H}}F_{\mathrm{H}}$ is locally bounded on the parameter domain.

Then we verify that the population field $D_{\mathrm{H}}F_{\mathrm{H}}$ has a closed graph. Take convergent sequences $\theta_j \to \theta$ and $v_j \to v$ such that $v_j \in D_{\mathrm{H}}F_{\mathrm{H}}(\theta_j)$. For $q \in \mathbb{R}^{n_\theta}$, let $\sigma_S(q) := \sup_{s \in S} \langle q, s \rangle$ denote the support function of a set S. Since taking the closed convex hull preserves the supremum of a continuous linear functional, we can obtain the following inequality,

$$\langle q, v_j \rangle \le \sup_{\xi(z) \in \Gamma_{\mathrm{H}}(\theta_j, z)} \langle q, \mathbb{E}[\xi(z)] \rangle = \sup_{\xi} \mathbb{E}[\langle q, \xi(z) \rangle] \le \mathbb{E}\left[\sigma_{\Gamma_{\mathrm{H}}(\theta_j, z)}(q)\right]. \tag{42}$$

Specializing α_j in (41) to 0 gives the outer limit conlusion $\mathrm{Limsup}_{j \to \infty} \Gamma_{\mathrm{H}}(\theta_j, z) \subseteq \Gamma_{\mathrm{H}}(\theta, z)$ for $\mathcal{D}$-almost every z. This implies the upper bound on the support functions below.

$$\limsup_{j \to \infty} \sigma_{\Gamma_{\mathrm{H}}(\theta_j, z)}(q) \le \sigma_{\Gamma_{\mathrm{H}}(\theta, z)}(q) \quad \text{a.e.} \tag{43}$$

Because $\Gamma_{\mathrm{H}}(\theta_j, z)$ is compact valued, its support function is attained by an admissible measurable selection. The uniform second-moment bound on the admissible sample-wise directions implies uniform integrability of $\sigma_{\Gamma_{\mathrm{H}}(\theta_j, z)}(q)$, which yields $\limsup_{j \to \infty} \mathbb{E}\left[\sigma_{\Gamma_{\mathrm{H}}(\theta_j, z)}(q)\right] \le \mathbb{E}\left[\sigma_{\Gamma_{\mathrm{H}}(\theta, z)}(q)\right]$. Taking $j \to \infty$ in (42) then gives $\langle q, v \rangle \le \mathbb{E}\left[\sigma_{\Gamma_{\mathrm{H}}(\theta, z)}(q)\right] = \sigma_{D_{\mathrm{H}}F_{\mathrm{H}}(\theta)}(q)$. Since this inequality holds for every q, the separating-hyperplane theorem ensures $v \in D_{\mathrm{H}}F_{\mathrm{H}}(\theta)$, confirming that $D_{\mathrm{H}}F_{\mathrm{H}}$ has a closed graph.

It remains to verify the conservative property. Define the instance-wise orthogonal-projection objective as $f_{\mathrm{H},z}(\theta) := \ell(\Pi_{\mathcal{C}}(g_\theta(z)), z)$. By the conservative chain rule together with Lemma 9, the set-valued map $\Gamma_{\mathrm{H}}(\cdot, z)$ is a conservative field for $f_{\mathrm{H},z}$ for $\mathcal{D}$-almost every z. Let $\vartheta : [0, T] \to \mathcal{K}_\theta$ be any absolutely continuous curve. For every admissible measurable selection $\xi(t, z) \in \Gamma_{\mathrm{H}}(\vartheta(t), z)$, the following identity holds for almost every t and $\mathcal{D}$-almost every z,

$$\frac{d}{dt} f_{\mathrm{H},z}(\vartheta(t)) = \left\langle \xi(t, z), \dot{\vartheta}(t) \right\rangle.$$

The uniform second-moment bound in (36) provides the required integrability to interchange differentiation and expectation under the integral sign. Thus, for almost every t, we obtain

$$\frac{d}{dt}F_{\mathrm{H}}(\vartheta(t)) = \mathbb{E}_{z\sim\mathcal{D}}\left[\frac{d}{dt}f_{\mathrm{H},z}(\vartheta(t))\right] = \left\langle \mathbb{E}_{z\sim\mathcal{D}}[\xi(t,z)], \dot{\vartheta}(t)\right\rangle.$$

According to (38), $D_{\mathrm{H}}F_{\mathrm{H}}(\vartheta(t))$ is defined as the closed convex hull of these expected directions. Because the inner product with $\dot{\vartheta}(t)$ is linear and continuous, the same equality extends to all $v \in D_{\mathrm{H}}F_{\mathrm{H}}(\vartheta(t))$, yielding the results below.

$$\frac{d}{dt}F_{\mathrm{H}}(\vartheta(t)) = \left\langle v, \dot{\vartheta}(t)\right\rangle.$$

Therefore, $D_{\mathrm{H}}F_{\mathrm{H}}$ is a conservative field for F_{H}. □

B.3 Convergence proof

Proof of Theorem 5. Let $\mathcal{F}_k$ denote the filtration representing the information available before sampling $\mathcal{B}_k$. Conditional independence of the minibatch samples combined with (15) implies that the mean direction satisfies $\mathbb{E}[\widehat{G}_k \mid \mathcal{F}_k] = H_{\alpha_k}(\theta_k)$. Define the noise sequence as $M_{k+1} := \widehat{G}_k - H_{\alpha_k}(\theta_k)$, which fulfills $\mathbb{E}[M_{k+1} \mid \mathcal{F}_k] = 0$. Consequently, $\{M_{k+1}\}$ forms a martingale difference sequence. Let $C_d := \sup_{\theta\in\mathcal{K}_\theta,\ \alpha\in[0,1]} \mathbb{E}_{z\sim\mathcal{D}}\left[\|d_\alpha(\theta,z)\|_2^2\right] < \infty$ denote the uniform second-moment bound from Proposition 11. Since $\widehat{G}_k$ is the average of the sample-wise directions in the mini-batch, Jensen's inequality gives the following inequality,

$$\mathbb{E}[\|\widehat{G}_k\|_2^2 \mid \mathcal{F}_k] \le C_d.$$

This yields the conditional variance bound $\mathbb{E}[\|M_{k+1}\|_2^2 \mid \mathcal{F}_k] \le C_d$. Again by Jensen's inequality, we have $\|H_{\alpha_k}(\theta_k)\|_2^2 \le C_d$. Thus, the mean drift sequence $\{H_{\alpha_k}(\theta_k)\}$ is uniformly bounded.

We next establish the graph consistency of the mean drift. Consider any subsequence $\{k_j\}$ such that $\theta_{k_j} \to \theta$ and $H_{\alpha_{k_j}}(\theta_{k_j}) \to h$. By the definition of the OA correction weight, satisfies $0 \le \alpha_k \le \eta_k$ with $\eta_k \to 0$, which forces $\alpha_{k_j} \to 0$. Proposition 11 then ensures $h \in D_{\mathrm{H}}F_{\mathrm{H}}(\theta)$, confirming that $\{H_{\alpha_k}(\theta_k)\}$ is graph consistent with $D_{\mathrm{H}}F_{\mathrm{H}}$.

The OA parameter update can now be written as follows,

$$\theta_{k+1} = \theta_k - \gamma_k\big(H_{\alpha_k}(\theta_k) + M_{k+1}\big).$$

This is precisely of the form (28) with $x_k = \theta_k, h_k = H_{\alpha_k}(\theta_k), D = D_{\mathrm{H}}F_{\mathrm{H}}, f = F_{\mathrm{H}}$. Assumptions 1 and 2 verify the required step-size and bounded-iterate conditions in Lemma 10. Furthermore, Proposition 11 establishes that $D_{\mathrm{H}}F_{\mathrm{H}}$ is a locally bounded and closed-graph conservative field with nonempty compact convex values, while validating the required graph consistency of the mean drift. Finally, the assumed empty-interior property of $F_{\mathrm{H}}(\mathcal{S}_{\mathrm{H}})$ is exactly the weak-Sard condition in Lemma 10. All hypotheses of Lemma 10 are fulfilled, guaranteeing that $F_{\mathrm{H}}(\theta_k)$ converges almost surely and the conclusion relationship below.

$$\mathrm{clust}(\theta_k) \subseteq \mathcal{S}_{\mathrm{H}} \qquad \text{almost surely.}$$

If $\mathcal{S}_{\mathrm{H}}$ is finite, θ_k converges almost surely to a single point of $\mathcal{S}_{\mathrm{H}}$. □

C Proofs of KKT Consistency

C.1 Local normal visibility and output-space consistency

Let $\mathcal{L} := \mathrm{span}(\mathcal{A} - \mathcal{A})$ denote the linear subspace parallel to the affine hull of $\mathcal{A}$. For a face F, we define its relative-normal subspace by $\mathcal{N}_F := \mathcal{L} \cap T_F^\perp$, with $P_{\mathcal{N}_F}$ representing the associated orthogonal projector. The inclusion $N_{\mathcal{C}}(q_{\mathrm{R}}(\bar{u})) \subseteq T_F^\perp$ ensures that all active radial facet normals, when restricted to $\mathcal{L}$, lie within $\mathcal{N}_F$. Leveraging this geometric configuration, the following proposition establishes a common escape direction across all local radial branches, while the residual of the orthogonal projection remains bounded away from zero.

Proposition 12 (Normal visibility, residual separation, and escape)**.** *Let $(\bar{u}, \bar{y})$ be a projection-induced spurious stationary pair with active face F. Suppose that $\ell(\cdot, z)$ is continuously differentiable and $N_{\mathcal{C}}(q_{\mathrm{R}}(\bar{u})) \subseteq T_F^{\perp}$. Then the normal direction can be retained by the SR component of OA projection as follows,*

$$P_{\mathcal{N}_F} J_{\mathrm{SR}}(\bar{u})^{\mathsf{T}} \nabla_y \ell(\bar{y}, z) \neq 0. \tag{44}$$

Moreover, there exist a neighborhood $\mathcal{U}$ of $\bar{u}$, a unit vector $v^{\mathrm{esc}} \in \mathcal{N}_F$, and constants $\bar{\alpha} \in (0, 1]$, $\underline{r} > 0$, and $c_{\mathrm{esc}} > 0$ such that, for every $u \in \mathcal{U}$ and $\alpha \in (0, \bar{\alpha})$,

$$\operatorname{dist}(-\nabla_y \ell(\Pi_{\mathcal{C}}(u), z), N_{\mathcal{C}}(\Pi_{\mathcal{C}}(u))) \geq \underline{r}, \tag{45}$$

$$\left\langle -J_{\mathrm{OA}}(u; \alpha)^{\mathsf{T}} \nabla_y \ell(p_{\mathrm{OA}}(u; \alpha), z), v^{\mathrm{esc}} \right\rangle \geq c_{\mathrm{esc}} \alpha. \tag{46}$$

Proof of Proposition 12. By Definition 1, the spurious stationary pair $(\bar{u}, \bar{y})$ satisfies $\bar{y} \in \operatorname{relint}(F)$ and $\bar{u} - \bar{y} \in \operatorname{ri} N_{\mathcal{C}}(F)$. Consequently, there exists a relative neighborhood $\mathcal{U}_0$ of $\bar{u}$ on which $\Pi_{\mathcal{C}}(u) \in \operatorname{relint}(F)$, the Jacobian reduces to $J_{\mathrm{H}}(u) = P_{T_F}$ and the normal cone remains invariant i.e. $N_{\mathcal{C}}(\Pi_{\mathcal{C}}(u)) = N_{\mathcal{C}}(F)$. To examine the directional components relative to the affine hull, set $v := \bar{u} - y_0$, $v_N := P_{\mathcal{N}_F} v$ and $g := P_{\mathcal{L}} \nabla_y \ell(\bar{y}, z)$. The tangential stationarity condition $P_{T_F} \nabla_y \ell(\bar{y}, z) = 0$ implies $g \in \mathcal{L} \cap T_F^{\perp} = \mathcal{N}_F$. Moreover, $g \neq 0$, since otherwise we would have $-\nabla_y \ell(\bar{y}, z) \in \mathcal{L}^{\perp} \subseteq N_{\mathcal{C}}(\bar{y})$, which contradicts the assumed KKT violation at the spurious stationary point.

We write the facet representation of $\mathcal{C}$ relative to $\mathcal{A}$ as follows,

$$\mathcal{C} = \{y \in \mathcal{A} : \langle a_j, y - y_0 \rangle \leq \delta_j, \quad j = 1, \ldots, m\},$$

where $a_j \in \mathcal{L} \setminus \{0\}$ and $\delta_j > 0$ because $y_0 \in \operatorname{ri}(\mathcal{C})$. Spurious stationarity implies that F is a proper face and $\bar{u} \notin \mathcal{C}$. Consequently, the radial projection point can be represented as $\bar{q} := q_{\mathrm{R}}(\bar{u}) = y_0 + \bar{\lambda} v$ with scale factor $\bar{\lambda} := \lambda^{\star}(\bar{u}) \in (0, 1)$. Let $\mathcal{I}_{\mathrm{R}} := \{j : \langle a_j, \bar{q} - y_0 \rangle = \delta_j\}$ index the facets active at $\bar{q}$. For every $j \in \mathcal{I}_{\mathrm{R}}$, the vector satisfies $a_j \in N_{\mathcal{C}}(\bar{q}) \cap \mathcal{L}$. The assumed inclusion $N_{\mathcal{C}}(\bar{q}) \subseteq T_F^{\perp}$ therefore gives $a_j \in \mathcal{L} \cap T_F^{\perp} = \mathcal{N}_F$.

For each $j \in \mathcal{I}_{\mathrm{R}}$, define the local radial expression as $q_{\mathrm{R},j}(u) := y_0 + \lambda_j(u)(u - y_0)$ with $\lambda_j(u) := \frac{\delta_j}{\langle a_j, u - y_0 \rangle}$. This expression agrees with $q_{\mathrm{R}}(u)$ wherever facet j determines the radial endpoint. Specifically, at $\bar{u}$, we have $\lambda_j(\bar{u}) = \bar{\lambda}$ and $\langle a_j, v \rangle = \delta_j / \bar{\lambda} > 0$. Let $J_{\mathrm{SR},j}$ denote the Jacobian of the map $u \mapsto y_0 + \rho(\|u - y_0\|_2^2)\big(q_{\mathrm{R},j}(u) - y_0\big)$. Differentiation along $\mathcal{L}$ yields the following result,

$$J_{\mathrm{SR},j}(\bar{u}) = \bar{\lambda} \left[\rho(\|v\|_2^2) I_{\mathcal{L}} + v \left(2\rho'(\|v\|_2^2) v - \rho(\|v\|_2^2) \frac{a_j}{\langle a_j, v \rangle} \right)^{\mathsf{T}} \right]. \tag{47}$$

To identify a common escape direction across all active local branches, we set $s := \langle v_N, g \rangle = \langle v, g \rangle$ and examine the projected gradient response $w_j := P_{\mathcal{N}_F} J_{\mathrm{SR},j}(\bar{u})^{\mathsf{T}} g$. Since $a_j \in \mathcal{N}_F$ and $g \in \mathcal{N}_F$, explicit computation gives the following result,

$$w_j = P_{\mathcal{N}_F} J_{\mathrm{SR},j}(\bar{u})^{\mathsf{T}} g = \bar{\lambda} \left[\rho g + \left(2\rho' v_N - \rho \frac{a_j}{\langle a_j, v \rangle} \right) s \right]. \tag{48}$$

Taking the inner product of w_j with v_N and noting that $\langle a_j, v_N \rangle = \langle a_j, v \rangle$, we arrive at the pivotal identity $\langle w_j, v_N \rangle = 2\bar{\lambda}\rho' \|v_N\|_2^2 s$ for $j \in \mathcal{I}_{\mathrm{R}}$. The right-hand side $2\bar{\lambda}\rho' \|v_N\|_2^2 s$ is independent of j, allowing us to define a common escape unit vector $v^{\mathrm{esc}} \in \mathcal{N}_F$ via the following case analysis,

- If $s \neq 0$, set $v^{\mathrm{esc}} := -\operatorname{sgn}(s) \frac{v_N}{\|v_N\|_2}$. Then $\langle w_j, v_N \rangle = 2\bar{\lambda}\rho' \|v_N\|_2^2 s$ yields

$$-\langle w_j, v^{\mathrm{esc}} \rangle = 2\bar{\lambda}\rho'(\|v\|_2^2) \|v_N\|_2 |s| > 0.$$

- If $s = 0$, (48) simplifies to $w_j = \bar{\lambda}\rho(\|v\|_2^2) g$. Setting $v^{\mathrm{esc}} := -\frac{g}{\|g\|_2}$ yields

$$-\langle w_j, v^{\mathrm{esc}} \rangle = \bar{\lambda}\rho(\|v\|_2^2) \|g\|_2 > 0.$$

Thus, in either case, there exist a unit vector $v^{\mathrm{esc}} \in \mathcal{N}_F$ and a constant $c_0 > 0$ such that

$$-\left\langle J_{\mathrm{SR},j}(\bar{u})^{\mathsf{T}} \nabla_y \ell(\bar{y}, z), v^{\mathrm{esc}} \right\rangle \geq c_0 \qquad \text{for every } j \in \mathcal{I}_{\mathrm{R}}. \tag{49}$$

The ambient gradient can replace g because each $J_{\mathrm{SR},j}$ has range in $\mathcal{L}$. Every facet outside $\mathcal{I}_{\mathrm{R}}$ has strict slack at $\bar{q}$. Continuity of q_{R} therefore ensures that any facet active at $q_{\mathrm{R}}(u)$ for u sufficiently close to $\bar{u}$ belongs to $\mathcal{I}_{\mathrm{R}}$. Furthermore, each $J_{\mathrm{SR},j}$ is continuous near $\bar{u}$, since its denominator remains positive and ρ' is continuous. By Clarke's generalized Jacobian definition, $\mathcal{J}_{\mathrm{SR}}(\bar{u}) \subseteq \operatorname{conv}\{J_{\mathrm{SR},j}(\bar{u}) : j \in \mathcal{I}_{\mathrm{R}}\}$. Taking convex combinations in (49) confirms that $P_{\mathcal{N}_F} J_{\mathrm{SR}}(\bar{u})^{\mathsf{T}} \nabla_y \ell(\bar{y}, z) \neq 0$, establishing (44).

The same argument provides a uniform escape bound near $\bar{u}$. By joint continuity of $J_{\mathrm{SR},j}(\cdot)$, $\nabla_y \ell(\cdot, z)$, and the OA state map $(u, \alpha) \mapsto p_{\mathrm{OA}}(u; \alpha)$, we can shrink $\mathcal{U}_0$ to a neighborhood $\mathcal{U}$, and select $\bar{\alpha} \in (0, 1]$ and $c_{\mathrm{esc}} > 0$ such that $-\left\langle J_{\mathrm{SR},j}(u)^{\mathsf{T}} \nabla_y \ell(p_{\mathrm{OA}}(u; \alpha), z), v^{\mathrm{esc}} \right\rangle \geq c_{\mathrm{esc}}$ for every $u \in \mathcal{U}$, $\alpha \in [0, \bar{\alpha}]$, and $j \in \mathcal{I}_{\mathrm{R}}$. Since any generalized Jacobian element $J \in \mathcal{J}_{\mathrm{SR}}(u)$ is a convex combination of such limiting branch Jacobians, the inequality extends directly to J.

Next, recall that $v^{\mathrm{esc}} \in \mathcal{N}_F \subseteq T_F^{\perp}$, which implies $J_{\mathrm{H}}(u) v^{\mathrm{esc}} = P_{T_F} v^{\mathrm{esc}} = 0$ for all $u \in \mathcal{U}$. Evaluating the full OA projection Jacobian $J_{\mathrm{OA}}(u; \alpha) = (1 - \alpha) J_{\mathrm{H}}(u) + \alpha J$, we obtain

$$\left\langle -J_{\mathrm{OA}}^{\mathsf{T}} \nabla_y \ell(p_{\mathrm{OA}}(u; \alpha), z), v^{\mathrm{esc}} \right\rangle = -\alpha \left\langle J^{\mathsf{T}} \nabla_y \ell(p_{\mathrm{OA}}(u; \alpha), z), v^{\mathrm{esc}} \right\rangle \geq c_{\mathrm{esc}} \alpha,$$

which establishes the escape margin inequality (46).

Finally, we prove the residual separation inequality (45). Define the projection residual function $r_{\mathrm{H}}(u) \triangleq \operatorname{dist}(-\nabla_y \ell(\Pi_{\mathcal{C}}(u), z), N_{\mathcal{C}}(\Pi_{\mathcal{C}}(u)))$. At $\bar{u}$, we have $r_{\mathrm{H}}(\bar{u}) = \operatorname{dist}(-\nabla_y \ell(\bar{y}, z), N_{\mathcal{C}}(F)) > 0$ due to the KKT violation. This function is continuous, so shrinking $\mathcal{U}$ if necessary guarantees a lower bound $\underline{r} > 0$ such that $r_{\mathrm{H}}(u) \geq \underline{r}$ for every $u \in \mathcal{U}$. This completes the proof of (45). □

Proof of Theorem 6. Suppose, for the sake of contradiction, that $\bar{y}$ violates the KKT condition. By hypothesis, $(\bar{u}, \bar{y})$ is a projection-induced spurious stationary pair with active face F satisfying $N_{\mathcal{C}}(q_{\mathrm{R}}(\bar{u})) \subseteq T_F^{\perp}$. Proposition 12 then provides a neighborhood $\mathcal{U}$ of $\bar{u}$, a unit escape direction v^{esc}, and positive constants $\bar{\alpha}$, $\underline{r}$, and c_{esc} satisfying the residual separation (45) and escape margin (46).

Since $u_k \to \bar{u}$ and $0 \leq \alpha_k \leq \eta_k \to 0$, there exists an integer $K \geq 1$ such that $u_k \in \mathcal{U}$ and $\alpha_k \in [0, \bar{\alpha}]$ for all $k \geq K$. By the separation bound (45), $r_k(z) \geq \underline{r}$ for all $k \geq K$, which by monotonicity of ψ implies $\alpha_k = \eta_k \psi(r_{k-1}(z)) \geq \psi(\underline{r}) \eta_k$ for all $k \geq K + 1$. Taking the inner product of (17) with v^{esc} and applying (46) yields the inequality below.

$$\langle u_{k+1} - u_k, v^{\mathrm{esc}} \rangle \geq c_{\mathrm{esc}} \gamma_k \alpha_k \geq c_{\mathrm{esc}} \psi(\underline{r}) \gamma_k \eta_k.$$

Summing this inequality from $k = K + 1$ to $m - 1$ produces

$$\langle u_m - u_K, v^{\mathrm{esc}} \rangle \geq c_{\mathrm{esc}} \psi(\underline{r}) \sum_{k=K}^{m-1} \gamma_k \eta_k.$$

Taking $m \to \infty$, the non-summability condition $\sum_k \gamma_k \eta_k = \infty$ causes the right-hand side to diverge to $+\infty$, whereas the left-hand side converges to the finite value $\langle \bar{u} - u_{K+1}, v^{\mathrm{esc}} \rangle$. This contradiction completes the proof, establishing that $\bar{y}$ satisfies the KKT consistency limit (19). □

C.2 Linear probing and empirical KKT consistency

In linear probing, the full-row-rank feature condition allows each output-space escape direction to be realized by a parameter perturbation that leaves the other training outputs unchanged. Combining this exact lifting with a weighted recurrence analysis under stochastic mini-batching ensures that any sample-level KKT violation induces an unbounded directional parameter drift, establishing empirical KKT consistency. Throughout this section, mini-batches $\mathcal{B}_k$ are sampled independently across iterations with bounded batch sizes $1 \leq |\mathcal{B}_k| \leq B_{\max} < \infty$ and uniform lower bounds on inclusion probabilities $\mathbb{P}(z_i \in \mathcal{B}_k) \geq p_{\min} > 0$ for all i and k. Let $U(\theta) := [g_\theta(z_1)^\top, \ldots, g_\theta(z_N)^\top]^\top \in \mathbb{R}^{Nd}$ denote the stacked output map across all training samples.

Lemma 13 (Geometry of the linear-probing output map)**.** *For the linear probing predictor in (20), the output Jacobian $J := J_U(\theta)$ is constant. If $\widetilde{H}\widetilde{H}^{\mathsf{T}} \succ 0$, then J has full row rank with*

$$J J^{\mathsf{T}} = (\widetilde{H}\widetilde{H}^{\mathsf{T}}) \otimes I_d \succ 0. \tag{50}$$

Furthermore, for every $V \in \mathbb{R}^{Nd}$, the parameter direction $d(V) := J^\mathsf{T}(JJ^\mathsf{T})^{-1}V$ satisfies $Jd(V) = V$.

Proof. By column-wise vectorization of $\widetilde{W}$, the output on sample z_i expands as $g_\theta(z_i) = \big(\widetilde{h}_0(z_i)^\mathsf{T} \otimes I_d\big)\theta$, which yields $U(\theta) = J\theta$ with constant Jacobian $J = \widetilde{H} \otimes I_d$. The mixed-product property of Kronecker products directly confirms (50). Since JJ^T is positive definite, J has full row rank, and substitution into $d(V) := J^\mathsf{T}(JJ^\mathsf{T})^{-1}V$ gives $Jd(V) = JJ^\mathsf{T}(JJ^\mathsf{T})^{-1}V = V$. □

Lemma 14 (Weighted recurrence under independent mini-batches)**.** *Let $\{E_k\}$ be independent events with $\mathbb{P}(E_k) \geq p > 0$. For any deterministic sequence $a_k \geq 0$ with $\sum_{k=0}^{\infty} a_k = \infty$,*

$$\sum_{k=1}^{\infty} a_k \mathbf{1}_{E_{k-1} \cap E_k} = \infty \qquad \textit{almost surely}. \tag{51}$$

Proof. At least one of the two series $\sum_j a_{2j+1}$ and $\sum_j a_{2j+2}$ diverges. Selecting the corresponding indices k_j, we define $b_j := a_{k_j}$ and $A_j := E_{k_j-1} \cap E_{k_j}$. Because the index pairs $\{k_j - 1, k_j\}$ are pairwise disjoint, the events $\{A_j\}$ are mutually independent with $\mathbb{P}(A_j) \geq p^2$. Defining the partial sum $S_n := \sum_{j=0}^{n} b_j \mathbf{1}_{A_j}$, independence implies

$$\mathbb{E}[e^{-S_n}] = \prod_{j=0}^{n} \left[1 - \mathbb{P}(A_j)(1 - e^{-b_j})\right] \leq \exp\left(-p^2 \sum_{j=0}^{n} (1 - e^{-b_j})\right).$$

Since $1 - e^{-b} \geq (1 - e^{-1}) \min\{b, 1\}$ for $b \geq 0$, the divergence of $\sum_j b_j$ implies $\sum_j (1 - e^{-b_j}) = \infty$. Thus $\mathbb{E}[e^{-S_n}] \to 0$. As S_n increases to $S_\infty \in [0, \infty]$, bounded convergence yields $\mathbb{E}[e^{-S_\infty}] = 0$, establishing $S_\infty = \infty$ almost surely. This divergent subseries is contained in the series in (51), proving the claim. □

Proof of Theorem 7. For each i, let $E_k^i := \{z_i \in \mathcal{B}_k\}$ denote the event that sample i is selected at iteration k. Applying Lemma 14 with $a_k = \gamma_k \eta_k$ establishes that $\sum_{k=1}^{\infty} \gamma_k \eta_k \mathbf{1}_{E_{k-1}^i \cap E_k^i} = \infty$ almost surely. Because the dataset is finite, this holds simultaneously for all i on an single event of probability one. We restrict our attention to a fixed realization within this set on which parameter convergence $\theta_k \to \bar{\theta}$, objective stationarity, and geometric conditions all hold.

We first derive tangential stationarity for each limiting output $\bar{y}_i$. The conditions $\bar{y}_i \in \operatorname{relint}(F_i)$ and $\bar{u}_i - \bar{y}_i \in \operatorname{ri} N_{\mathcal{C}}(F_i)$ ensure that $\Pi_{\mathcal{C}}$ is locally affine near $\bar{u}_i$, with Jacobian $P_{T_{F_i}}$. Consequently, F_{H}^N is continuously differentiable near $\bar{\theta}$. Denoting $h_i := P_{T_{F_i}} \nabla_y \ell(\bar{y}_i, z_i)$ and stacking $h := \big[h_1^\top \dots h_N^\top\big]^\top$, stationarity of $\bar{\theta}$ yields $0 = \nabla_\theta F_{\mathrm{H}}^N(\bar{\theta}) = \frac{1}{N} J^\mathsf{T} h$. By Lemma 13, J^T is injective, which means $h = 0$ and thus verifies the sample-wise tangential stationarity below.

$$P_{T_{F_i}} \nabla_y \ell(\bar{y}_i, z_i) = 0, \qquad i = 1, \dots, N. \tag{52}$$

Suppose, for the sake of contradiction, that $\bar{y}_i$ violates the KKT condition for some sample i. Tangential stationarity (52) and the active face assumptions render $(\bar{u}_i, \bar{y}_i)$ a projection-induced spurious stationary pair. Combining this with $N_{\mathcal{C}}(q_{\mathrm{R}}(\bar{u}_i)) \subseteq T_{F_i}^{\perp}$, Proposition 12 provides a neighborhood $\mathcal{U}_i$, a unit escape vector $v_i^{\mathrm{esc}} \in \mathcal{N}_{F_i}$, and positive constants $\bar{\alpha}_i, \underline{r}_i, c_i > 0$.

Let $V_i \in \mathbb{R}^{Nd}$ be the block vector placing v_i^{esc} in its i-th d-dimensional component and zeros elsewhere, and define the lifted parameter direction $d_i := J^\mathsf{T}(JJ^\mathsf{T})^{-1}V_i$. By Lemma 13, $J_i d_i = v_i^{\mathrm{esc}}$ for the i-th block Jacobian $J_i := J_{g_\theta}(z_i)$, while $J_j d_i = 0$ for all $j \neq i$, isolating the directional effect to sample i.

Because $\theta_k \to \bar{\theta}$ and $0 \leq \alpha_k \leq \eta_k \to 0$, sample i falls within $\mathcal{U}_i$ for all sufficiently large k. Whenever E_{k-1}^i occurs, its sample residual is bounded below by $\underline{r}_i$, driving the root-mean-square batch residual $\mathcal{R}_{k-1} \geq \underline{r}_i / \sqrt{|\mathcal{B}_{k-1}|} \geq \underline{r}_i / \sqrt{B_{\max}}$. Monotonicity of ψ then implies

$$\mathcal{R}_{k-1} \geq \frac{\underline{r}_i}{\sqrt{|\mathcal{B}_{k-1}|}} \geq \frac{\underline{r}_i}{\sqrt{B_{\max}}}.$$

By monotonicity of ψ, the following inequality holds for all sufficiently large k.

$$\alpha_k \geq \eta_k \psi\left(\frac{r_i}{\sqrt{B_{\max}}}\right) \mathbf{1}_{E^i_{k-1}} \tag{53}$$

For each training sample, writing $G_{k,j} := J_{\mathrm{OA},k}(u_k(z_j))^\mathsf{T} \nabla_y \ell(y_k(z_j), z_j)$, the directional parameter step along d_i evaluates to

$$\langle \theta_{k+1} - \theta_k, d_i \rangle = -\frac{\gamma_k}{|\mathcal{B}_k|} \sum_{z_j \in \mathcal{B}_k} \langle G_{k,j}, J_j d_i \rangle.$$

Due to the uncoupling property $J_j d_i = 0$ ($j \neq i$), only terms corresponding to sample i survive. When $z_i \in \mathcal{B}_k$ (i.e., E^i_k occurs), applying the escape lower bound yields

$$\langle \theta_{k+1} - \theta_k, d_i \rangle \geq \frac{\gamma_k c_i}{B_{\max}} \alpha_k \mathbf{1}_{E^i_k} \geq C_i \gamma_k \eta_k \mathbf{1}_{E^i_{k-1} \cap E^i_k}, \tag{54}$$

where $C_i := \frac{c_i}{B_{\max}} \psi\left(\frac{r_i}{\sqrt{B_{\max}}}\right) > 0$. Summing (54) from a sufficiently large iteration index K to $m-1$ gives

$$\langle \theta_m - \theta_K, d_i \rangle \geq C_i \sum_{k=K}^{m-1} \gamma_k \eta_k \mathbf{1}_{E^i_{k-1} \cap E^i_k}.$$

Taking $m \to \infty$, the divergence of the right-hand side via weighted recurrence (51) contradicts the convergence of the left-hand side to the finite limit $\langle \bar{\theta} - \theta_K, d_i \rangle$. Thus, no training sample can violate KKT conditions, confirming that (21) holds almost surely for all $i = 1, \ldots, N$. □

D Extended Experimental Results

This section provides the complete experimental specifications underlying Section 5. We first describe the common evaluation protocol, baseline implementations, hyperparameter selection, and the linear-probing architecture used throughout the study. We then provide the full construction and additional results for the two output-space experiments and the three parameterized benchmarks. Unless stated otherwise, all methods within the same experiment use identical problem instances, initializations, data splits, and outer-loop optimization budgets.

D.1 Experimental Setup

D.1.1 Evaluation metrics and common protocol

For a problem instance with context ξ, let $y^\star_\xi$ denote a high-accuracy reference solution of the original constrained problem and let $\widehat{y}_m(\xi)$ denote the decision returned by method m, which is an abstract representation of each baseline. We measure solution quality by the objective gap

$$\mathrm{Gap}_m(\xi) := f_\xi(\widehat{y}_m(\xi)) - f_\xi\left(y^\star_\xi\right). \tag{55}$$

Reported gaps are averaged across evaluation instances as well as initialization seeds when multiple initialization seeds are used. For feasible decisions, the gap is nonnegative up to numerical solver tolerance. Consequently, a negative gap from an approximate baseline does not indicate a better constrained solution and instead reflects an infeasible decision whose objective falls below the constrained optimum.

Feasibility is always evaluated in the original decision coordinates rather than in an internal reduced parameterization. For linear constraints $A_{\mathrm{eq}} y = b_{\mathrm{eq}}$ and $A_{\mathrm{ineq}} y \leq b_{\mathrm{ineq}}$, we record the violations $v_{\mathrm{eq}}(y) := \|A_{\mathrm{eq}} y - b_{\mathrm{eq}}\|_\infty$ and $v_{\mathrm{ineq}}(y) := \|[A_{\mathrm{ineq}} y - b_{\mathrm{ineq}}]_+\|_\infty$, declaring an output feasible once both metrics lie below the specified benchmark tolerance. This unified evaluation ensures a fair comparison.

Reference solutions are computed using exact solvers, such as OSQP. The same reference solutions are used by all methods. For methods that enforce feasibility only approximately, objective gaps are reported without filtering infeasible instances, while feasibility is shown separately. This convention exposes cases in which an apparently favorable objective value is obtained by violating the constraints.

D.1.2 BASELINES AND FAIR-COMPARISON SETTINGS

We compare OA projection against representative differentiable constraint-enforcement methods. OP maps the raw decision to its Euclidean projection onto the feasible set and uses the projection backward rule described in Section 3. SR projection (Schneider & Kuhn, 2026) maps the ambient space into the relative interior of the feasible set through a radial boundary map followed by a smooth radial contraction. RAYEN (Tordesillas et al., 2026) enforces convex constraints constructively by rescaling a raw direction according to its distance to the feasible-set boundary. DC3 (Donti et al., 2021b) combines exact equality completion with differentiable iterative correction of inequality violations. FSNet (Nguyen & Donti, 2025) initializes a differentiable feasibility-seeking procedure at the learned prediction and minimizes an unconstrained objective formed from constraint violations. Finally, ΠNet (Grontas et al., 2026) implements OP through Douglas-Rachford operator splitting together with an implicit backward pass.

A central fairness requirement concerns the SR component of OA projection. The standalone SR baseline and the SR branch inside OA projection use exactly the same mapping. Thus, if p_{SR} denotes the standalone SR implementation, the SR component entering $p_{\mathrm{OA},k} = (1-\alpha_k)\Pi_{\mathcal{C}} + \alpha_k p_{\mathrm{SR}}$ is numerically identical to that baseline.

Whenever equality constraints are eliminated before training, all compatible methods operate on the same reduced affine coordinates and are mapped back to the same original decision space for evaluation. For DC3, we implement equality completion using an orthonormal null-space parameterization. Specifically, let the columns of N form an orthonormal basis of $\ker(A_{\mathrm{eq}})$, and let $y_0(\xi)$ be a particular solution satisfying $A_{\mathrm{eq}} y_0(\xi) = b_{\mathrm{eq}}(\xi)$. We then write $y = y_0 + Nz$ with $N^\top N = I$. The original inequalities induce the reduced system $Gz \le h(\xi)$ with $G = A_{\mathrm{ineq}} N$ and $h(\xi) = b_{\mathrm{ineq}} - A_{\mathrm{ineq}} y_0(\xi)$. DC3 applies its differentiable inequality-correction steps to z in this reduced system. Because every iterate is reconstructed through the same affine map, the equality constraints are satisfied by construction. During training, the correction steps are unrolled and gradients are propagated through them.

Baseline-specific internal parameters, such as the number and step size of DC3 corrections, the feasibility-seeking budget of FSNet, and the forward/backward iteration counts of ΠNet, are tuned independently on the same validation split. The outer learning setup is fixed across methods, whereas each baseline uses the validation-selected configuration of its own constraint-handling procedure.

The output-space experiments operate without a learning model and directly optimize the raw decision variable. For each $m \in \{\mathrm{OP}, \mathrm{SR}, \mathrm{OA}, \mathrm{RAYEN}, \Pi\mathrm{Net}\}$, we solve $\min_u f\big(p_m(u)\big)$ from the same raw initialization using a common outer gradient-based optimization protocol and budget. This setting fixes the optimization variable, objective, initialization, and outer update procedure while varying the mechanism used to enforce the constraints and propagate gradients. The resulting update is governed by $J_{p_m}(u)^\top \nabla f(p_m(u))$, where J_{p_m} denotes the Jacobian or method-specific backward operator. These experiments therefore isolate how constraint handling reshapes the effective gradient field and affects the optimization trajectory and convergence, independently of model architecture, representation learning, and generalization.

Under this protocol, OP and SR expose the two geometric components combined by OA projection, while RAYEN provides an alternative radial parameterization. ΠNet implements OP through operator splitting and an implicit backward pass, enabling a comparison between different forward and backward realizations of projection under the same outer optimization protocol. DC3 and FSNet are evaluated in the parameterized benchmarks, as their designs intrinsically depend on learning-based prediction, auxiliary loss functions, or solver-dependent correction steps that cannot be decoupled into a pure output-space mapping.

D.1.3 OA CONFIGURATION AND HYPERPARAMETER SELECTION

The OA projection employs a residual-regulated gain $\alpha_k = \eta_k \psi(\mathcal{R}_{k-1})$, where the residual state $\mathcal{R}_{k-1}$ is detached from the autograd computation graph at iteration k. Across all experiments, the base learning rate follows a power-law schedule $\eta_k = (k+1)^{-a}$, and the gain function $\psi(r)$ is selected from the following saturating family:

$$\psi(r) \in \left\{ \frac{r}{r+\lambda},\ \tanh\left(\frac{r}{\lambda}\right),\ 1 - \exp\left(-\frac{r}{\lambda}\right) \right\}.$$

For the SR branch, the contraction function $\rho(t)$ adopts the standard configuration family from the SR baseline:

$$\rho(t) \in \left\{ \varepsilon + (1-\varepsilon)\frac{t}{t+\lambda},\ \varepsilon + (1-\varepsilon)\tanh\left(\frac{t}{\lambda}\right),\ \varepsilon + (1-\varepsilon)\left(1-\exp\left(-\frac{t}{\lambda}\right)\right) \right\}.$$

Consequently, OA projection and the standalone SR baseline share identical contraction configurations within each experimental setup. Depending on the setting, the decay schedule is clocked by either the iteration index or the normalized optimization progress.

Hyperparameters are selected strictly on the validation set using grid search grounded in official baseline configurations, and remain fixed across all test evaluations. Table 3 summarizes the hyperparameter search spaces used in our experiments.

Table 3: Hyperparameter grids used for model selection.

Method	Hyperparameter	Candidate values
SR	ρ, scale	$\{\text{frac}, \tanh, \exp\} \times \{1, 10, 50, 200\}$
RAYEN	radial function	Official linear-polytope map
DC3	correction step, stepsize, momentum	$\{5\} \times \{10^{-4}\} \times \{0.5\}$
FSNet	maximum iterations	50
ΠNet	forward/backward iterations	$\{20, 50, 100, 300\} \times \{25, 50, 100\}$
OA	ψ, η	$\{\text{frac}, \tanh, \exp\} \times \{0.1, 0.2, 0.4, 0.5\}$

All methods use the same conservative outer step size, chosen to ensure stable gradient-based optimization across the evaluated methods. The step size is fixed throughout validation and test evaluation, so the comparison primarily reflects the effect of the different constraint-handling maps rather than differences in optimizer tuning. Validation is used only to select the method-specific projection and mixing hyperparameters reported in Table 3. The numerical parameters of ΠNet are calibrated separately for projection accuracy without changing the common outer step size. All selected configurations remain fixed during test evaluation.

D.1.4 Linear-probing implementation

The three application benchmarks use a common linear-probing architecture. Each problem supplies a fixed feature map $\Phi : \mathcal{X} \to \mathbb{R}^m$, and only an affine head is optimized:

$$s_{W,b}(\xi) = W\Phi(\xi) + b, \qquad \widehat{z}_m(\xi) = p_m(s_{W,b}(\xi)). \tag{56}$$

The feature map is frozen throughout training, so gradients update only (W, b). All methods use the same feature vectors, head dimensions, head initialization, minibatch sequence, and outer optimizer. Their only difference is the constraint-handling mechanism p_m and any internal computation required by that mechanism.

When affine equalities are present, we parameterize the corresponding affine hull as

$$y = y_{\text{eq}} + Nz, \qquad A_{\text{eq}}N = 0, \qquad N^{\mathsf{T}}N = I,$$

and let the trainable head predict z. This avoids asking the learned head to rediscover deterministic equality structure and places all compatible constraint methods on the same reduced decision space. Inequality constraints are transformed accordingly and the final output is mapped back to the original coordinates before computing the objective and evaluation metrics.

D.2 Output-Space Optimization

The following experiments remove the learning model entirely and optimize the raw variable entering the constraint map. They therefore isolate differences in first-order geometry from representation capacity or feature learning.

D.2.1 CONVEX QP PROGRAM

Problem generation. We follow the QP generation scheme of DC3 (Donti et al., 2021b). A convex QP family is defined by

$$\begin{aligned} \min_{y\in\mathbb{R}^n} \quad & \frac{1}{2}y^\mathsf{T}Qy + p^\mathsf{T}y, \\ \text{s.t.} \quad & Ay = x, \qquad Gy \le h, \end{aligned} \tag{57}$$

where Q, p, A, G, h are fixed within a family and the equality right-hand side x varies between instances. We use $n = 100$, $n_{\text{eq}} = 50$ and $n_{\text{ineq}} = 50$. Following the DC3 construction, Q is diagonal with entries sampled from $\text{Unif}[0, 1]$, p is sampled from $\text{Unif}[0, 1]^n$, the entries of A and G are standard normal, and $x \sim \text{Unif}[-1, 1]^{n_{\text{eq}}}$. The inequality bound is defined by $h_i = \sum_j \left|(GA^\dagger)_{ij}\right|$, which guarantees that $y = A^\dagger x$ is feasible for every generated context.

For each structural QP family, we generate $10{,}000$ contexts, using the first $1{,}000$ for model selection and the remaining $9{,}000$ for evaluation. Each evaluation instance is optimized from five shared raw initializations. We repeat the complete construction for structural seeds $\{17, 23, 42\}$. Thus, each family-level statistic aggregates $45{,}000$ optimization trajectories. Reference objectives are obtained with OSQP with absolute and relative tolerances of 10^{-8}. Feasibility is assessed in the original y-coordinates with tolerance 10^{-7}.

Main-family results. Table 4 presents the central result. OA projection achieves the smallest mean and median objective gaps for every structural family while retaining complete feasibility. Averaged across families, OA projection reduces the mean objective gap from 0.0701 for the next-best method, SR, to 0.0455, and reduces the median objective gap from 0.0680 to 0.0113, corresponding to reductions of 35.1% and 83.4%, respectively. The same ordering across independently generated families shows that the improvement is not tied to a particular realization of the QP geometry.

Table 4: Output-space optimization on the parametric QP benchmark. Within each family, statistics are computed over $9{,}000$ instances and five shared initializations. Reported gaps are the mean $\pm$ standard deviation of the corresponding family-level statistics over three structural seeds.

Method	Mean gap	Median gap	Feasible (%)
OP	1.2807 ± 0.2369	0.7829 ± 0.2618	100.0
SR	0.0701 ± 0.0169	0.0680 ± 0.0172	100.0
RAYEN	0.0761 ± 0.0162	0.0738 ± 0.0166	100.0
ΠNet	1.2585 ± 0.2274	0.7737 ± 0.2612	100.0
OA	$\mathbf{0.0455 \pm 0.0303}$	$\mathbf{0.0113 \pm 0.0080}$	100.0

First-order optimality. To determine whether the objective improvement is accompanied by better first-order optimality, we evaluate the projected-gradient natural residual

$$\mathcal{R}_\gamma(z) = \frac{1}{\gamma}\left\|z - \Pi_{\mathcal{Z}_x}\left(z - \gamma\nabla\phi_x(z)\right)\right\|_2, \tag{58}$$

where ϕ_x is the reduced QP objective and the same conservative γ is used for all methods within a family. This residual vanishes at a first-order solution and does not require estimating an active constraint set. It is computed only after optimization and therefore does not affect any trajectory.

$$\mathcal{R}_\gamma(z) = 0 \quad \Longleftrightarrow \quad 0 \in \nabla\phi_x(z) + N_{\mathcal{Z}_x}(z) \quad \Longleftrightarrow \quad z \text{ satisfies the KKT conditions.}$$

OA projection attains the lowest aggregate mean and median natural residual among the five methods. Together with its lower objective gap, this result indicates that OA projection reaches decisions that are closer to first-order optimality within the common optimization budget.

D.2.2 NONCONVEX PROGRAM

This benchmark examines whether the optimization advantage of OA projection persists when the constrained landscape contains genuine nonconvex curvature and multiple local basins. Starting from

Table 5: Natural-residual results on the parametric QP benchmark. Within each family, statistics are computed over $9{,}000$ instances and five shared initializations. Entries report the mean $\pm$ standard deviation of the corresponding family-level statistics over three structural seeds.

Method	Mean natural residual	Median natural residual
OP	1.1679 ± 0.1452	1.0957 ± 0.1774
SR	0.1694 ± 0.0399	0.1608 ± 0.0388
RAYEN	0.1824 ± 0.0388	0.1732 ± 0.0376
ΠNet	1.1575 ± 0.1426	1.0895 ± 0.1766
OA	$\mathbf{0.1543 \pm 0.0857}$	$\mathbf{0.1102 \pm 0.0686}$

the parametric QP construction in Section D.2.1, we retain the same feasible sets and add a structured oscillatory perturbation to the reduced objective as follows,

$$\begin{aligned} \min_{z \in \mathbb{R}^{50}} \quad & F_i(z) = z^\mathsf{T} H z + g_i^\mathsf{T} z + c_i - \frac{\beta}{m} \sum_{j=1}^{m} \cos\left(\frac{\omega}{\ell} u_j^\mathsf{T} z + \phi_{ij}\right), \\ \text{s.t.} \quad & Gz \le h_i. \end{aligned} \tag{59}$$

Here, i indexes the problem context, $m = 4$, $\omega = 2\pi$, and $\{u_j\}_{j=1}^m$ are shared orthonormal directions. The phases ϕ_{ij} vary across problem instances but are fixed across methods. The length scale ℓ is determined from feasible directional radii on the validation split. The amplitude β is selected separately for each structural family so that the oscillatory curvature is 1.25 times the median quadratic curvature along the selected directions. Under this construction, every validation instance admits a feasible point with negative directional curvature.

We use the same three structural seeds, $10{,}000$ contexts per family, validation–evaluation split, and five shared raw initializations as in the convex QP benchmark. The raw initializations are sampled between four and eight boundary radii from the feasible anchor and are verified to lie outside the feasible set. All methods optimize the same reduced variables for $3{,}000$ gradient steps using a common conservative step size derived from an upper bound on the gradient Lipschitz constant. RAYEN and ΠNet use their official implementations and the same validation protocol as in the preceding experiment.

Reference objectives are computed using SCIP spatial branch-and-bound on an equivalent reformulation of each nonconvex QP. Only references with relative optimality gap at most 10^{-4} are retained. The reported global objective gap is the difference between the final objective and this SCIP reference.

Table 6 reports the central comparison. OA projection obtains the smallest mean and median global objective gaps in every structural family while maintaining complete feasibility. Across families, OA projection reduces the mean gap from 0.0682 for the next-best method, SR, to 0.0242, and reduces the median gap from 0.0634 to 0.0187. These correspond to reductions of 64.5% and 70.5%, respectively. RAYEN closely tracks SR, whereas OP and ΠNet terminate in substantially higher-objective basins under the same optimization budget.

Table 6: Output-space optimization on the nonconvex benchmark. Within each family, statistics are computed over $9{,}000$ instances and five shared initializations. Entries report the mean $\pm$ standard deviation of the family-level statistics over three structural seeds.

Method	Mean global gap	Median global gap	Feasible (%)
OP	9.7081 ± 0.6309	9.2744 ± 0.6268	100.0
SR	0.0682 ± 0.0164	0.0634 ± 0.0166	100.0
RAYEN	0.0695 ± 0.0162	0.0649 ± 0.0164	100.0
ΠNet	9.9816 ± 0.6045	9.4066 ± 0.6060	100.0
OA	$\mathbf{0.0242 \pm 0.0017}$	$\mathbf{0.0187 \pm 0.0011}$	100.0

We further evaluate the natural residual defined in (58), using $\gamma = 1/L$ with the same family-specific Lipschitz bound used to determine the outer step size. Its zero set remains equivalent to the KKT first-order condition, but in the nonconvex setting, this condition characterizes stationarity rather than

global optimality. We therefore use the global objective gap as the primary metric and the natural residual as supporting evidence of first-order convergence.

As shown in Table 7, OA projection also achieves the smallest mean and median natural residual. In particular, it reduces the aggregate median residual from 0.2305 for SR to 0.1005. The lower objective gap is therefore accompanied by closer convergence to a first-order solution, rather than arising solely from a different feasible parameterization.

Table 7: Natural-residual results on the nonconvex benchmark. Aggregation follows Table 6.

Method	Mean natural residual	Median natural residual
OP	3.9832 ± 0.1755	3.9812 ± 0.1804
SR	0.2381 ± 0.0525	0.2305 ± 0.0536
RAYEN	0.2423 ± 0.0517	0.2350 ± 0.0525
ΠNet	4.0364 ± 0.1774	4.0139 ± 0.1804
OA	$\mathbf{0.1182 \pm 0.0027}$	$\mathbf{0.1005 \pm 0.0021}$

D.3 Linear-Probing Benchmarks

To evaluate whether the optimization advantages observed directly in output space persist under learned parameterizations, we evaluate all methods within a unified linear-probing framework. Linear probing is particularly suited for transfer settings where a pretrained representation is already informative and adaptation is confined to the output mapping. Compared with full fine-tuning, this paradigm drastically reduces trainable parameters, preserves pretrained feature structures, and prevents generalization collapse induced by feature distortion (Kumar et al., 2022).

Across all three benchmarks, each predictor combines a task-specific Multilayer Perceptron (MLP) backbone with an affine output head. We adopt a two-stage protocol by first pretraining the backbone using an auxiliary regression loss against solver-generated optimal decisions, and then freezing its parameters to optimize only the linear head under the respective constraint-handling layers. Crucially, within each benchmark, all constraint methods share identical pretrained representations and head initializations. This protocol provides a strictly controlled evaluation. Because the feature representation is held fixed, observed performance differences isolate how the constraint geometry of each method shapes the gradients received by the same trainable head, decoupled from backbone dynamics.

D.3.1 DC OPF

Problem and scenario generation. We use the MATPOWER IEEE 57-bus system with seven generators. Network topology, generator limits, and quadratic cost parameters are instantiated from MATPOWER, with thermal branch capacities and angle-difference limits supplemented by the PGLib-OPF case57 instance. For a given load scenario ξ, the decision vector consists of active generator outputs $P_g \in \mathbb{R}^7$ and bus voltage angles $\theta \in \mathbb{R}^{57}$. Each instance solves the following program,

$$\begin{aligned} \min_{P_g,\theta} \quad & \sum_{j=1}^{7} \left[c_{2,j} P_{g,j}^2 + c_{1,j} P_{g,j} + c_{0,j} \right], \\ \text{s.t.} \quad & C_g P_g - P_d(\xi) = B_{\text{bus}} \theta + P_{\text{businj}}, \\ & P_g^{\min} \le P_g \le P_g^{\max}, \\ & -\bar{f} \le B_f \theta + P_{\text{finj}} \le \bar{f}, \\ & \underline{\delta} \le A_\Delta \theta \le \overline{\delta}, \qquad \theta_r = 0, \end{aligned} \tag{60}$$

where C_g maps generators to buses, B_{bus} is the DC nodal susceptance matrix, B_f maps angles to branch flows, and A_Δ computes line angle differences. While the network parameters and cost coefficients remain static, instance variability is driven by the stochastic demand vector $P_d(\xi)$. The input context $\xi \in \mathbb{R}^{42}$ represents the standardized demand perturbations at the 42 buses with positive base loads. To generate spatial correlation and guarantee non-negative load profiles, the raw demand vector $P_d \in \mathbb{R}^{42}$ is constructed via a Gaussian copula as $P_{d,j} = F_j^{-1}\Big(\Phi(Z_j)\Big)$, where $Z \sim \mathcal{N}(\mathbf{0}, \Sigma)$

is a latent Gaussian vector with an equicorrelated covariance matrix $\Sigma_{ij} = 0.5 + 0.5\delta_{ij}$, and $\Phi(\cdot)$ denotes the standard normal CDF. The marginal CDF F_j corresponds to a lower-truncated normal distribution $\mathcal{TN}(0, \infty; \mu_j, \sigma^2)$ with latent scale $\sigma = 20\,\mathrm{MW}$. The latent location parameter μ_j is uniquely determined such that the post-truncation expectation equals the MATPOWER base load, i.e., $\mathbb{E}[P_{d,j}] = P_{d,j}^{\mathrm{base}}$. Finally, ground-truth optimal solutions are computed using OSQP, configured with solution polishing enabled and absolute/relative optimality tolerances set to 10^{-9}.

Linear-probing protocol. By eliminating the 57 power-balance equalities and fixing the reference angle, the decision space reduces from 64 to 6 dimensions, subject to 334 remaining generator, line-flow, and angle-difference inequalities. A linear program provides a shared strictly interior anchor for each instance. All methods operate on these reduced coordinates and are evaluated after reconstructing the original (P_g, θ) space.

For each data seed $s \in \{0, 1, 2, 3, 4\}$, we generate 5,000 training and 2,000 IID test instances. The training set is partitioned into a 4,000-instance fit split and a 1,000-instance validation split. A two-hidden-layer MLP (width 128, $\tanh$ activations) is pretrained on the fit split for 100 epochs to regress the OSQP ground-truth solutions. We then freeze the feature backbone, discard the auxiliary output head, and attach a fresh affine head. Across all constraint-handling methods, this head is trained for 200 epochs on the fit split using Adam with a learning rate of 3×10^{-3} and a batch size of 256, sharing identical initializations and minibatch sequences to guarantee controlled comparisons.

Hyperparameters for SR and OA projection are tuned on the validation split after a 50-epoch candidate training phase. Crucially, while OA method trains the head via the OA operator, both model selection and test-time inference are performed using OP. This setup isolates the explicit impact of OA projection on the optimization trajectory while retaining OP as the standard deployed feasibility layer. For DC3, we adopt the same equality-eliminated coordinate space and configure it with 5 correction steps (step size 10^{-4}, momentum 0.5, and stopping tolerance 10^{-4}). Feasibility for DC3 is evaluated in the original (P_g, θ) physical space with a tolerance of 10^{-5}.

To evaluate solution quality, we measure the relative objective gap $\Delta_{\mathrm{rel}}(z) = \frac{\phi_\xi(z) - \phi_\xi(z^\star)}{|\phi_\xi(z^\star)|}$, which quantifies the percentage cost increase over the reference OSQP optimum $z^\star$. For a feasible decision, the relative gap is nonnegative up to solver tolerance. We report feasibility separately because an infeasible decision can attain an artificially low and possibly negative relative gap. We additionally report the scaled projected-gradient natural residual, which vanishes exactly at a KKT solution of the reduced convex QP and is computed only after training.

Results. Table 8 summarizes the linear-probing performance on the IEEE 57-bus DC-OPF benchmark. OA projection achieves the lowest relative objective gap and natural residual among all strictly feasible methods. Specifically, relative to standard OP, OA projection reduces the mean relative gap from 0.4772% to 0.3538% with a 25.9% cost improvement, while simultaneously lowering the natural residual from 0.0520 to 0.0443. SR and RAYEN also retain full feasibility but yield suboptimal dispatches with significantly larger objective gaps (0.7117% and 0.5707%, respectively). ΠNet approaches the OP gap but incurs an average constraint violation of 1.71×10^{-5}. FSNet and DC3 exhibit both larger objective gaps (3.52% and 1.45%) and violations up to 8.55×10^{-3}. The objective improvement of OA projection is therefore accompanied by closer first-order optimality rather than a relaxation of the physical constraints.

D.3.2 Constrained trajectory optimization

Problem and scenario generation. We evaluate a safe-corridor trajectory optimization benchmark adapted from Tordesillas et al. (2026), which models collision-free, dynamically constrained motion planning for mobile robots. The continuous trajectory is parameterized by a clamped uniform quadratic B-spline over six time intervals, defined by eight position control points $c = [c_1^\mathsf{T}, \ldots, c_8^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{16}$. Following the benchmark specification, the spline spans 11 knots across a fixed horizon $t_f = 35\,\mathrm{s}$ through three overlapping convex polyhedral corridors. For a scenario context $\xi = (\gamma_v, \gamma_a, \gamma_j, p_f)$, each instance solves the program below.

$$\begin{aligned} \min_{c \in \mathbb{R}^{16}} \quad & J_\xi(c) := \|p(t_f; c) - p_f\|_2^2 + \gamma_v V(c) + \gamma_a A(c) + \gamma_j J(c), \\ \text{s.t.} \quad & A_{\mathrm{eq}} c = b_{\mathrm{eq}}(p_0, v_0), \qquad A_{\mathrm{ineq}} c \le b_{\mathrm{ineq}}(s, \lambda_v, \lambda_a). \end{aligned} \tag{61}$$

Table 8: Linear-probing evaluation on IEEE 57-bus DC-OPF. Each seed contains 2,000 IID test instances. Entries are the mean ± standard deviation of seed-level statistics over five data seeds. Relative gaps are reported in percent. Mean constraint violation averages the per-instance maximum constraint violation over the test set.

Method	Relative gap (%)	Natural residual	Mean constraint violation
OP	0.4772 ± 0.2145	0.0520 ± 0.0119	$(1.86 \pm 0.71) \times 10^{-12}$
SR	0.7117 ± 0.0746	0.0733 ± 0.0026	0
RAYEN	0.5707 ± 0.0747	0.0642 ± 0.0033	$(2.79 \pm 0.06) \times 10^{-17}$
IINet	0.4709 ± 0.2175	0.0524 ± 0.0116	$(1.71 \pm 0.48) \times 10^{-5}$
FSNet	3.5196 ± 3.5494	0.2072 ± 0.1095	$(4.80 \pm 2.26) \times 10^{-5}$
DC3	1.4477 ± 0.1531	0.1233 ± 0.0068	$(8.55 \pm 1.02) \times 10^{-3}$
OA	$\mathbf{0.3538 \pm 0.0565}$	$\mathbf{0.0443 \pm 0.0028}$	$(9.70 \pm 3.56) \times 10^{-13}$

The equality constraints enforce initial and terminal kinematic conditions, whereas the inequality constraints impose velocity and acceleration bounds alongside corridor containment. By the convex-hull property of MINVO control points, enforcing containment of c within the polyhedra strictly guarantees continuous trajectory safety over $t \in [0, t_f]$ rather than only at discrete time checkpoints.

Scenario generation varies both objective weights and physical boundaries. Penalty weights $\gamma_v, \gamma_a, \gamma_j$ are sampled independently from $\mathrm{Unif}[0, 2]$. Corridor geometry is randomized by scaling half-space offsets as $b_j(s_j) = s_j b_j + (1 - s_j) A_j \mu_j$, where μ_j is the corridor centroid, $s_j = 0.75 + 0.25 B_j$, and $B_j \sim \mathrm{Beta}(2, 2)$. Initial positions p_0 and target locations p_f are drawn uniformly by area within the scaled initial and terminal corridors, respectively. Kinematic limits $v_{\max} = 2.8\lambda_v\,\mathrm{m/s}$ and $a_{\max} = 1.5\lambda_a\,\mathrm{m/s^2}$ are parameterized by $\lambda_v \sim \mathrm{Unif}[0.65, 1.05]$ and $\lambda_a \sim \mathrm{Unif}[0.15, 1.05]$.

Linear-probing protocol. Following the five-seed protocol established in Section D.3.1, each data seed comprises 5,000 training and 2,000 IID test instances. Certified ground-truth optimal solutions are computed using OSQP with solution polishing enabled and both absolute and relative tolerances set to 10^{-9}. Model selection on the validation split identifies a $\tanh$ radial SR operator configured with an ε parameter of 0.05 and a scale of 10, paired with a $\tanh$ OA operator using a decay power of 0.1. Regarding baseline setups, DC3 executes 5 correction steps with a step size of 10^{-4} and a momentum of 0.5, whereas FSNet utilizes 20 L-BFGS correction iterations. Finally, model performance is evaluated across the objective gap, natural residual, and mean constraint violation under a unified feasibility tolerance of 10^{-5}.

Results. Table 9 summarizes the linear-probing performance across five data seeds. Among all feasible methods that maintain near-zero constraint violations, OA projection achieves the lowest mean objective gap of 6.2863. Compared to standard OP, which yields an objective gap of 7.4778, OA projection reduces this gap by 15.9% while maintaining a natural residual of 0.2259 that remains closely aligned with OP's 0.2212. Although SR and RAYEN also preserve exact feasibility, they converge to substantially higher objective gaps of 10.9879 and 11.0111, respectively. Notably, OA projection achieves a 42.9% reduction in objective gap relative to RAYEN. In contrast, IINet incurs an average constraint violation of 4.47×10^{-5}, whereas FSNet and DC3 suffer from severe constraint violations ranging between 10^{-4} and 10^{-2} alongside significantly worse objective gaps.

D.3.3 Minimum-cost graph flow

Problem and scenario generation. The final benchmark evaluates capacitated transport over a fixed layered logistics network consisting of 3 warehouse nodes, 5 transfer hubs, 7 demand nodes, and 40 directed arcs. Eliminating one redundant flow conservation equation yields a full-row-rank incidence matrix $D \in \mathbb{R}^{14 \times 40}$, accompanied by a net supply vector $b \in \mathbb{R}^{14}$ and arc capacities $u \in \mathbb{R}^{40}_{++}$. For a given scenario context $q_t \in \mathbb{R}^{10}$, the arc flow vector $x_t \in \mathbb{R}^{40}$ solves the following quadratic flow problem

$$\begin{aligned} \min_{x_t \in \mathbb{R}^{40}} \quad & \frac{1}{2} x_t^\mathsf{T} \,\mathrm{Diag}(H_t) x_t + c_t^\mathsf{T} x_t, \\ \text{s.t.} \quad & D x_t = b, \qquad 0 \le x_t \le u. \end{aligned} \tag{62}$$

Table 9: Linear-probing evaluation on the parametric trajectory benchmark. Each seed contains 2,000 independent IID test instances. Entries are the mean ± standard deviation of seed-level statistics over five data seeds. Mean objective gaps are reported. Mean constraint violation averages the per-instance maximum constraint violation over the test set.

Method	Mean objective gap	Natural residual	Mean constraint violation
OP	7.4778 ± 0.2716	$\mathbf{0.2212 \pm 0.0069}$	$(6.82 \pm 1.18) \times 10^{-12}$
IINet	7.6840 ± 0.2782	0.2349 ± 0.0871	$(4.47 \pm 0.53) \times 10^{-5}$
SR	10.9879 ± 0.6505	0.3015 ± 0.0092	$(5.73 \pm 0.44) \times 10^{-17}$
RAYEN	11.0111 ± 0.6698	0.3012 ± 0.0092	$(9.00 \pm 0.60) \times 10^{-17}$
FSNet	16.0044 ± 6.4540	0.3677 ± 0.0773	$(5.26 \pm 2.96) \times 10^{-4}$
DC3	20.7637 ± 0.2097	0.5021 ± 0.0232	$(6.56 \pm 0.95) \times 10^{-2}$
OA	$\mathbf{6.2863 \pm 0.2817}$	0.2259 ± 0.0061	$(7.32 \pm 1.80) \times 10^{-14}$

The vectors $H_t \in \mathbb{R}^{40}_{++}$ and $c_t \in \mathbb{R}^{40}_{+}$ parameterize scenario-dependent congestion and transportation costs. By enforcing $(H_t)_e > 0.05$ for every arc e, the objective function remains strictly convex, guaranteeing a unique optimal flow vector for each scenario. Graph topology, node supplies, demands, and arc capacities remain invariant across all contexts.

Linear-probing protocol. In alignment with the five-seed evaluation protocol outlined in Section D.3.1, every random seed incorporates $5{,}000$ training samples alongside $2{,}000$ independent and identically distributed test instances. We generate exact target optimal flows via OSQP by enabling solution polishing and enforcing stringent absolute and relative solver tolerances of 10^{-9}. Hyperparameter search on the validation set selects an exponential radial SR operator featuring an ε value of 0.05 and unit scaling, working alongside a fractional OA operator configured with a decay exponent of 0.5. For the competing baselines, DC3 runs 20 correction iterations using a validation-tuned step size of 0.2, momentum of 0.5, and stopping tolerance of 10^{-4}, whereas FSNet implements 20 L-BFGS correction steps. Model quality is assessed through the relative objective gap, natural residual, and mean constraint violation, all benchmarked against a uniform feasibility tolerance of 10^{-5}.

Results. Table 10 summarizes the linear-probing performance across five data seeds. Among all evaluated methods, OA projection achieves both the lowest mean objective gap of 0.0810 and the lowest natural residual of 0.1647, while maintaining strict feasibility across all test predictions with constraint violations on the scale of 10^{-13}. Compared to standard OP, which yields an objective gap of 0.2260 and a natural residual of 0.2745, OA projection achieves a substantial 64.1% reduction in objective cost. Meanwhile, IINet obtains the second-lowest objective gap of 0.0934, but incurs a minor mean constraint violation of 2.52×10^{-4} due to incomplete convergence during its finite projection steps. DC3 achieves tighter adherence to feasibility with a low mean constraint violation of 5.13×10^{-7}, though its objective gap remains markedly higher at 2.1392. Finally, although SR and RAYEN preserve exact feasibility with constraint violations bounded below 10^{-15}, both methods converge to substantially higher-cost dispatches with a mean objective gap of 5.0558.

Table 10: Linear-probing evaluation on the minimum-cost flow benchmark. Each seed contains 2,000 independent IID test instances. Entries are the mean ± standard deviation of seed-level statistics over five data seeds. Mean objective gaps are reported. Mean constraint violation averages the per-instance maximum constraint violation over the test set.

Method	Objective gap	Natural residual	Mean constraint violation
OP	0.2260 ± 0.0067	0.2745 ± 0.0040	$(2.38 \pm 0.06) \times 10^{-14}$
SR	5.0558 ± 0.1586	1.7934 ± 0.0327	$(5.89 \pm 0.14) \times 10^{-16}$
RAYEN	5.0558 ± 0.1586	1.7934 ± 0.0327	$(5.93 \pm 0.15) \times 10^{-16}$
IINet	0.0934 ± 0.0102	0.1808 ± 0.0107	$(2.52 \pm 0.69) \times 10^{-4}$
FSNet	2.6864 ± 0.0393	1.1413 ± 0.0121	$(2.39 \pm 0.20) \times 10^{-3}$
DC3	2.1392 ± 0.0652	0.9116 ± 0.0188	$(5.13 \pm 4.08) \times 10^{-7}$
OA	$\mathbf{0.0810 \pm 0.0070}$	$\mathbf{0.1647 \pm 0.0086}$	$(2.25 \pm 0.27) \times 10^{-13}$