Soft-Radial Projection for Constrained End-to-End Learning

Philipp J. Schneider **Daniel Kuhn**
EPFL
{philipp.schneider, daniel.kuhn}@epfl.ch

Abstract

Constrained end-to-end learning trains neural networks whose outputs must satisfy feasibility constraints, such as resource, budget, or operational limits. While projection and optimization layers guarantee constraint satisfaction, boundary-based projections can introduce unfavorable optimization geometry: regions of the unconstrained output space are mapped to the same active face of the feasible set. This makes the resulting map locally rank-deficient, suppressing gradients in constrained directions and potentially impairing optimization—a bottleneck known as gradient saturation. We propose *Soft-Radial Projection (SRP)*, a differentiable reparameterization layer that maps unconstrained network outputs into the relative interior of a convex feasible set by smoothly rescaling rays from a strictly feasible anchor point. Unlike hard radial or orthogonal projections, the resulting map is one-to-one and has a full-rank Jacobian almost everywhere, while guaranteeing strict feasibility. We prove that constrained networks equipped with SRP retain universal approximation guarantees. For polytopes and balls, the layer admits a closed-form forward pass, while for general convex sets, the ray–boundary intersection reduces to a one-dimensional search. Experiments on decision-focused learning benchmarks demonstrate strong solution quality at low computational cost, with particularly favorable results in portfolio optimization and substantially lower inference latency than differentiable optimization layers in the evaluated settings.

1 Introduction

Many decision-making systems operate under hard constraints during both training and deployment—for example, safety envelopes in autonomous driving, actuator limits in robotics [1], or budget and capacity constraints in operations. While neural networks are powerful function approximators, they are not inherently constraint-aware, and naive usage can lead to *infeasible* outputs precisely in scenarios where constraint violations have serious consequences. We formalize constrained learning objectives and propose a strict-interior projection layer that enforces feasibility *by construction*.

Setup. Let $\mathcal{Z} \subseteq \mathbb{R}^d$ be the input space, $\mathcal{Y} \subseteq \mathbb{R}^n$ the target space, and $\mathcal{C} \subseteq \mathbb{R}^n$ the feasible set of decisions, assumed closed and convex with a nonempty interior. This assumption is without loss of generality, as we can restrict the ambient space to the affine hull of the convex set and formulate all rank and Jacobian statements in the corresponding coordinates. Data are drawn from a distribution $\mathbb{P}$ on $\mathcal{Z} \times \mathcal{Y}$ with marginal $\mathbb{P}_Z$ on $\mathcal{Z}$. We define the policy space Π as the set of all measurable mappings $\pi : \mathcal{Z} \to \mathbb{R}^n$ that map an input z to a decision.

Supervised learning. Let $(z, y) \sim \mathbb{P}$. We learn an optimal policy by solving

$$\min_{\pi \in \Pi} \quad \left\{ \mathbb{E}_{(z,y)\sim\mathbb{P}} \big[\ell\big(\pi(z), y\big) \big] : \pi(z) \in \mathcal{C} \text{ for } \mathbb{P}_Z\text{-a.e. } z \right\}, \tag{1}$$

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

where $\ell : \mathcal{C} \times \mathcal{Y} \to \mathbb{R}$ is a chosen loss. This formulation covers standard regression tasks as well as imitation learning, where the label y represents a hindsight-optimal decision $\pi^\star(z)$. In such cases, ℓ quantifies the distance to the optimal decision or a regret surrogate.

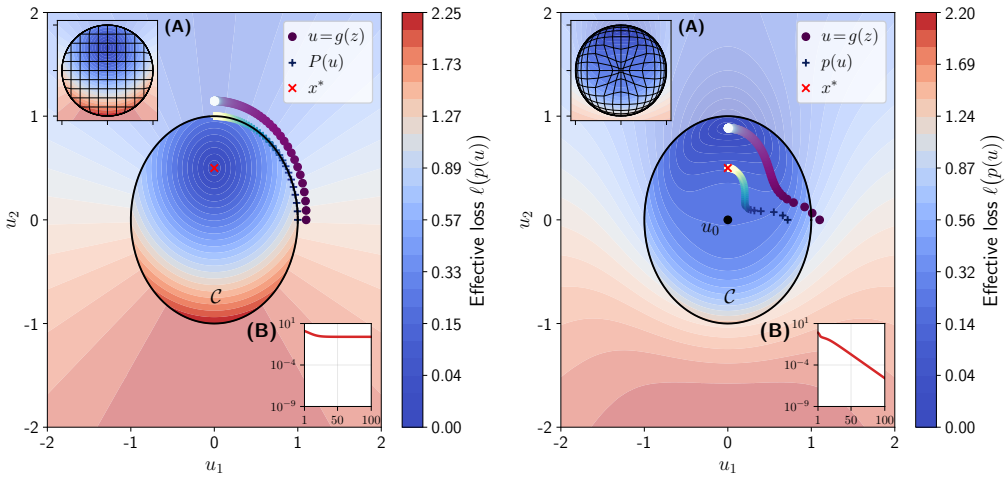
Task-driven learning. Without labels, we measure performance directly via a task cost $c : \mathcal{Z} \times \mathcal{C} \to \mathbb{R}$ and optimize

$$\min_{\pi \in \Pi} \quad \left\{ \mathbb{E}_{z \sim \mathbb{P}_Z} \left[c\big(z, \pi(z)\big) \right] : \pi(z) \in \mathcal{C} \text{ for } \mathbb{P}_Z\text{-a.e. } z \right\}. \tag{2}$$

This formulation aligns the training objective with the downstream operational cost, thereby circumventing the potential objective mismatch inherent in surrogate supervised targets [2–5].

Feasibility by construction. A common way to enforce feasibility is to compose an unconstrained candidate map $g : \mathcal{Z} \to \mathbb{R}^n$ with a fixed constraint-enforcement map $\mathrm{Proj} : \mathbb{R}^n \to \mathcal{C}$, yielding a feasible policy $\pi(z) = \mathrm{Proj}(g(z))$. This avoids penalties and post-hoc repair, but the geometry of Proj strongly affects optimization. Boundary-valued maps, such as orthogonal projections or hard radial/gauge projections, enforce feasibility by collapsing infeasible points onto $\partial\mathcal{C}$. As a result, they are typically many-to-one outside the feasible set and can induce rank-deficient Jacobians, suppressing gradient components in the collapsed directions.

The standard example is the orthogonal projection $P(u) = \arg\min_{v \in \mathcal{C}} \|u - v\|^2$. Related hard radial and gauge-based layers project along rays from an interior anchor to the boundary [6–8]. These constructions are computationally attractive and form an important basis for our work. However, these maps remain many-to-one outside the feasible set: hard radial and gauge-based projections collapse points along exterior rays, while orthogonal projection collapses points along outward normal directions. Consequently, the gradient-saturation phenomenon illustrated in Figure 1 is not specific to the orthogonal projection.



(a) Orthogonal projection $P(u)$.

(b) Soft-radial projection $p(u)$ into $\mathrm{Int}(\mathcal{C})$.

Figure 1: Impact of projection geometry on optimization. Comparison of (a) Orthogonal and (b) Soft-Radial Projection on a 2D constrained task (target $\times$). Main plots visualize the trajectory of the *unconstrained candidate* ($\circ$) versus the *feasible decision* ($+$). Insets show the coordinate grid warping (A) and training loss (B). Soft-radial projection preserves nonzero radial derivatives for infeasible inputs.

Soft-Radial Projection to the interior. We propose *Soft-Radial Projection* (SRP), a strict-interior alternative to hard radial clipping. Starting from the same ray-based geometry as gauge and radial projection layers, SRP replaces boundary clipping with a soft radial contraction that maps every $u \in \mathbb{R}^n$ into $\mathrm{Int}(\mathcal{C})$. Under mild assumptions, the resulting map is injective, defines a homeomorphism $p : \mathbb{R}^n \to \mathrm{Int}(\mathcal{C})$, and has an invertible Jacobian wherever it is differentiable, avoiding the many-to-one collapse of hard radial, gauge-based, and orthogonal projections. For polytopes and balls, SRP admits a closed-form forward pass, while for general convex sets, the ray–boundary intersection reduces to a one-dimensional search.

Related work. The challenge of obtaining feasibility guarantees for neural networks has attracted broad interest, from safety-critical control [9] to decision-focused analytics [10]. Existing methodologies split into *optimization-based layers* [*e.g.*, 11–13], which differentiate through iterative solvers, and *constructive layers* that enforce feasibility by design [*e.g.*, 14–16, 8]. Most closely related are ray- and gauge-based constructive layers, such as RAYEN [8] and the ray-projection method of Konstantinov and Utkin [14], which also exploit radial geometry. Our contribution is to replace hard boundary clipping with a strict-interior homeomorphic reparameterization, preserving the computational appeal of ray-based enforcement while avoiding the rank-deficient projection Jacobians induced by clipping. A detailed comparison is provided in Appendix A.

Contributions. Our main contributions are:

- We introduce *Soft-Radial Projection* (SRP), a ray-based enforcement layer for closed convex sets that replaces boundary clipping with a homeomorphic reparameterization of $\mathbb{R}^n$ onto $\mathrm{Int}(\mathcal{C})$.
- We prove that SRP has an invertible Jacobian almost everywhere and establish value equivalence, interior stationarity, bounded-latent value gaps, and constrained universal approximation.
- We demonstrate strong solution quality at low computational cost on decision-focused learning benchmarks, with particularly favorable results in portfolio optimization.

We proceed by formally introducing the soft-radial projection mechanism.

Notation. We use $\|u\|$ for the Euclidean norm of vectors and $\|A\|$ for the spectral norm of matrices. For functions $\psi : \mathcal{Z} \to \mathbb{R}^n$, we denote the uniform norm by $\|\psi\|_\infty := \sup_{z\in\mathcal{Z}} \|\psi(z)\|$. We write $\mathrm{Int}(\mathcal{C})$ and $\partial\mathcal{C}$ for the interior and boundary of the feasible set. The Jacobian of a differentiable map ϕ is denoted by J_ϕ. We reserve u for unconstrained vectors (pre-projection) and use x or $\pi(z)$ for feasible vectors in $\mathcal{C}$.

2 Soft-Radial Projection Layer

Many models output a raw vector in $\mathbb{R}^n$ that must satisfy hard constraints at inference time—*e.g.*, neural networks and decoders, control policies, differentiable optimization layers, or amortized solvers. We place a *soft-radial projection layer* $p : \mathbb{R}^n \to \mathrm{Int}(\mathcal{C})$ downstream of such modules so that, regardless of how the raw output is produced (even if infeasible), the final decision lies in the feasible set. Concretely, we instantiate the unconstrained candidate g as a parameterized map $g_\theta : \mathcal{Z} \to \mathbb{R}^n$ (*e.g.*, a neural network). Given the raw output $u = g_\theta(z)$, we set the policy as $\pi_\theta = p \circ g_\theta$, which ensures $\pi_\theta(z) \in \mathrm{Int}(\mathcal{C})$ during training and deployment without penalty tuning or post-hoc repair. The learning objectives (1) and (2) are then instantiated by composing with p (*e.g.*, $\ell(\pi_\theta(z), y)$ or $c(z, \pi_\theta(z))$). This composition turns any base predictor into a constraint-satisfying model while keeping gradients usable for end-to-end training. The remainder of this section develops the geometry and regularity of p needed for gradient-based methods. We first formalize the soft-radial projection p and its ray-wise components. The construction relies on a smooth contraction toward an anchor point, mapping the entire ambient space into the interior of $\mathcal{C}$ while retaining differentiability almost everywhere.

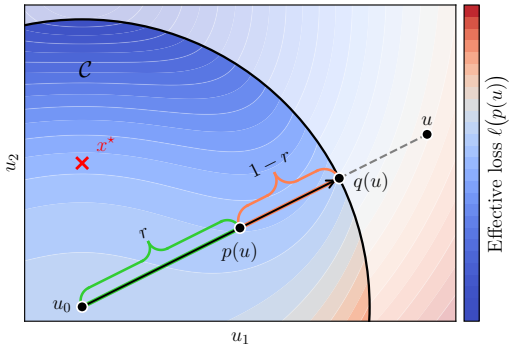


Figure 2: Geometric intuition. For an infeasible input u, the ray from anchor u_0 through u intersects the boundary $\partial\mathcal{C}$ at the hard radial projection $q(u)$. Our soft-radial mapping $p(u)$ strictly enforces feasibility by smoothly interpolating along the segment $[u_0, q(u)]$ via the radial weight r.

Definition 2.1 (Soft-Radial Projection)**.** Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a closed, convex set with nonempty interior, and fix an anchor point $u_0 \in \mathrm{Int}(\mathcal{C})$. Let $r : \mathbb{R}_+ \to [0, 1)$ be continuous and strictly increasing with $r(0) > 0$ and $\lim_{\rho\to\infty} r(\rho) = 1$.

For any $u \in \mathbb{R}^n$, the *soft-radial projection* of u onto $\mathcal{C}$ is

$$p(u) \;=\; u_0 \;+\; r\big(\|u - u_0\|^2\big)\,\big(q(u) - u_0\big),$$

where the *hard radial projection* $q(u)$ is defined by $q(u) = u$ if $u \in \mathcal{C}$, and $q(u) = u_0 + \alpha^\star(u)(u - u_0)$ otherwise, with the scaling factor $\alpha^\star(u) = \sup\{\alpha \in [0,1] : u_0 + \alpha(u - u_0) \in \mathcal{C}\} \in [0,1]$.

Remark 2.2. The mapping q projects u onto $\mathcal{C}$ along the ray emanating from the anchor u_0, in contrast to the usual Euclidean (orthogonal) projection. The soft-radial projection p then interpolates between u_0 and $q(u)$, with interpolation factor governed by $r(\|u - u_0\|^2)$ (see Fig. 2). If we drop the strict-interior requirement and allow $r \equiv 1$, then $p(u) = q(u)$ and the soft-radial projection takes the form of the hard radial projection.

Coordinate convention. The construction of p is invariant under translations of both the input and the constraint set: replacing u by $u - u_0$ and $\mathcal{C}$ by $\mathcal{C} - \{u_0\}$ leaves the construction unchanged up to translation. Hence, throughout the subsequent geometric and optimization analysis and the corresponding proofs, we work in translated coordinates and assume without loss of generality that $0 \in \mathrm{Int}(\mathcal{C})$ is the anchor. Under this convention, $q(u) = u$ for $u \in \mathcal{C}$, $q(u) = \alpha^\star(u)u$ otherwise, and $p(u) = r(\|u\|^2)q(u)$. The general-anchor form is recovered by translating back by u_0.

Regularity of components. We first record basic properties of the ray map q and the scalar $\alpha^\star$. These facts will be used later to build a raywise parametrization of $\mathrm{Int}(\mathcal{C})$ and to analyze $\ell \circ p$ with first-order methods. In particular, continuity and local Lipschitz continuity of q provide the regularity needed for the subsequent analysis of p.

Lemma 2.3. (Ray intersection and continuity). *Let $\mathcal{C} \subset \mathbb{R}^n$ be nonempty, closed, and convex with nonempty interior, and assume $0 \in \mathrm{Int}(\mathcal{C})$. For any $u \in \mathbb{R}^n$, the set $\{\alpha \in [0,1] : \alpha u \in \mathcal{C}\}$ is a nonempty closed interval $[0, \alpha^\star(u)]$ for a unique $\alpha^\star(u) \in [0,1]$, and $q(u) \in \mathcal{C}$. Moreover, $\alpha^\star$ is globally Lipschitz continuous and q is locally Lipschitz (and therefore differentiable almost everywhere by Rademacher's theorem; see, e.g., [17], Thm. 3.1.6).*

Structural properties of p. We now analyze the global geometry of p. A critical requirement is that p must preserve the *representational capacity* of the model g. Unlike orthogonal projections, which suffer from dimensional collapse by mapping the exterior space onto the boundary surface (a many-to-one mapping), our construction guarantees that p induces a *one-to-one parametrization* of $\mathrm{Int}(\mathcal{C})$ by strictly rescaling radii along rays emanating from the origin.

Assumption 2.4 (Radial monotonicity)**.** The function $r : \mathbb{R}_+ \to [0,1)$ is C^1, strictly increasing with $r'(\rho) > 0$ for all $\rho > 0$, and satisfies $r(0) > 0$ and $\lim_{\rho \to \infty} r(\rho) = 1$.

Assumption 2.4 ensures that the soft-radial projection is strictly monotonic along rays, providing the topological foundation for the one-to-one mapping. This prevents distinct points on the same ray from collapsing to the same interior location. Specifically, the scalar map $t \mapsto r(t^2)t$ has derivative $\frac{d}{dt}\big[r(t^2)t\big] = r(t^2) + 2t^2 r'(t^2)$. Since r and r' are positive, this derivative is strictly positive for all $t \geq 0$. Beyond the boundary in a direction v, the radial profile is $r(t^2)\bar{t}(v)$, whose derivative $2t\, r'(t^2)\bar{t}(v)$ is also strictly positive. Thus, p is strictly monotone along each ray and, since it preserves ray directions, is injective.

The mapping p admits a radial decomposition into angular and radial components, simplifying the analysis. Since the projection preserves the direction of the ray $\mathbb{R}_+ v$ (the angular component), its behavior is characterized by its action on the scalar distance from the origin—the radial component.

Definition 2.5 (Boundary distance)**.** Fix a unit direction $v \in \mathbb{R}^n$ ($\|v\| = 1$). We define the *boundary distance* $\bar{t}(v) \in (0, \infty]$ as the distance from the origin to the boundary of $\mathcal{C}$ along v, given by $\bar{t}(v) := \sup\{t \geq 0 : tv \in \mathcal{C}\}$.

With the boundary established, the radial profile of the projection is defined by the function ψ_v. This map distinguishes between the interior regime, where the ray remains within $\mathcal{C}$, and the exterior regime, where the input is scaled relative to the boundary.

Definition 2.6 (Scalar projection map)**.** For a fixed unit direction v, define the map $\psi_v : [0, \infty) \to [0, \infty)$ by $\psi_v(t) := r(t^2)\min\{t, \bar{t}(v)\}$.

By construction, the soft-radial projection satisfies $p(tv) = \psi_v(t)v$. With this decomposition, we now establish the regularity of these components. The following lemma verifies that the boundary distance varies continuously with direction and that the scalar projection is strictly monotonic—properties that are essential for proving the soft-radial projection is a global bijection.

Lemma 2.7. (Regularity of components). *The boundary distance map $\bar{t}$ is continuous on the unit sphere $\{v : \|v\| = 1\}$ (with values in $(0, \infty]$). Moreover, under Assumption 2.4, for any unit direction v, the scalar map ψ_v is continuous and strictly increasing, and its range is exactly $[0, \bar{t}(v))$.*

The raywise analysis in Lemma 2.7 provides the sufficient conditions to characterize the global topology of p. We now formally state the main result: p is a homeomorphism, acting as a reversible deformation of the entire ambient space $\mathbb{R}^n$ onto the feasible interior $\mathrm{Int}(\mathcal{C})$.

Theorem 2.8. (Homeomorphism). *Under Assumption 2.4, the soft-radial projection $p : \mathbb{R}^n \to \mathrm{Int}(\mathcal{C})$ is a homeomorphism.*

This theorem establishes a continuous, bijective correspondence between the latent output space $\mathbb{R}^n$ and the feasible interior $\mathrm{Int}(\mathcal{C})$, with a continuous inverse.

Differential regularity. Finally, to enable gradient-based learning, we establish differentiability of the soft-radial projection and full-rank Jacobians wherever differentiable.

Theorem 2.9. (Differentiability and invertibility). *Under Assumption 2.4, the soft-radial projection p is differentiable almost everywhere on $\mathbb{R}^n$. The Jacobian $J_p(u)$ is invertible at every point u where it exists.*

The invertibility of J_p ensures that the mapping has *full rank* almost everywhere. Consequently, soft-radial projection avoids the rank deficiency characteristic of orthogonal and hard radial projections.

3 Optimization Guarantees

After establishing the geometry of the projection layer, we now analyze its implications for *unconstrained optimization*. Consider a standard constrained task where we seek to minimize a loss $\ell(x)$ over the feasible set $\mathcal{C}$. By equipping the model with the soft-radial projection, we transform this problem into the unconstrained minimization of the composite objective $f(u) \coloneqq \ell(p(u))$, where $u \in \mathbb{R}^n$ denotes the unconstrained candidate vector, for instance the model output $u = g(z)$.

Rather than developing specialized optimization algorithms, our goal is to translate the geometric properties of p into first-order optimization properties of the composite objective. Similar unconstrained-training viewpoints appear in prior optimization-layer and enforcement-layer approaches [11, 13, 8, 16]. Our specific contribution is to characterize what the strict-interior soft-radial geometry implies for optimal value equivalence, interior stationarity, the value gap under bounded latent norms, and the inherent failure of global PL-type bounds.

Optimal value equivalence. Our first requirement is that the unconstrained formulation should preserve the value of the original constrained problem. Since p is a homeomorphism from $\mathbb{R}^n$ onto $\mathrm{Int}(\mathcal{C})$, the unconstrained parametrization covers exactly the feasible interior. Continuity of ℓ then ensures that optimizing over the interior has the same infimum as optimizing over the closed feasible set.

Theorem 3.1. (Equivalence of optimal values). *Let $\ell : \mathcal{C} \to \mathbb{R}$ be continuous. With $f \coloneqq \ell \circ p$, we have*

$$\inf_{u \in \mathbb{R}^n} f(u) \;=\; \inf_{x \in \mathrm{Int}(\mathcal{C})} \ell(x) \;=\; \inf_{x \in \mathcal{C}} \ell(x).$$

Moreover, $\arg\min f \neq \emptyset$ *if and only if* $\arg\min_{\mathcal{C}} \ell$ *intersects* $\mathrm{Int}(\mathcal{C})$. *In that case, $u^\star \in \arg\min f$ if and only if $p(u^\star) \in \arg\min_{\mathcal{C}} \ell$.*

First-order correspondence. Beyond agreement of optimal values, optimization algorithms require that first-order stationarity of the unconstrained objective $f = \ell \circ p$ be interpretable in terms of first-order conditions for the original constrained problem. Since $p(\mathbb{R}^n) \subset \mathrm{Int}(\mathcal{C})$, the correspondence we establish is necessarily an *interior* one: it relates stationary points of f to points $x \in \mathrm{Int}(\mathcal{C})$ satisfying $\nabla \ell(x) = 0$.

Proposition 3.2. (Correspondence of stationary points). *Assume ℓ is C^1 on $\mathrm{Int}(\mathcal{C})$. Wherever p is differentiable, with $f = \ell \circ p$ we have $\nabla f(u) = J_p(u)^\top \nabla \ell\big(p(u)\big)$.*

Since $J_p(u)$ is invertible almost everywhere by Theorem 2.9, the stationary conditions are equivalent at such points $\nabla f(u) = 0 \iff \nabla \ell\big(p(u)\big) = 0$. Thus, the full-rank Jacobian ensures that the

unconstrained landscape introduces no projection-induced spurious stationary points within the feasible interior wherever p is differentiable. If the constrained problem has no interior minimizer, any minimizing sequence for f must satisfy $\|u_k\| \to \infty$, since the soft-radial projection approaches boundary optima only asymptotically.

Value gap under bounded latent norm. As reaching a boundary minimizer strictly requires the unconstrained prediction to diverge, we must consider practical scenarios where network outputs are constrained by a latent norm budget M, for instance through weight decay, clipping, or architectural limits. This restricts the reachable objective values to $\ell_M^\star := \inf_{\|u\| \le M} f(u)$. Proposition 3.3 quantifies the resulting suboptimality gap relative to the true constrained minimum $\ell^\star := \inf_{x \in \mathcal{C}} \ell(x)$.

Proposition 3.3. (Value gap under a latent norm budget). *Assume the problem admits a boundary minimizer $x^\star \in \arg\min_{x \in \mathcal{C}} \ell(x) \cap \partial\mathcal{C}$ with $x^\star = \bar{t}(v^\star)\, v^\star$ for some unit vector $v^\star$, and assume that ℓ is globally L-Lipschitz on $\mathcal{C}$. Define the boundary deficit $\delta_M(v) := \bar{t}(v) - \psi_v(M)$. Then $0 \le \ell_M^\star - \ell^\star \le L\, \delta_M(v^\star)$. In particular, if $M \ge \bar{t}(v^\star)$, then $\delta_M(v^\star) = (1 - r(M^2))\, \bar{t}(v^\star)$, and hence $0 \le \ell_M^\star - \ell^\star \le L(1 - r(M^2))\, \bar{t}(v^\star)$.*

Proposition 3.3 makes the trade-off explicit: any mechanism ensuring $\sup_z \|g_\theta(z)\| \le M$ restricts the attainable decisions and induces at most $L\, \delta_M(v^\star)$ suboptimality. Once $M \ge \bar{t}(v^\star)$, the bias is controlled directly by the radial saturation error $1 - r(M^2)$. Conversely, if the constrained problem admits an interior minimizer $x^\circ \in \mathrm{Int}(\mathcal{C})$, then $u^\circ := p^{-1}(x^\circ)$ has finite norm, so any budget $M \ge \|u^\circ\|$ incurs no bias.

Global optimality for convex losses. When ℓ is convex, any *interior* stationary point is globally optimal. Since $p(\mathbb{R}^n) = \mathrm{Int}(\mathcal{C})$ and Proposition 3.2 establishes the equivalence of stationary conditions wherever J_p is invertible, stationary points of f at such locations correspond to global minimizers of ℓ over $\mathcal{C}$.

Corollary 3.4. (Global optimality of interior stationary points for convex losses). *Assume ℓ is convex and C^1 on $\mathrm{Int}(\mathcal{C})$, and let $f := \ell \circ p$. If p is differentiable at u and $\nabla f(u) = 0$, then $p(u)$ is a global minimizer of ℓ over $\mathcal{C}$.*

Remark (Anchor). At the anchor $u_0 \in \mathrm{Int}(\mathcal{C})$ we have $J_p(u_0) = r(0)I$, hence $\nabla f(u_0) = r(0)\, \nabla \ell(u_0)$. Thus, the condition $r(0) > 0$ in Assumption 2.4 prevents the anchor from becoming an artificial stationary point.

On PL-type conditions and bounded iterates. Because the soft-radial map saturates as $\|u\| \to \infty$, gradients can become arbitrarily small even when the objective value remains bounded away from its infimum. This rules out a global Polyak–Łojasiewicz (PL) inequality in general.

Lemma 3.5. (Absence of a global PL inequality in general). *Even if $\mathcal{C} := B_1(0)$, $\ell(x) := \|x\|^2$ and $u_0 := 0$, there exists no function r satisfying Assumption 2.4 such that $f = \ell \circ p$ satisfies a global PL inequality on $\mathbb{R}^n$.*

Although the global PL condition fails, this obstruction is global. In practice, optimization is confined to a bounded region, for example through weight decay, clipping, trust regions, or projected updates. On bounded subsets of its input space, p is locally Lipschitz and differentiable a.e. with a full-rank Jacobian wherever differentiable. Consequently, whenever the composite training objective $F(\theta) := \mathbb{E}_z\big[\ell(p(g_\theta(z)))\big]$ satisfies the usual compactness, smoothness, and integrability assumptions, it falls within standard stochastic first-order frameworks, yielding nonconvex SGD convergence in the smooth regime. In the nonsmooth regime, standard stochastic subgradient results guarantee convergence to Clarke-stationary points under local-Lipschitzness and standard stochastic-oracle assumptions [18, Cor. 5.9]. Importantly, smoothness of F is an assumption on the expected composite objective itself; it is not guaranteed merely because the input distribution is absolutely continuous, since the network outputs may still concentrate on nonsmooth regions of p. We defer the formal regularity assumptions, proof intuition, and precise statement of Proposition 3.6 to Appendix B.2.

Robustness to degeneracy. A potential concern is the degenerate scenario where the network output distribution concentrates on a nondifferentiable set of the soft-radial map. In this case, the analysis of convergence requires additional assumptions from nonsmooth optimization, as discussed above. Unlike orthogonal projections that collapse exterior regions onto lower-dimensional boundary faces, soft-radial projection maps strictly to the interior and maintains an invertible Jacobian wherever differentiable. This avoids the rank deficiency characteristic of orthogonal projections, while accounting for the possibility of nondifferentiable network outputs.

4 Constrained Universal Approximation

We next demonstrate that soft-radial projection preserves the expressivity of the base model. While constrained universality has been established for input-dependent affine layers [16] and probabilistic transformer constructions [19], our guarantee relies on a distinct topological mechanism: the global homeomorphism $p : \mathbb{R}^n \to \mathrm{Int}(\mathcal{C})$. Let $\mathcal{Z} \subset \mathbb{R}^d$ be compact and let $\mathcal{G} \coloneqq \{g_\theta : \mathcal{Z} \to \mathbb{R}^n\}$ be a universal approximator in $C(\mathcal{Z}, \mathbb{R}^n)$, such as sufficiently wide or deep neural networks [20, 21]. We show that the constrained class $p \circ \mathcal{G}$ can uniformly approximate any continuous target $h : \mathcal{Z} \to \mathcal{C}$. Since p maps strictly into the open set $\mathrm{Int}(\mathcal{C})$, the key point is to handle targets whose image intersects the boundary $\partial\mathcal{C}$.

Theorem 4.1. (Universal approximation on $\mathcal{C}$). *Let $\mathcal{Z} \subset \mathbb{R}^d$ be compact and let $\mathcal{G}$ be a universal approximator on $\mathcal{Z}$. For every continuous function $h : \mathcal{Z} \to \mathcal{C}$ and every $\varepsilon > 0$, there exists a function $g \in \mathcal{G}$ such that the constrained predictor $p \circ g$ satisfies $\sup_{z\in\mathcal{Z}} \| p(g(z)) - h(z) \| \leq \varepsilon$.*

Proof intuition. The argument proceeds by density and composition. Convexity of $\mathcal{C}$ ensures that strictly interior functions are dense in $C(\mathcal{Z}, \mathcal{C})$, so any target h admits a uniformly close interior approximation h_ε. Since p is a homeomorphism onto $\mathrm{Int}(\mathcal{C})$, there exists a continuous pre-image ϕ such that $p \circ \phi = h_\varepsilon$. The result follows by applying the universality of $\mathcal{G}$ to approximate ϕ.

Remark (Unbounded sets $\mathcal{C}$). Compactness of $\mathcal{Z}$ ensures that $h(\mathcal{Z})$ is bounded even if $\mathcal{C}$ is unbounded. Consequently, the relevant pre-images in $\mathbb{R}^n$ remain within a compact subset, where uniform approximation bounds apply.

Approximation stability. Uniform error bounds transfer to the constrained model up to a constant factor. Let $K \subset \mathbb{R}^n$ be compact and let L_K be a Lipschitz constant of p on K. For any continuous $\phi : \mathcal{Z} \to \mathbb{R}^n$ and any $g \in \mathcal{G}$ satisfying $\phi(\mathcal{Z}), g(\mathcal{Z}) \subset K$, we have $\sup_{z\in\mathcal{Z}} \|p(g(z)) - p(\phi(z))\| \leq L_K \sup_{z\in\mathcal{Z}} \|g(z) - \phi(z)\|$. Thus, composition with p preserves the expressivity of $\mathcal{G}$ and transfers uniform error bounds from $\mathcal{G}$ to the constrained class $p \circ \mathcal{G}$, up to the Lipschitz factor L_K.

5 Numerical Experiments

While Section 1 emphasized safety, strict constraints are equally critical in operations where violations are prohibitive. We evaluate our framework on formulation (2) via portfolio optimization and ride-sharing dispatch; see Appendix C for extended experimental details. For our implementation, we treat the projection as a differentiable layer, allowing us to train the model end-to-end by backpropagating gradients directly through the projection operation to the base network g_θ.[1]

5.1 Implementation Details

A critical computational advantage of the soft-radial projection is that the hard radial scaling factor, $\alpha^\star(u)$, often admits an efficient closed-form solution. This contrasts sharply with standard layers based on orthogonal projection ($\min_{v\in\mathcal{C}} \|u - v\|^2$). While orthogonal projection onto a single ball or simplex admits efficient direct algorithms, projection onto polyhedral sets formed by intersecting linear constraints generally requires solving a constrained quadratic program (QP) for every sample in the batch. These QPs typically lack closed-form solutions, necessitating iterative numerical solvers—such as Dykstra's algorithm for intersections of convex sets [22, 23] or ADMM for splitting complex composite constraints [24]—that significantly slow down the forward pass. For constraint families admitting closed-form radial scaling, our method circumvents this bottleneck by replacing the iterative QP with a direct, vectorized calculation.

Choice of radial contraction r. Recall from Definition 2.1 that the radial contraction r maps the squared distance $\rho \coloneqq \|u - u_0\|^2$ to a scaling factor in $[0, 1)$. We parameterize this contraction using smooth sigmoidal mappings satisfying Assumption 2.4. Fixing parameters $\varepsilon \in (0, 1)$ and $\lambda > 0$, we consider three families: (Rational) $r(\rho) = \varepsilon + (1 - \varepsilon)\rho/(\rho + \lambda)$, (Exponential) $r(\rho) = \varepsilon + (1 - \varepsilon)(1 - e^{-\rho/\lambda})$, and (Hyperbolic) $r(\rho) = \varepsilon + (1 - \varepsilon)\tanh(\rho/\lambda)$. While all three forms guarantee strict interior feasibility, they differ in their asymptotic saturation rate (see Appendix C.1). Inside $\mathcal{C}$, the radial derivative $r(\rho) + 2\rho\, r'(\rho)$ remains bounded away from zero, ensuring that the Jacobian retains full rank, including at the anchor u_0. Along rays with finite boundary distance, the radial derivative tends to zero as the unconstrained input moves farther beyond the feasible boundary.

[1] Code: `https://github.com/RAO-EPFL/SRP`

Computing the radial map. The evaluation of $p(u)$ reduces to computing the scaling factor $\alpha^\star(u)$ of the hard radial projection $q(u) = u_0 + \alpha^\star(u)(u - u_0)$. By construction, $\alpha^\star(u) = 1$ for feasible inputs, while for infeasible inputs $\alpha^\star(u)$ determines the intersection of the ray from u_0 through u with the boundary $\partial\mathcal{C}$.

Case 1: Polyhedra (Linear constraints). For sets defined by linear inequalities $Ax \leq b$ (*e.g.*, the simplex or capped simplex), the hard radial scaling factor is obtained exactly without iteration. Since u_0 is strictly interior ($b - Au_0 > 0$), we obtain

$$\alpha^\star(u) = \min\left\{1, \min_{i:\, a_i^\top(u-u_0)>0} \frac{b_i - a_i^\top u_0}{a_i^\top (u - u_0)}\right\},$$

where the inner minimum is $+\infty$ if the index set is empty. The computation is fully parallelizable and requires $\mathcal{O}(mn)$ operations for a dense constraint matrix $A \in \mathbb{R}^{m\times n}$.
Case 2: Euclidean balls. For spherical constraints $\|x - c\|_2 \leq R$, the positive root $t^\star$ of $\|(u_0 + t(u - u_0)) - c\|^2 = R^2$ admits an analytic solution, yielding $\alpha^\star(u) = \min\{1, t^\star\}$ for $u \neq u_0$, with $\alpha^\star(u_0) = 1$.
Case 3: General convex sets. For a general convex set $\{x \in \mathbb{R}^n : h(x) \leq 0\}$ with convex h, computing $\alpha^\star(u)$ reduces to a one-dimensional search along the ray, with $\phi(t) = h(u_0 + t(u - u_0))$. If $u \in \mathcal{C}$, then $\alpha^\star(u) = 1$. Otherwise, convexity ensures that the feasible values of $t \in [0, 1]$ form an interval containing zero, whose endpoint $\alpha^\star(u)$ can be determined by bisection.

Choice of the interior point u_0. Any $u_0 \in \mathrm{Int}(\mathcal{C})$ is sufficient for the construction, but conditioning depends on the margin $\delta := \mathrm{dist}(u_0, \partial\mathcal{C})$. Smaller margins weaken the Lipschitz bounds for the radial map, motivating the use of high-slack anchors in practice, such as barycenters for simplices, normalized cap vectors for capped simplices, Chebyshev centers for general polytopes, or analytic/barrier centers when available.[2]

Baselines. To provide insights with different neural network architectures, we consider for $g_\theta(z)$ a multi-layer perceptron (MLP) and a Long Short-Term Memory (LSTM) [25] architecture. To benchmark the efficacy of the soft-radial projection, we compare against state-of-the-art constraint enforcement layers ($\mathrm{Proj}(g_\theta(z))$) across four categories:

- *Softmax (Simplex):* For standard simplex constraints, we utilize the Softmax function with temperature scaling τ. While computationally cheap, it does not generalize to arbitrary convex sets.
- *Orthogonal projection:* We compare against the standard Euclidean projection. As discussed in Section 1, this method is idempotent—preserving the landscape for already-feasible points—but suffers from gradient saturation when inputs lie outside $\mathcal{C}$.
- *Gauge/radial projection:* We compare against *RAYEN* [8], a hard gauge-based layer that maps infeasible outputs along rays to $\partial\mathcal{C}$. Unlike SRP, it lacks a soft interior contraction and is many-to-one outside the feasible set.
- *Optimization-based layers:* We evaluate adapted versions of *Deep Constraint Completion and Correction* (DC3) [13], which enforces feasibility via gradient-based corrections, and *HardNet* [16], a recent architecture specialized for affine constraints.[3]

5.2 Portfolio Optimization

We consider a financial market with N assets over a finite horizon T. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space equipped with a filtration $(\mathcal{F}_t)_{t\geq 0}$. At each time step t, the agent observes signals z_t–comprising raw returns augmented with rolling volatility and correlation to the market proxy computed over lookback H—and selects a portfolio weight vector $w_t \in \mathcal{C} \subset \mathbb{R}^N$. Between steps $t-1$ and t, simple asset returns $y_t \in \mathbb{R}^N$ induce a passive drift, evolving weights to $w_{t-1}^+ = (\mathrm{diag}(\mathbb{1} + y_t)\, w_{t-1})/((\mathbb{1} + y_t)^\top w_{t-1})$. We account for the transaction cost incurred when selecting w_t at time t and the portfolio return realized over the subsequent period. The objective is to maximize the Sharpe ratio of the resulting net returns R_{t+1}^{net}

$$\max_\theta \ \frac{\mathbb{E}[R_{t+1}^{\text{net}}]}{\sqrt{\mathrm{Var}(R_{t+1}^{\text{net}})}} \quad \text{s.t.} \quad R_{t+1}^{\text{net}} = w_t^\top y_{t+1} - \frac{\gamma}{2}\|w_t - w_{t-1}^+\|_1, \ \ w_t = \mathrm{Proj}_{\mathcal{C}}(g_\theta(z_t)) \ \ \forall t. \quad (3)$$

[2] We isolate anchor sensitivity and boundary optima in Appendix C.3.
[3] App. C.2 describes the implementation of the respective models for the capped simplex.

Here, γ is the transaction cost parameter and the L_1 norm captures turnover costs.[4] We employ the *component-wise capped simplex*, which generalizes the standard simplex by enforcing asset-specific capacities $c \in (N^{-1}, 1]^N$ to control risk concentration $\mathcal{C} = \{w \in \mathbb{R}^N_+ \mid \mathbb{1}^\top w = 1,\ w \le c\}$.

Results analysis. Table 1 evaluates the net Sharpe ratio and turnover of an LSTM policy trained end-to-end under various constraint enforcement methods. On the standard simplex (Δ), the Softmax baseline achieves a Sharpe ratio of 0.65 without the additional per-asset capacity constraints. Under the capped simplex ($\mathcal{C}$), O-Proj achieves a Sharpe ratio of 0.44, while the adapted HardNet and DC3 baselines attain 0.46 and 0.25, respectively. Conversely, our soft-radial projection achieves the highest mean Sharpe ratio (0.70), with low turnover (0.04) and substantially lower inference cost than the correction-based methods. Notably, the hard gauge-based RAYEN matches our computational efficiency by design but yields a lower Sharpe ratio (0.60) with higher turnover. Overall, soft-radial projection combines the low latency of ray-based enforcement with the best observed solution quality on this task.

Table 1: Portfolio results for the *Global* dataset, aggregated over five seeds (mean $\pm$ std) with $N = 11$ assets, per-asset capacity $c = 0.25$, and transaction cost of 10 bps. DC3 and HardNet include evaluation-time feasibility correction; see App. C.2.

Set	Method	SR (Net)↑	Turnover↓	Training [s]↓	Inference [ms]↓
Δ	Softmax	0.65 (0.09)	0.11 (0.02)	11.16 (0.19)	0.34 (0.08)
$\mathcal{C}$	O-Proj	0.44 (0.13)	0.16 (0.02)	46.29 (0.97)	5.81 (0.19)
	DC3	0.25 (0.18)	0.23 (0.01)	23.52 (0.41)	8.44 (0.36)
	HardNet	0.46 (0.21)	0.16 (0.04)	17.27 (0.92)	6.87 (0.10)
	RAYEN	0.60 (0.06)	0.11 (0.01)	14.09 (0.25)	0.68 (0.01)
	SRP	0.70 (0.06)	0.04 (0.02)	14.73 (0.14)	0.76 (0.01)

5.3 Ride-Sharing Dispatch

To demonstrate generalizability, we apply our framework to ride-sharing dispatch with a *time-varying budget*. Unlike the fixed portfolio sum, the available fleet size $S_t \in \mathbb{N}_+$ fluctuates over time, causing the feasible set $\mathcal{C}_t$ to expand and contract. In practice, such constraints enforce per-zone capacity limits (*e.g.*, congestion or emissions control) [26, 27]. At each time step t, the agent observes a demand history window z_t (augmented with sine/cosine encodings for day and week) and a current supply proxy S_t. We seek an allocation $a_t \in \mathbb{R}^N_+$ over N zones to maximize the *served rate* for the upcoming interval. We optimize the empirical average:

$$\max_\theta \ \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\sum_{i=1}^N \min_\tau(a_{i,t}, d_{i,t})}{\sum_{i=1}^N d_{i,t}}\right] \qquad \text{s.t.} \qquad a_t = \mathrm{Proj}_{\mathcal{C}_t}(g_\theta(z_t)) \ \ \forall t. \tag{4}$$

where d_t is the realized demand in the interval following the observation window. We replace the hard minimum with the smooth SoftMin, $\min_\tau(x, y) = -\tau \log(e^{-x/\tau} + e^{-y/\tau})$, to maintain non-zero gradients even when supply exceeds demand ($a_{i,t} > d_{i,t}$). The constraint set is a scaled capped simplex $\mathcal{C}_t = \{a \in \mathbb{R}^N_+ \mid \mathbb{1}^\top a = S_t,\ a \le \kappa S_t\}$. We implement Proj via a *scaling transformation* (homothety): the network output is normalized by S_t, projected onto the canonical unit capped simplex, and subsequently rescaled by S_t. This effectively maps the dynamic constraint geometry to a fixed reference polytope.

Results analysis. Table 2 evaluates the served rate (demand fulfillment) of an MLP policy. Unlike the highly stochastic portfolio domain, the ride-sharing environment exhibits more regular demand patterns, allowing the unconstrained Softmax baseline to achieve a served rate of 0.84. Under dynamic capacity constraints ($\mathcal{C}_t$), DC3, HardNet,

Table 2: Ride-sharing dispatch results, aggregated over five random seeds (mean $\pm$ std), for the top $N = 150$ activity zones and zone capacity $\kappa = 0.1$; see App. C.6.

Set	Method	Served Rate↑	Training [s]↓	Inference [ms]↓
Δ_t	Softmax	0.84 (0.00)	12.53 (0.72)	0.34 (0.06)
$\mathcal{C}_t$	O-Proj	0.72 (0.01)	80.25 (4.11)	10.65 (0.33)
	DC3	0.84 (0.00)	61.39 (1.30)	17.15 (2.04)
	HardNet	0.84 (0.00)	44.88 (0.69)	13.57 (0.74)
	RAYEN	0.84 (0.00)	17.92 (0.65)	0.79 (0.07)
	SRP	0.84 (0.00)	15.52 (0.30)	0.89 (0.01)

[4] During training, we approximate the L_1 norm using the smooth Pseudo-Huber loss; evaluation uses exact turnover.

RAYEN, and our soft-radial projection all match this performance (0.84). This demonstrates that soft-radial projection achieves the best observed solution quality while enforcing feasibility at a fraction of the computational cost of correction-based methods. Conversely, orthogonal projection underperforms (0.72), consistent with the loss of gradient information along directions suppressed by active constraints. Finally, while hard gauge-based methods such as RAYEN perform equally well in this structured setting, our soft-radial mapping additionally guarantees injectivity and a full-rank Jacobian almost everywhere, avoiding the many-to-one collapse characteristic of boundary-valued radial projections.

6 Conclusion

This paper introduced *Soft-Radial Projection* to address the gradient saturation associated with boundary-valued projection layers. By constructing a radial homeomorphism to the interior of a convex set, we guarantee strict feasibility and preserve model expressivity (Theorem 4.1) without relying on iterative optimization for common constraint families. Empirically, the method achieves competitive solution quality at low computational cost, with particularly strong results in portfolio optimization. These findings establish Soft-Radial Projection as an efficient alternative to conventional projection layers for enforcing hard constraints in end-to-end learning.

Limitations. While Soft-Radial Projection applies to general convex feasible sets, input-dependent constraints introduce additional computational overhead when the interior anchor and hard radial projection must be computed for each instance. Furthermore, boundary optima are approached asymptotically, introducing a trade-off between approximation accuracy and gradient conditioning. Since Soft-Radial Projection is not idempotent, the method is designed for end-to-end training rather than post-hoc correction of pretrained predictions. Developing efficient implementations for input-dependent constraints and further understanding this trade-off offer promising directions for future research.

Acknowledgments

We thank Yves Rychener for his valuable insights and contributions to the initial conceptualization of this work. This work was supported as part of NCCR Automation, a National Centre of Competence in Research, funded by the Swiss National Science Foundation (grant number 51NF40_225155).

References

[1] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):411–444, 2022.

[2] Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, pages 5490–5500. Curran Associates, Inc., 2017.

[3] Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022.

[4] Yves Rychener, Daniel Kuhn, and Tobias Sutter. End-to-end learning for stochastic optimization: A Bayesian perspective. In *International Conference on Machine Learning*, pages 29455–29472, 2023.

[5] Abhilash Reddy Chenreddy and Erick Delage. End-to-end conditional robust optimization. In *Uncertainty in Artificial Intelligence*, pages 736–748, 2024.

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

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

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

[9] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(42):1437–1480, 2015.

[10] James Kotary, Ferdinando Fioretto, Pascal Van Hentenryck, and Bryan Wilder. End-to-end constrained optimization learning: A survey. In *International Joint Conference on Artificial Intelligence*, pages 4475–4482, 2021. Survey Track.

[11] Brandon Amos and J Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR, 2017.

[12] A. Agrawal, B. Amos, S. Barratt, S. Boyd, S. Diamond, and Z. Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems*, pages 9562–9574. Curran Associates, Inc., 2019.

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

[14] Andrei V Konstantinov and Lev V Utkin. A new computationally simple approach for implementing neural networks with output hard constraints. In *Doklady Mathematics*, volume 108, pages S233–S241. Springer, 2023.

[15] 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.

[16] Youngjae Min and Navid Azizan. HardNet: Hard-constrained neural networks with universal approximation guarantees. *arXiv preprint arXiv:2410.10807*, 2024.

[17] Herbert Federer. *Geometric Measure Theory*, volume 153 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, Germany, 1969.

[18] Damek Davis, Dmitriy Drusvyatskiy, Sham Kakade, and Jason D Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20(1): 119–154, 2020.

[19] Anastasis Kratsios, Behnoosh Zamanlooy, Tianlin Liu, and Ivan Dokmanić. Universal approximation under constraints is possible with transformers. In *International Conference on Learning Representations*, 2022.

[20] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[21] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

[22] Richard L Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983.

[23] James P. Boyle and Richard L. Dykstra. A method for finding projections onto the intersection of convex sets in Hilbert spaces. In Richard Dykstra, Tim Robertson, and Farroll T. Wright, editors, *Advances in Order Restricted Statistical Inference*, pages 28–47, New York, NY, 1986. Springer New York.

[24] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2010.

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997.

[26] Javier Alonso-Mora, Samitha Samaranayake, Alex Wallar, Emilio Frazzoli, and Daniela Rus. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, 114(3):462–467, 2017.

[27] Gioele Zardini, Nicolas Lanzetti, Marco Pavone, and Emilio Frazzoli. Analysis and control of autonomous mobility-on-demand systems. *Annual Review of Control, Robotics, and Autonomous Systems*, 5(1):633–658, 2022.

[28] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer, 2009.

[29] 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 *International Conference on Learning Representations*, 2026.

[30] Gonzalo E Constante-Flores, Hao Chen, and Can Li. Enforcing Hard Linear Constraints in Deep Learning Models with Decision Rules. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025.

[31] Ke Deng, Hanwen Zhang, Jin Lu, and Haijian Sun. HoP: Homeomorphic polar learning for hard constrained optimization. *arXiv preprint arXiv:2502.00304*, 2025.

[32] Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Imposing hard constraints on deep networks: Promises and limitations. *arXiv preprint arXiv:1706.02025*, 2017. CVPR 2017 Workshop on Negative Results in Computer Vision.

[33] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[34] L.E.J. Brouwer. Über Abbildung von Mannigfaltigkeiten. *Mathematische Annalen*, 71:97–115, 1912.

[35] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[36] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[37] Frank H. Clarke. *Optimization and Nonsmooth Analysis*, volume 5 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 1990.

[38] Zihao Zhang, Stefan Zohren, and Stephen Roberts. Deep learning for portfolio optimization. *Journal of Financial Data Science*, 2(4):8–20, 2020.

[39] Harry Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.

[40] Richard O Michaud. The Markowitz optimization enigma: Is 'optimized' optimal? *Financial Analysts Journal*, 45(1):31–42, 1989.

[41] Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106 (30):12267–12272, 2009.

[42] Mark HA Davis and Andrew R Norman. Portfolio selection with transaction costs. *Mathematics of Operations Research*, 15(4):676–713, 1990.

[43] Hongtai Zeng, Chao Yang, Yanzhen Zhou, Cheng Yang, and Qinglai Guo. GLinSAT: The general linear satisfiability neural network layer by accelerated gradient descent. *Advances in Neural Information Processing Systems*, 37:122584–122615, 2024.

[44] Haoyu Zhu, Yao Zhang, Jiashen Ren, and Qingchun Hou. T-SKM-Net: Trainable neural network framework for linear constraint satisfaction via sampling Kaczmarz-Motzkin method. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(17):14431–14439, 2026.

[45] Sogol Babaeinejadsarookolaee, Adam Birchfield, Richard D Christie, Carleton Coffrin, Christopher DeMarco, Ruisheng Diao, Michael Ferris, Stephane Fliscounakis, Scott Greene, Renke Huang, et al. The power grid library for benchmarking AC optimal power flow algorithms. *arXiv preprint arXiv:1908.02788*, 2019.

A Related Work

Enforcing hard constraints in neural networks is a fundamental challenge in safety-critical learning, requiring architectures that guarantee output feasibility without compromising optimization stability. Optimization over the space of constrained functions is generally intractable. However, we can simplify the task-driven formulation (2) by leveraging the *Interchangeability Principle* [28, Theorem 14.60]. Applied to this setting, the principle establishes that minimizing the expected cost is equivalent to minimizing the cost pointwise for almost every input z, formally $\inf_{\pi \in \Pi,\, \pi(z) \in \mathcal{C}} \mathbb{E}_{z \sim \mathbb{P}_Z}[c(z, \pi(z))] = \mathbb{E}_{z \sim \mathbb{P}_Z}\left[\inf_{x \in \mathcal{C}} c(z, x)\right]$. This equivalence motivates solving feasibility pointwise by reparameterizing the decision policy: we optimize an unconstrained candidate function $g : \mathcal{Z} \to \mathbb{R}^n$ composed with a fixed projection operator, such that $\pi(z) = \mathrm{Proj}(g(z))$.

The literature can be categorized by the mechanism of enforcement—ranging from soft relaxations and iterative solvers to rigorous geometric constructions. Table 3 summarizes the properties of representative state-of-the-art approaches.

Broadly, constraint layers are distinguished by their dependency on the input context. *Input-independent constraints* involve a feasible set $\mathcal{C}$ that is fixed across all samples (*e.g.*, static actuator limits or portfolio weights), whereas *input-dependent constraints* involve a dynamic geometry $\mathcal{C}(z)$ conditioned on the input. Our work primarily targets the former, establishing a rigorous geometric framework for fixed convex sets. While strict constraint satisfaction is essential, the quality of the learning signal—specifically, the ability of backpropagation to provide informative gradients in high-noise regimes—remains under-explored. Existing approaches prioritize inference speed and feasibility during training and execution. However, convergence behavior toward the true solution and the ultimate optimality gap are less understood, particularly when the enforcement layer significantly alters the unconstrained network output. While this review is not exhaustive—with ongoing developments ranging from accelerated orthogonal projections [29] to the use of decision rules for constraint enforcement [30]—different applications and noise profiles often demand alternatives to standard orthogonal projection to maintain superior optimization properties.

Table 3: **Comparison with state-of-the-art constraint layers.** We position Soft-Radial Projection among constructive geometric layers that guarantee feasibility without high-dimensional iterative solvers.

Method	Constraint Scope	Input Dep.	Satisfaction	Computation	Gradient Landscape	Univ. Approx.
Soft-Penalty	Any	Yes	No (Penalty)	Closed-Form	Smooth (infeasible)	Yes
DC3 [13]	General	Yes	Approx.*	Iterative (GD)	Step-size sensitive	Unknown
RAYEN [8]	Convex (Lin/QC/SOC/LMI)	Limited	Strict	Closed-Form	Clipped / Flat	Unknown
HardNet-Aff [16]	Affine	Yes	Strict	Closed-Form	Saturated (Boundary)	Yes
HP [15]	Ball-Homeomorphic	Yes	Strict	Bisection	Iterative search	Yes
HoP (Polar) [31]	Star-Convex	Yes	Strict (Int.)	Closed-Form†	Polar stagnation	Unknown
RP [14]	Convex (Lin/Quad)	Limited	Strict (Int.)	Closed-Form	Smooth (Saturating)	Unknown
SRP (Ours)	**Convex**	**No**‡	**Strict (Int.)**	**C.-F. / 1D Root**	**Full-rank Jacobian a.e.**	**Yes**

*Asymptotic (linear constraints); heuristic/approximate otherwise; truncated correction in practice.

†HoP requires a per-sample interior anchor and a ray–boundary intersection query.

‡Current focus is on fixed sets $\mathcal{C}$, though formulation supports dynamic anchors.

A.1 Approximation and Optimization Approaches

Soft penalties and activations. The most straightforward approach, favored by many industry practitioners, relaxes strict constraints into the training objective. Penalty-based methods augment the loss function with a regularization term proportional to the magnitude of constraint violations [32]. While this preserves the smoothness of the optimization landscape, it fundamentally fails to guarantee feasibility during inference, particularly in high-dimensional spaces where the volume of the feasible set relative to the ambient space can become negligible. Similarly, standard activation functions (*e.g.*, Sigmoid, Softmax) offer computationally efficient bounds for elementary box or simplex constraints but lack the topological expressivity to model complex coupled constraints or polyhedral boundaries.

Differentiable optimization and iterative correction. To ensure strict satisfaction, a second family of methods embeds optimization solvers directly into the network architecture. Differentiable optimization layers, such as OptNet [11] and CVXPYLayers [12], compute the orthogonal projection of a prediction onto $\mathcal{C}$ by solving a convex program during the forward pass. DC3 [13] instead utilizes a

completion–correction scheme, where equality constraints are satisfied via variable completion and inequality violations are reduced through iterative gradient-based correction. While mathematically general, these approaches introduce significant computational overhead. Furthermore, backpropagating through these layers typically relies on implicit differentiation, which assumes the inner solver has reached a stationary point. When truncated for real-time efficiency (*e.g.*, fixed-iteration limits), this assumption is violated, leading to conflict-prone gradient updates where the descent direction of the task loss contradicts the incomplete constraint correction.

A.2 Constructive Geometric Layers

To circumvent the computational cost of iterative solvers, recent research focuses on "Feasibility by Construction" via closed-form surjective mappings $\mathrm{Proj} : \mathbb{R}^n \to \mathcal{C}$. We classify these geometric approaches into three distinct families:

Gauge and boundary projections. RAYEN [8] enforces convex constraints by selecting a fixed interior point and *rescaling* a predicted direction so that the endpoint remains feasible; the mapping is identity when the step stays inside the set, and otherwise returns the first ray–boundary intersection. HardNet-Aff [16] targets *input-dependent affine* inequalities via a closed-form pseudoinverse-based correction that clamps violated constraints to their bounds while moving in directions parallel to already-satisfied constraint boundaries. While both layers are computationally efficient, their piecewise ray/clamping structure can suppress learning signals when constraints are active (*e.g.*, RAYEN becomes insensitive to step magnitude after rescaling, and HardNet-Aff can attenuate gradient components in constraint-normal directions). This suppression can bias training toward boundary-feasible outputs and reduce interior exploration.

Topological diffeomorphisms. To utilize the set's interior, topological methods reshape the output space. Homeomorphic Projection (HP) [15] learns a bi-Lipschitz *invertible neural network* (INN) that approximates a minimum-distortion homeomorphism between the constraint set and a unit ball. At inference, HP maps an infeasible prediction into the ball via the INN inverse and then performs a k-step bisection along the ray to the ball center, requiring repeated INN forward evaluations and feasibility checks; the resulting boundary-feasible point trades computation for accuracy through the bisection depth and depends on the learned mapping distortion. Polar coordinate methods like HoP [31] attempt an analytic solution but introduce polar stagnation, singularities at the origin where angular components are undefined, resulting in numerical instability near the geometric center.

Radial interior mappings. Closely related to our work, Konstantinov and Utkin [14] propose a "ray-projection" layer that constructs feasible points by scaling a direction vector r from an interior anchor. While effective, this requires a specialized architecture that learns separate polar coordinates $(g(r, s))$ rather than a direct map of the ambient Euclidean space. Furthermore, these constructions typically do not furnish a global bijection between $\mathbb{R}^n$ and $\mathrm{Int}(\mathcal{C})$, potentially resulting in saturating radial behavior due to the choice of bounded scaling functions (*e.g.*, sigmoids).

Our contribution: Soft-radial projection. We propose a drop-in reparameterization layer for fixed closed convex sets with a non-empty interior. Our mapping acts directly on the unconstrained ambient prediction $u \in \mathbb{R}^n$ as a unified radial homeomorphism $\mathbb{R}^n \to \mathrm{Int}(\mathcal{C})$. For standard convex geometries—such as balls, ellipsoids, and boxes—the mapping admits a closed-form solution. For general convex sets, we utilize an efficient monotonic 1D root-finding procedure. Crucially, we ensure the mapping is differentiable with an invertible Jacobian almost everywhere, preventing the rank-collapse characteristic of boundary projections. This avoids exact projection-induced rank collapse, although the radial derivative diminishes as the unconstrained input moves farther beyond the feasible boundary. The construction also preserves the universal approximation property of the base network (Theorem 4.1).

B Proofs

In the following sections, we provide complete proofs for the theoretical results. To facilitate readability, we restate each claim prior to its proof.

B.1 Proofs of Section 2

We begin by establishing the regularity of the individual projection components.

Lemma 2.3. (Ray intersection and continuity). *Let $\mathcal{C} \subset \mathbb{R}^n$ be nonempty, closed, and convex with nonempty interior, and assume $0 \in \operatorname{Int}(\mathcal{C})$. For any $u \in \mathbb{R}^n$, the set $\{\alpha \in [0,1] : \alpha u \in \mathcal{C}\}$ is a nonempty closed interval $[0, \alpha^\star(u)]$ for a unique $\alpha^\star(u) \in [0,1]$, and $q(u) \in \mathcal{C}$. Moreover, $\alpha^\star$ is globally Lipschitz continuous and q is locally Lipschitz (and therefore differentiable almost everywhere by Rademacher's theorem; see, e.g., [17], Thm. 3.1.6).*

Proof. Ray feasibility and intersection. Fix $u \in \mathbb{R}^n$ and consider scalars $\alpha \in [0,1]$ such that $\alpha u \in \mathcal{C}$. The map $\alpha \mapsto \alpha u$ is continuous and $\mathcal{C}$ is closed, so the set of feasible $\alpha \in [0,1]$ is closed in $[0,1]$ and nonempty (since $\alpha = 0$ is always feasible). Convexity of $\mathcal{C}$ implies this set is convex in $\mathbb{R}$, and any nonempty closed convex subset of $\mathbb{R}$ is an interval; hence it is of the form $[0, \alpha^\star(u)]$ for a unique $\alpha^\star(u) \in [0,1]$.

Global Lipschitz continuity of α. Let $\gamma_{\mathcal{C}} : \mathbb{R}^n \to [0,\infty)$ denote the Minkowski functional (gauge) of $\mathcal{C}$,

$$\gamma_{\mathcal{C}}(x) \coloneqq \inf\{t > 0 : x \in t\mathcal{C}\}.$$

Standard properties of the gauge (see, *e.g.*, Rockafellar [33], Thm. 14.5 and Chap. 13) imply that $\gamma_{\mathcal{C}}$ is sublinear (convex and positively homogeneous) and finite on $\mathbb{R}^n$ when $0 \in \operatorname{Int}(\mathcal{C})$. Because $0 \in \operatorname{Int}(\mathcal{C})$, there exists $\delta > 0$ with $B(0,\delta) \subset \mathcal{C}$, which yields $\gamma_{\mathcal{C}}(x) \le \|x\|/\delta$ for all x and, by subadditivity of $\gamma_{\mathcal{C}}$,

$$\gamma_{\mathcal{C}}(x) - \gamma_{\mathcal{C}}(y) \le \gamma_{\mathcal{C}}(x-y) \le \frac{\|x-y\|}{\delta} \quad \forall\, x, y \in \mathbb{R}^n,$$

and

$$\gamma_{\mathcal{C}}(y) - \gamma_{\mathcal{C}}(x) \le \gamma_{\mathcal{C}}(y-x) \le \frac{\|y-x\|}{\delta} \quad \forall\, x, y \in \mathbb{R}^n.$$

Taking the maximum of the two inequalities shows

$$|\gamma_{\mathcal{C}}(x) - \gamma_{\mathcal{C}}(y)| \le \frac{\|x-y\|}{\delta} \quad \forall\, x, y \in \mathbb{R}^n,$$

so $\gamma_{\mathcal{C}}$ is Lipschitz with constant $1/\delta$.

For $u \neq 0$, the feasibility of $\alpha u \in \mathcal{C}$ is equivalent to $\alpha \le 1/\gamma_{\mathcal{C}}(u)$. Therefore, the feasible $\alpha \in [0,1]$ form the interval $[0, \min\{1, 1/\gamma_{\mathcal{C}}(u)\}]$, implying that $\alpha^\star(u) = \min\big(1, 1/\gamma_{\mathcal{C}}(u)\big)$, for $u \neq 0$, while for $u = 0$ we clearly have $\alpha^\star(u) = 1$. Define a helper function $\psi : [0,\infty) \to (0,1]$ by $\psi(t) \coloneqq \min(1, 1/t)$ for $t > 0$ and $\psi(0) \coloneqq 1$. The function ψ is 1–Lipschitz on $[0,\infty)$, so for all $u, v \in \mathbb{R}^n$,

$$\begin{aligned} |\alpha^\star(u) - \alpha^\star(v)| &= |\psi(\gamma_{\mathcal{C}}(u)) - \psi(\gamma_{\mathcal{C}}(v))| \\ &\le |\gamma_{\mathcal{C}}(u) - \gamma_{\mathcal{C}}(v)| \le \frac{\|u-v\|}{\delta}. \end{aligned}$$

Thus $\alpha^\star$ is globally Lipschitz with constant $1/\delta$.

Local Lipschitz continuity of q. We now bound the variation of $q(u) = \alpha^\star(u)u$ on bounded sets. Let $R > 0$ and assume $\|v\| \le R$ and $\|u\| \le R$. Then

$$\begin{aligned} \|q(u) - q(v)\| &= \|\alpha^\star(u)u - \alpha^\star(v)v\| \\ &\le \|\alpha^\star(u)(u-v)\| + \|v(\alpha^\star(u) - \alpha^\star(v))\| \\ &\le \|u-v\| + \|v\| \cdot |\alpha^\star(u) - \alpha^\star(v)| \\ &\le \|u-v\| + \frac{\|v\|}{\delta}\,\|u-v\| \\ &\le \Big(1 + \frac{R}{\delta}\Big)\|u-v\|. \end{aligned}$$

Hence q is $(1 + R/\delta)$–Lipschitz on the ball $\{u : \|u\| \leq R\}$, which implies that q is locally Lipschitz on $\mathbb{R}^n$, completing the proof. □

Next, we analyze the structural properties of the constructed projection map.

Lemma 2.7. (Regularity of components). *The boundary distance map $\bar{t}$ is continuous on the unit sphere $\{v : \|v\| = 1\}$ (with values in $(0, \infty]$). Moreover, under Assumption 2.4, for any unit direction v, the scalar map ψ_v is continuous and strictly increasing, and its range is exactly $[0, \bar{t}(v))$.*

Proof. Continuity of $\bar{t}$. We relate $\bar{t}(v)$ to the Minkowski functional (gauge) $\gamma_{\mathcal{C}}$. By definition, $\bar{t}(v) = 1/\gamma_{\mathcal{C}}(v)$ (with the convention $1/0 = \infty$). Since $\gamma_{\mathcal{C}}$ is continuous on $\mathbb{R}^n$ (Lemma 2.3) and the inversion map is continuous on $[0, \infty)$ into $(0, \infty]$, the composition $v \mapsto \bar{t}(v)$ is continuous on the unit sphere.

Properties of ψ_v. For $t \leq \bar{t}(v)$, we have $\psi_v(t) = r(t^2)t$. Strict monotonicity on this interval was established immediately following Assumption 2.4. For $t > \bar{t}(v)$ (which implies $\bar{t}(v) < \infty$), we have $\psi_v(t) = r(t^2)\bar{t}(v)$, which is strictly increasing simply because r is increasing. Continuity holds at $t = \bar{t}(v)$ as both expressions coincide. Finally, since $\psi_v(0) = 0$ and $\lim_{t\to\infty} \psi_v(t) = \bar{t}(v)$, the range is exactly $[0, \bar{t}(v))$. □

Building on these established properties, we prove that p is indeed a homeomorphism.

Theorem 2.8. (Homeomorphism). *Under Assumption 2.4, the soft-radial projection $p : \mathbb{R}^n \to \mathrm{Int}(\mathcal{C})$ is a homeomorphism.*

Proof. We first establish that p is a bijection.

Injectivity. Let $u_1, u_2 \in \mathbb{R}^n$ with $p(u_1) = p(u_2)$. Since p preserves directions (*i.e.*, $p(u) \in \mathbb{R}_+ u$), u_1 and u_2 must lie on the same ray, so $u_i = t_i v$ for some unit vector v and $t_i \geq 0$. The scalar map ψ_v from Definition 2.6 satisfies $p(t_i v) = \psi_v(t_i)v$. By Lemma 2.7, ψ_v is strictly increasing, so $\psi_v(t_1) = \psi_v(t_2)$ implies $t_1 = t_2$, and hence $u_1 = u_2$.

Surjectivity. Let $x \in \mathrm{Int}(\mathcal{C})$. If $x = 0$, $x = p(0)$. If $x \neq 0$, set $v = x/\|x\|$ and $s = \|x\|$. Since $x \in \mathrm{Int}(\mathcal{C})$, we strictly have $s < \bar{t}(v)$. By Lemma 2.7, the continuous scalar map ψ_v has range $[0, \bar{t}(v))$. By the Intermediate Value Theorem, there exists a unique t such that $\psi_v(t) = s$, yielding $p(tv) = x$. Moreover, p is continuous on $\mathbb{R}^n$ as it is composed of the continuous scaling function r (Assumption 2.4) and the continuous hard radial projection q (Lemma 2.3). Thus, p is a continuous bijection from $\mathbb{R}^n$ onto $\mathrm{Int}(\mathcal{C})$.

Continuity of Inverse. Since p is a continuous injection from $\mathbb{R}^n$ into an open subset of $\mathbb{R}^n$, the Invariance of Domain theorem [34] guarantees that p is an open map. Because p is a continuous bijection that maps open sets to open sets, it is a homeomorphism between $\mathbb{R}^n$ and its image $\mathrm{Int}(\mathcal{C})$. □

Finally, we analyze the differentiability of p, establishing that it is a diffeomorphism almost everywhere.

Theorem 2.9. (Differentiability and invertibility). *Under Assumption 2.4, the soft-radial projection p is differentiable almost everywhere on $\mathbb{R}^n$. The Jacobian $J_p(u)$ is invertible at every point u where it exists.*

Proof. By Lemma 2.3, q is locally Lipschitz and therefore differentiable almost everywhere by Rademacher's theorem. Since r is C^1 and strictly positive, the identities $p(u) = r(\|u\|^2)q(u)$ and $q(u) = p(u)/r(\|u\|^2)$ show that p and q have the same domain of differentiability. Whenever $J_p(u)$ exists, the product rule gives

$$J_p(u) \;=\; r(\|u\|^2)\, J_q(u) \;+\; 2\, r'(\|u\|^2)\, q(u)\, u^\top. \tag{5}$$

At a boundary point $u = t_0 v$ with $\|v\| = 1$ and $t_0 = \bar{t}(v)$, the left and right derivatives of the radial profile ψ_v differ by $r(t_0^2) > 0$. Hence p is not differentiable on $\partial\mathcal{C}$. We now examine the Jacobian in the interior and exterior of the feasible set.

Interior points. For $u \in \operatorname{Int}(\mathcal{C})$ we have $q(u) = u$ and $J_q(u) = I$, hence

$$J_p(u) = r(\|u\|^2)\, I + 2\, r'(\|u\|^2)\, uu^\top.$$

For $u \neq 0$, this matrix has eigenvector $u/\|u\|$, with corresponding eigenvalue $r(\|u\|^2) + 2\,\|u\|^2\, r'(\|u\|^2)$ and all vectors orthogonal to u are eigenvectors with eigenvalue $r(\|u\|^2)$. Assumption 2.4 gives $r(\rho) > 0$ and $r(\rho) + 2\rho\, r'(\rho) > 0$ for all $\rho \geq 0$, so $J_p(u) \succ 0$ is invertible for all $u \in \operatorname{Int}(\mathcal{C}) \setminus \{0\}$. At the origin, q coincides with the identity in a neighborhood of 0 (because $0 \in \operatorname{Int}(\mathcal{C})$), so $J_q(0) = I$ and the rank-one term vanishes, giving $J_p(0) = r(0)\, I$, which is invertible since $r(0) > 0$.

Exterior points. For $u \notin \mathcal{C}$, the gauge $\gamma_{\mathcal{C}}(u) > 1$ and $q(u) = u/\gamma_{\mathcal{C}}(u) \in \partial\mathcal{C}$. Fix such a point where p is differentiable. Then q is differentiable, and the identity $\gamma_{\mathcal{C}}(u) = \|u\|/\|q(u)\|$, valid throughout the exterior, implies that the gauge is differentiable there as well. Write $\lambda \coloneqq \gamma_{\mathcal{C}}(u) > 1$. Differentiating the exterior expression for q yields

$$J_q(u) \;=\; \frac{1}{\lambda} I - \frac{1}{\lambda^2}\, u\, \nabla\gamma_{\mathcal{C}}(u)^\top.$$

We establish that $J_p(u)$ is invertible by showing its kernel is trivial. Let $h \in \mathbb{R}^n$ satisfy $J_p(u)h = 0$. We decompose h with respect to the subspace $W = \{w \in \mathbb{R}^n \mid u^\top w = 0\}$. Any h can be uniquely written as $h = h_W + \beta u$, where $h_W \in W$ and $\beta \in \mathbb{R}$.

First, we compute the product of $J_q(u)$ with the radial component. Since $\gamma_{\mathcal{C}}$ is positively homogeneous of degree 1 by definition, Euler's homogeneous function theorem implies $\nabla\gamma_{\mathcal{C}}(u)^\top u = \gamma_{\mathcal{C}}(u) = \lambda$. Using this relation, we find

$$J_q(u)u \;=\; \frac{1}{\lambda}u - \frac{1}{\lambda^2}u\underbrace{(\nabla\gamma_{\mathcal{C}}(u)^\top u)}_{\lambda} \;=\; \frac{1}{\lambda}u - \frac{1}{\lambda}u \;=\; 0.$$

For the orthogonal component h_W, we have

$$J_q(u)h_W \;=\; \frac{1}{\lambda}h_W - \frac{\nabla\gamma_{\mathcal{C}}(u)^\top h_W}{\lambda^2}u.$$

Substituting $h = h_W + \beta u$ and the expression for $J_p(u)$ (5) into the condition $J_p(u)h = 0$ yields

$$r(\|u\|^2)J_q(u)h_W + 2\beta\|u\|^2 r'(\|u\|^2)q(u) \;=\; 0.$$

Substituting the expansion of $J_q(u)h_W$ allows us to group terms proportional to u. The equation becomes

$$\frac{r(\|u\|^2)}{\lambda}h_W \;+\; \mu u = 0,$$

where $\mu \in \mathbb{R}$ captures all radial coefficients:

$$\mu = 2\beta\|u\|^2 r'(\|u\|^2)\frac{1}{\lambda} - \frac{r(\|u\|^2)\nabla\gamma_{\mathcal{C}}(u)^\top h_W}{\lambda^2}.$$

Projecting the vector equation onto the subspace W eliminates μu, yielding

$$\frac{r(\|u\|^2)}{\lambda}h_W = 0.$$

Since $r > 0$ and $\lambda > 1$, we must have $h_W = 0$. Consequently, the radial term must also vanish, so $\mu u = 0$, implying $\mu = 0$. With $h_W = 0$, the expression for μ simplifies to

$$2\beta\|u\|^2 r'(\|u\|^2)\frac{1}{\lambda} = 0.$$

Assumption 2.4 ensures $r'(\rho) > 0$ for $\rho > 0$, so we must have $\beta = 0$. Thus $h = 0$, proving that $J_p(u)$ is invertible. Combining the results for interior and exterior points, we conclude that p is differentiable almost everywhere and $J_p(u)$ is invertible wherever it exists. □

B.2 Proofs of Section 3

We begin by establishing the equivalence of the optimal values between the unconstrained composite optimization $f(u)$ and the constrained optimization objective.

Theorem 3.1. (Equivalence of optimal values). *Let $\ell : \mathcal{C} \to \mathbb{R}$ be continuous. With $f \coloneqq \ell \circ p$, we have*

$$\inf_{u\in\mathbb{R}^n} f(u) \;=\; \inf_{x\in\mathrm{Int}(\mathcal{C})} \ell(x) \;=\; \inf_{x\in\mathcal{C}} \ell(x).$$

Moreover, $\arg\min f \neq \emptyset$ *if and only if* $\arg\min_{\mathcal{C}} \ell$ *intersects* $\mathrm{Int}(\mathcal{C})$. *In that case, $u^\star \in \arg\min f$ if and only if $p(u^\star) \in \arg\min_{\mathcal{C}} \ell$.*

Proof. By bijectivity of $p : \mathbb{R}^n \to \mathrm{Int}(\mathcal{C})$,

$$\inf_{u\in\mathbb{R}^n} f(u) = \inf_{x\in\mathrm{Int}(\mathcal{C})} \ell(x).$$

Since $u_0 \in \mathrm{Int}(\mathcal{C})$ and $\mathcal{C}$ is convex, for any $x \in \mathcal{C}$ and $\varepsilon \in (0, 1]$, $x_\varepsilon \coloneqq (1-\varepsilon)x + \varepsilon u_0 \in \mathrm{Int}(\mathcal{C})$ and $x_\varepsilon \to x$ as $\varepsilon \downarrow 0$. By continuity, $\ell(x_\varepsilon) \to \ell(x)$, hence $\inf_{\mathrm{Int}(\mathcal{C})} \ell \le \ell(x)$ for all $x \in \mathcal{C}$, implying $\inf_{\mathrm{Int}(\mathcal{C})} \ell \le \inf_{\mathcal{C}} \ell$. Since $\mathrm{Int}(\mathcal{C}) \subseteq \mathcal{C}$, we also have $\inf_{\mathcal{C}} \ell \le \inf_{\mathrm{Int}(\mathcal{C})} \ell$. Thus the optimal values coincide.

Finally, $\arg\min f \neq \emptyset$ holds if and only if there exists $x^\star \in \mathrm{Int}(\mathcal{C})$ with $\ell(x^\star) = \inf_{x\in\mathcal{C}} \ell(x)$, in which case $u^\star = p^{-1}(x^\star)$ exists and satisfies $u^\star \in \arg\min f$. Conversely, if $u^\star \in \arg\min f$, then $p(u^\star) \in \mathrm{Int}(\mathcal{C})$ satisfies $\ell(p(u^\star)) = \inf_{x\in\mathcal{C}} \ell(x)$, *i.e.*, $p(u^\star) \in \arg\min_{\mathcal{C}} \ell \cap \mathrm{Int}(\mathcal{C})$. □

Next, we establish the equivalence of stationary points between the two objectives.

Proposition 3.2. (Correspondence of stationary points). *Assume ℓ is C^1 on* $\mathrm{Int}(\mathcal{C})$. *Wherever p is differentiable, with $f = \ell \circ p$ we have $\nabla f(u) = J_p(u)^\top \nabla \ell\big(p(u)\big)$.*

Proof. The gradient identity follows from the chain rule, using that $p(u) \in \mathrm{Int}(\mathcal{C})$ for all u. If $J_p(u)$ is invertible and $\nabla f(u) = J_p(u)^\top \nabla \ell(p(u)) = 0$, then $\nabla \ell(p(u)) = 0$. The converse is immediate. □

Value gap under bounded latent norm. As discussed in Section 3, when the optimizer lies on the boundary $\partial\mathcal{C}$, exact attainment under a strict-interior parameterization requires an unbounded latent norm. This is a structural property of any homeomorphism $p : \mathbb{R}^n \to \mathrm{Int}(\mathcal{C})$: if a sequence $x_k \in \mathrm{Int}(\mathcal{C})$ converges to $x \in \partial\mathcal{C}$, then necessarily $\|p^{-1}(x_k)\| \to \infty$. Otherwise, a bounded subsequence $u_{k_j} \coloneqq p^{-1}(x_{k_j})$ would have a convergent subsequence with limit $u \in \mathbb{R}^n$, and continuity of p would imply $p(u) = x \in \partial\mathcal{C}$, contradicting $p(\mathbb{R}^n) = \mathrm{Int}(\mathcal{C})$.

The practically relevant question is therefore the suboptimality induced by a bounded latent budget. Following the coordinate convention in Section 2, we work in translated coordinates and assume the anchor is at the origin. Hence, for every unit vector v, we have

$$p(tv) = \psi_v(t)v, \qquad \psi_v(t) = r(t^2)\min\{t, \bar{t}(v)\},$$

where $\bar{t}(v)$ denotes the distance to the boundary along direction v. For a budget $M > 0$, define the latent budget set $B_M \coloneqq \{u \in \mathbb{R}^n : \|u\| \le M\}$, the reachable feasible set $\mathcal{C}_M \coloneqq p(B_M)$, and the restricted optimal value

$$\ell^\star_M \coloneqq \inf_{x\in\mathcal{C}_M} \ell(x) = \inf_{\|u\|\le M} \ell(p(u)).$$

Proposition 3.3. (Value gap under a latent norm budget). *Assume the problem admits a boundary minimizer $x^\star \in \arg\min_{x\in\mathcal{C}} \ell(x) \cap \partial\mathcal{C}$ with $x^\star = \bar{t}(v^\star)\, v^\star$ for some unit vector $v^\star$, and assume that ℓ is globally L-Lipschitz on $\mathcal{C}$. Define the boundary deficit $\delta_M(v) \coloneqq \bar{t}(v) - \psi_v(M)$. Then $0 \le \ell^\star_M - \ell^\star \le L\,\delta_M(v^\star)$. In particular, if $M \ge \bar{t}(v^\star)$, then $\delta_M(v^\star) = (1 - r(M^2))\,\bar{t}(v^\star)$, and hence $0 \le \ell^\star_M - \ell^\star \le L(1 - r(M^2))\,\bar{t}(v^\star)$.*

Proof. Let $x_M \coloneqq p(Mv^\star) = \psi_{v^\star}(M)\, v^\star$. Since $Mv^\star \in B_M$, we have $x_M \in \mathcal{C}_M$, so $\ell^\star_M \le \ell(x_M)$. Moreover, since $\mathcal{C}_M \subseteq \mathcal{C}$, we have $\ell^\star \le \ell^\star_M$. Using also $\ell^\star = \ell(x^\star)$, we obtain

$$0 \le \ell^\star_M - \ell^\star \le \ell(x_M) - \ell(x^\star) \le L\|x_M - x^\star\|.$$

Because x_M and $x^\star$ lie on the same ray, $\|x_M - x^\star\| = \bar{t}(v^\star) - \psi_{v^\star}(M) = \delta_M(v^\star)$. If $M \geq \bar{t}(v^\star)$, then $\psi_{v^\star}(M) = r(M^2)\bar{t}(v^\star)$, which gives $\delta_M(v^\star) = (1 - r(M^2))\bar{t}(v^\star)$ and hence the stated specialization. □

The same argument also shows that no value gap is induced when the constrained problem admits an interior minimizer that is reachable under the budget. Indeed, if $x^\circ \in \arg\min_{x \in \mathcal{C}} \ell(x) \cap \mathrm{Int}(\mathcal{C})$, then $u^\circ := p^{-1}(x^\circ)$ has finite norm. Hence, for every $M \geq \|u^\circ\|$, we have $x^\circ \in \mathcal{C}_M$ and therefore $\ell^\star_M = \ell^\star$.

We then extend this analysis to show that stationary points of the composite objective correspond to global minimizers of the constrained problem.

Corollary 3.4. (Global optimality of interior stationary points for convex losses). *Assume ℓ is convex and C^1 on $\mathrm{Int}(\mathcal{C})$, and let $f := \ell \circ p$. If p is differentiable at u and $\nabla f(u) = 0$, then $p(u)$ is a global minimizer of ℓ over $\mathcal{C}$.*

Proof. Let p be differentiable at u with $\nabla f(u) = 0$. Then $J_p(u)$ is invertible by Theorem 2.9. By Proposition 3.2, $\nabla \ell(p(u)) = 0$. Since $p(u) \in \mathrm{Int}(\mathcal{C})$ and ℓ is convex, $\nabla \ell(p(u)) = 0$ implies that $p(u)$ is a global minimizer of ℓ over $\mathrm{Int}(\mathcal{C})$, and hence over $\mathcal{C}$ by continuity. □

We now investigate the convergence rate and demonstrate the absence of a global Polyak-Łojasiewicz (PL) condition, which necessitates a more nuanced convergence analysis.

Lemma 3.5. (Absence of a global PL inequality in general). *Even if $\mathcal{C} := B_1(0)$, $\ell(x) := \|x\|^2$ and $u_0 := 0$, there exists no function r satisfying Assumption 2.4 such that $f = \ell \circ p$ satisfies a global PL inequality on $\mathbb{R}^n$.*

Proof. The unique minimizer of ℓ over $\mathcal{C}$ is $x^\star = 0$, and the corresponding optimal value $\ell^\star = 0$. Fix a unit vector v and consider $u(t) := tv$ with $t \geq 0$. For $\mathcal{C} = B_1(0)$ the hard radial projection is $q(u) = u$ if $\|u\| \leq 1$ and $q(u) = u/\|u\|$ otherwise; hence for $t \geq 1$ we have $q(u(t)) = v$ and therefore

$$p(u(t)) \;=\; r(t^2)\, v, \qquad f(u(t)) \;=\; \|p(u(t))\|^2 \;=\; r(t^2)^2.$$

Since $r(t^2) \to 1$ as $t \to \infty$, we obtain

$$f(u(t)) - f^\star \;=\; r(t^2)^2 \;\to\; 1 \quad (t \to \infty),$$

so along this ray the suboptimality converges to a strictly positive constant. Next, compute the gradient for $t > 1$. On $\{\|u\| > 1\}$, the hard radial projection $q(u) = u/\|u\|$ has Jacobian

$$J_q(u) \;=\; \frac{1}{\|u\|}\Big(I - \frac{uu^\top}{\|u\|^2}\Big),$$

so at $u(t) = tv$

$$J_q(u(t)) \;=\; \frac{1}{t}\big(I - vv^\top\big).$$

Substituting the expression for the Jacobian $J_p(u)$ given in (5), we obtain for $t > 1$

$$J_p(u(t)) = \frac{r(t^2)}{t}(I - vv^\top) \;+\; 2t\, r'(t^2)\, vv^\top.$$

By Proposition 3.2, the gradient is given by $\nabla f(u) = J_p(u)^\top \nabla \ell(p(u))$. Since $\ell(x) = \|x\|^2$ implies $\nabla \ell(x) = 2x$, and specifically $p(u(t)) = r(t^2)v$, we obtain

$$\nabla f(u(t)) \;=\; J_p(u(t))^\top \big(2r(t^2)v\big) \;=\; 2r(t^2) J_p(u(t))v,$$

where the last equality follows from the symmetry of $J_p(u(t))$. When applying the Jacobian matrix to the vector v, the tangential term vanishes because $(I - vv^\top)v = 0$. Only the radial term remains

$$J_p(u(t))v \;=\; \big(2tr'(t^2)vv^\top\big)\, v \;=\; 2tr'(t^2)v.$$

Substituting this back yields the final gradient norm

$$\|\nabla f(u(t))\| \;=\; \|2r(t^2) \cdot 2tr'(t^2)v\| \;=\; 4\, t\, r(t^2)\, r'(t^2).$$

It remains to show the existence of a sequence $t_k \to \infty$ such that $t_k r'(t_k^2) \to 0$. Since $\lim_{\rho\to\infty} r(\rho) = 1$, we have $\int_0^\infty r'(\rho)\,d\rho = 1 - r(0) < \infty$. We claim there exists a sequence $\rho_k \to \infty$ such that $\sqrt{\rho_k}\, r'(\rho_k) \to 0$. Indeed, if not, then there exist $\varepsilon > 0$ and $\rho_0 \geq 0$ such that $\sqrt{\rho}\, r'(\rho) \geq \varepsilon$ for all $\rho \geq \rho_0$, which implies

$$\int_0^\infty r'(\rho)\,d\rho \;\geq\; \varepsilon \int_{\rho_0}^\infty \frac{d\rho}{\sqrt{\rho}} \;=\; \infty,$$

a contradiction. Setting $t_k := \sqrt{\rho_k}$ yields $t_k \to \infty$ and $t_k r'(t_k^2) \to 0$. Along $u(t_k) = t_k v$ we therefore have $\|\nabla f(u(t_k))\| \to 0$ while $f(u(t_k)) - f^\star \to 1$. Consequently, no $\mu > 0$ can satisfy a global PL inequality $\|\nabla f(u)\|^2 \geq 2\mu\,(f(u) - f^\star)$ for all $u \in \mathbb{R}^n$. □

Although the global PL condition fails in general (Lemma 3.5), this does not preclude standard first-order guarantees on compact iterate sets. The key distinction is between generic differentiability and smoothness. Since the soft-radial map is locally Lipschitz, exact nondifferentiability events are typically avoided under continuous sampling assumptions; this justifies ordinary backpropagation at generic iterates. However, almost-everywhere differentiability is weaker than the L-smoothness required by the classical smooth SGD rate. When iterates may cross boundary or active-face switching surfaces of the soft-radial map, the natural regularity class is nonsmooth. Proposition 3.6 therefore states the compact-iterate nonsmooth guarantee, while the usual smooth SGD guarantee is recovered only as a special case when the reachable latent outputs remain in a compact region on which the soft-radial map is C^2; see, *e.g.*, Ghadimi and Lan [35], Bottou et al. [36], Clarke [37], and Davis et al. [18].

Generic differentiability. Let Σ_p denote the set of points where the projection p is non-differentiable. Because p is locally Lipschitz (Lemma 2.3), Rademacher's theorem dictates that Σ_p has a Lebesgue measure of zero. Let $\Phi(\theta, z) := g_\theta(z)$ represent the unconstrained network mapping. Assuming the parameters and inputs (θ, z) possess a joint density, and that the network mapping is non-degenerate (*i.e.*, its Jacobian maintains full rank, $\operatorname{rank} D_{(\theta,z)}\Phi = n$ almost everywhere), the coarea formula guarantees that the network's output $\Phi(\theta, z)$ has a distribution absolutely continuous with respect to Lebesgue measure. Because the network's output distribution is absolutely continuous and the non-differentiable set Σ_p has zero volume, it follows that $\mathbb{P}(g_\theta(z) \in \Sigma_p) = 0$. Under these assumptions, backpropagation through p is well-defined almost surely. For standard fully connected networks with a free output bias, this non-degenerate condition is immediate because the derivative of $g_\theta(z)$ with respect to the output bias has full row rank. If the law of θ_t remains absolutely continuous along training and fresh samples z_t are drawn from a density, the same argument applies at every finite iterate. This absolute-continuity preservation holds, for example, under full-dimensional additive continuous noise, or may be imposed as an assumption on the stochastic update map. Importantly, this almost-sure differentiability statement does not imply L-smoothness of the population objective; a locally Lipschitz function can be differentiable almost everywhere while still having gradient jumps.

Assumption B.1 (Compact stochastic approximation regime)**.** The iterates generated by stochastic subgradient descent remain in a compact set $K \subset \mathbb{R}^{d_\theta}$. The objective $F(\theta) := \mathbb{E}_z[\ell(p(g_\theta(z)))]$ is locally Lipschitz and tame on a neighborhood of K, and is bounded below on K. Moreover, the stochastic oracle satisfies the standard measurability, bounded-noise, and subgradient-oracle conditions of tame nonsmooth stochastic approximation.

Proposition 3.6. (Compact-iterate nonsmooth convergence). *Suppose Assumption B.1 holds. Then stochastic subgradient descent with step sizes $\eta_t \to 0$, $\sum_t \eta_t = \infty$, $\sum_t \eta_t^2 < \infty$ satisfies the subsequential convergence guarantee of Davis et al. [18, Cor. 5.9]: almost surely, every accumulation point of $\{\theta_t\}$ is Clarke stationary. Equivalently, for every convergent subsequence $\theta_{t_j} \to \bar{\theta}$, $0 \in \partial^\circ F(\bar{\theta})$* a.s.*, where $\partial^\circ F$ denotes the Clarke subdifferential.*

Proof. By Lemma 2.3, the soft-radial map p is locally Lipschitz. Therefore, whenever g_θ and ℓ are locally Lipschitz on compact sets, the single-sample composite $\theta \mapsto \ell(p(g_\theta(z)))$ is locally Lipschitz on compact parameter sets. In the empirical finite-sum setting,

$$F(\theta) = \frac{1}{m}\sum_{i=1}^m \varphi_i(\theta), \qquad \varphi_i(\theta) := \ell(p(g_\theta(z_i))),$$

local Lipschitzness is preserved under finite summation, so F is locally Lipschitz on a neighborhood of K. For general population expectations, local Lipschitzness can be verified using a uniform or integrable Lipschitz envelope, or assumed directly as in the proposition.

Tameness is also preserved under finite compositions and finite sums, provided the functions involved are definable in a common o-minimal structure. This covers standard semialgebraic settings, including ReLU networks, polyhedral feasible sets, and algebraic radial maps. Logistic sigmoid and $\tanh$-type activations can be handled in o-minimal structures containing the exponential function; otherwise, tameness should be verified or assumed directly.

On compact sets where the summands have a common Lipschitz constant, Clarke subgradients are uniformly bounded. Together with the compact-iterate, measurability, bounded-noise, and subgradient-oracle assumptions stated in the proposition, the tame nonsmooth stochastic approximation result of Davis et al. [18, Cor. 5.9] applies. Hence every accumulation point of the stochastic subgradient iterates is Clarke stationary almost surely. □

Smooth special case. The classical smooth SGD convergence rate can be recovered under a stronger "compact-patch" condition, which ensures the network's outputs remain within a smooth region of the projection layer. Let $V \supset K$ be an open convex neighborhood around the compact iterate set K, and let $\mathcal{Z}$ be compact with $\mathrm{supp}(\mathbb{P}_Z) \subseteq \mathcal{Z}$. We define the reachable latent set as $\mathcal{U}_V := \{g_\theta(z) : \theta \in \overline{V}, z \in \mathcal{Z}\}$. Suppose $\mathcal{U}_V$ is compactly contained in an open set U where the projection is twice continuously differentiable ($p \in C^2(U)$). Furthermore, assume the loss ℓ is C^2 on a neighborhood of $p(\mathcal{U}_V)$, and that the network g_θ has uniformly bounded first and second derivatives with respect to θ on $\overline{V} \times \mathcal{Z}$. Under these conditions, the composite objective F has a bounded Hessian on V, making F globally L-smooth on V for some finite L. To see why, let $h(u) := \ell(p(u))$. For a fixed input z, define the sample loss as $\psi_z(\theta) := h(g_\theta(z))$. Applying the chain rule yields the gradient and Hessian:

$$\nabla_\theta \psi_z(\theta) = D_\theta g_\theta(z)^\top \nabla h(g_\theta(z)),$$

$$\nabla^2_{\theta\theta} \psi_z(\theta) = D_\theta g_\theta(z)^\top \nabla^2 h(g_\theta(z)) D_\theta g_\theta(z) + \sum_i \partial_i h(g_\theta(z))\, \nabla^2_{\theta\theta} g_{\theta,i}(z).$$

Because of our compactness and bounded-derivative assumptions, all terms on the right-hand side are uniformly bounded on $\overline{V} \times \mathcal{Z}$. This allows us to interchange differentiation and expectation, proving that $\nabla^2 F$ is uniformly bounded on V. Consequently, if the stochastic-gradient oracle is unbiased and has bounded variance on K, the standard nonconvex SGD bound of Ghadimi and Lan [35] applies. For a horizon-dependent constant step size $\eta_t \equiv \eta = \Theta(T^{-1/2})$ with $\eta \le 1/L$, we obtain

$$\min_{0 \le t < T} \mathbb{E}\|\nabla F(\theta_t)\|^2 = \mathcal{O}(T^{-1/2}).$$

Geometric intuition. This compact-patch condition is fundamentally geometric. For smooth convex bodies, such as balls and ellipsoids, the ray-boundary intersection varies smoothly wherever it is transverse. However, for polyhedral sets—such as boxes, simplices, and capped simplices—the active boundary face changes depending on the ray direction, making p only piecewise smooth. In these cases, the smooth SGD guarantee only holds while the reachable set $\mathcal{U}_V$ stays a positive distance away from the boundary-switching and active-face switching surfaces. If the training iterates cross these surfaces, the objective becomes nonsmooth, and the Clarke-stationary guarantee detailed in Proposition 3.6 becomes the appropriate theoretical framework.

B.3 Proofs of Section 4

Finally, we prove that the constrained model inherits the universal approximation capabilities of the underlying unconstrained architecture.

Theorem 4.1. (Universal approximation on $\mathcal{C}$). *Let $\mathcal{Z} \subset \mathbb{R}^d$ be compact and let $\mathcal{G}$ be a universal approximator on $\mathcal{Z}$. For every continuous function $h : \mathcal{Z} \to \mathcal{C}$ and every $\varepsilon > 0$, there exists a function $g \in \mathcal{G}$ such that the constrained predictor $p \circ g$ satisfies* $\sup_{z \in \mathcal{Z}} \| p(g(z)) - h(z) \| \le \varepsilon$.

Proof. Without loss of generality, we assume $0 \in \mathrm{Int}(\mathcal{C})$ serves as the anchor point (see Coordinate Convention in Section 2). Throughout this proof, we denote the uniform norm of a function $\psi : \mathcal{Z} \to \mathbb{R}^n$ by $\|\psi\|_\infty := \sup_{z \in \mathcal{Z}} \|\psi(z)\|$. We separate the interior and boundary cases.

Interior case. Assume the target image satisfies $h(\mathcal{Z}) \subset \mathrm{Int}(\mathcal{C})$. Since $p : \mathbb{R}^n \to \mathrm{Int}(\mathcal{C})$ is a homeomorphism (Theorem 2.8), the function $\phi \coloneqq p^{-1} \circ h : \mathcal{Z} \to \mathbb{R}^n$ is continuous. Since $\mathcal{Z}$ is compact, the image $\phi(\mathcal{Z})$ is compact. Fix any $\delta_0 > 0$ (*e.g.*, $\delta_0 = 1$) and define the compact δ_0-thickening

$$K \coloneqq \big\{\xi \in \mathbb{R}^n \mid \mathrm{dist}\big(\xi, \phi(\mathcal{Z})\big) \le \delta_0\big\},$$

where $\mathrm{dist}(\xi, A) \coloneqq \inf_{a \in A} \|\xi - a\|$. By the local Lipschitz continuity of p on the compact set K, there exists a constant $L_K < \infty$ such that $\|p(a) - p(b)\| \le L_K \|a - b\|$ for all $a, b \in K$. By the universality of $\mathcal{G}$, there exists a function $g \in \mathcal{G}$ such that

$$\|g - \phi\|_\infty \;\le\; \delta, \qquad \delta \coloneqq \min\Big\{\delta_0,\; \frac{\varepsilon}{L_K}\Big\}.$$

This implies $g(\mathcal{Z}) \subset K$. Consequently, for all $z \in \mathcal{Z}$:

$$\begin{aligned} \|p(g(z)) - h(z)\| &= \|p(g(z)) - p(\phi(z))\| \\ &\le L_K \, \|g(z) - \phi(z)\| \\ &\le L_K \, \delta \;\le\; \varepsilon, \end{aligned}$$

which implies $\|p \circ g - h\|_\infty \le \varepsilon$, proving the claim for the interior case.

Boundary case. For a general target $h : \mathcal{Z} \to \mathcal{C}$, we define a strictly interior approximation h_ε by shrinking h towards the anchor. Define the shrink operator $S_\tau(x) \coloneqq (1 - \tau)x$. Since $\mathcal{C}$ is convex and $0 \in \mathrm{Int}(\mathcal{C})$, we have $S_\tau(\mathcal{C}) \subset \mathrm{Int}(\mathcal{C})$ for any $\tau \in (0, 1)$. Let $D \coloneqq \|h\|_\infty < \infty$. If $D = 0$, then $h \equiv 0$, and the interior case applies directly. Otherwise, we select a shrinkage factor τ sufficiently small to ensure h_ε remains within $\varepsilon/2$ of h. Specifically, set:

$$\tau \coloneqq \min\Big\{\tfrac{1}{2},\; \frac{\varepsilon}{2D}\Big\} \in (0, 1), \qquad h_\varepsilon \coloneqq S_\tau \circ h.$$

The uniform distance between the original and shrunken target is bounded by

$$\|h_\varepsilon - h\|_\infty \;=\; \sup_{z \in \mathcal{Z}} \|(1 - \tau)h(z) - h(z)\| \;=\; \tau D \;\le\; \frac{\varepsilon}{2}.$$

Since the new target satisfies $h_\varepsilon(\mathcal{Z}) \subset \mathrm{Int}(\mathcal{C})$, we can invoke the *Interior case* result with precision $\varepsilon/2$. This guarantees the existence of a function $g \in \mathcal{G}$ such that

$$\|p \circ g - h_\varepsilon\|_\infty \;\le\; \frac{\varepsilon}{2}.$$

Finally, applying the triangle inequality yields the total error bound

$$\|p \circ g - h\|_\infty \;\le\; \underbrace{\|p \circ g - h_\varepsilon\|_\infty}_{\le \varepsilon/2} + \underbrace{\|h_\varepsilon - h\|_\infty}_{\le \varepsilon/2} \;\le\; \varepsilon.$$

□

C Extended Numerical Results

In this section, we first provide additional insights into the model parameters, specifically how the radial contraction function $r(\cdot)$ and the scale parameter λ influence the effective optimization landscape. Next, we provide implementation details for the competitor algorithms, DC3 [13] and HardNet [16], which were adjusted to accommodate the (capped) simplex constraints considered in our experiments. We then detail the architectures and hyperparameters utilized within our study. Finally, we provide the experimental details needed to reproduce the reported benchmarks.

Computational resources. All experiments reported in the paper were run on a shared compute cluster using containerized single-GPU jobs. Each job was allocated one NVIDIA A100-SXM4-80GB GPU and 16 CPU cores. Training times (s) and inference times (ms) are reported for each experiment where applicable.

Affine-hull parametrization. For simplex and capped-simplex constraints, the feasible set lies in the affine hyperplane defined by the budget equality. For $\mathcal{C} \subseteq \{w \in \mathbb{R}^N : \mathbb{1}^\top w = 1\}$, where N denotes the number of assets or zones, a raw network output $u = g_\theta(z) \in \mathbb{R}^N$ defines the centered direction

$$d = (u - u_0) - \frac{1}{N}\mathbb{1}\mathbb{1}^\top(u - u_0).$$

The ray search is then performed along $u_0 + td$. Since $\mathbb{1}^\top d = 0$, the entire ray remains in the affine hyperplane $\mathbb{1}^\top w = 1$, and the boundary computation only needs to enforce the inequality constraints. The portfolio and ride-sharing experiments use this parametrization; ride-sharing applies it in the normalized unit-budget capped simplex before rescaling by the time-varying supply S_t.

C.1 Sensitivity Analysis: Radial Contraction and Scaling

In Sec. 5.1 we introduced radial contraction functions $r(\cdot)$ satisfying the properties of Assumption 2.4. For completeness, we restate them here. With fixed parameters $\varepsilon \in (0, 1)$ and $\lambda > 0$, we consider three families:

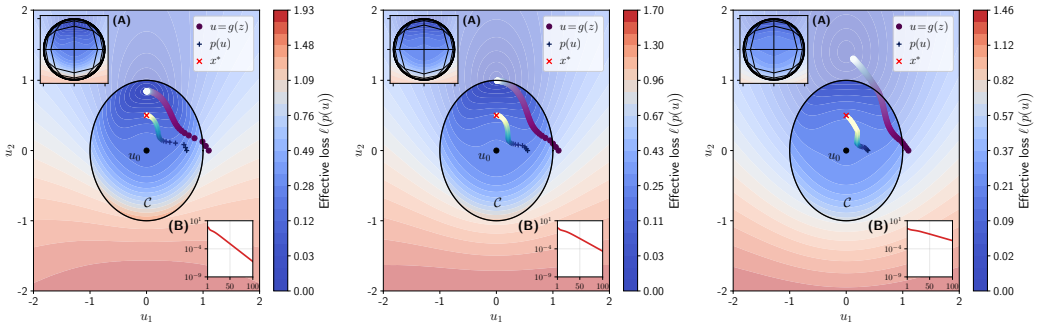
$$r(\rho) = \varepsilon + (1 - \varepsilon)\frac{\rho}{\rho + \lambda} \tag{6}$$

$$r(\rho) = \varepsilon + (1 - \varepsilon)\big(1 - e^{-\rho/\lambda}\big) \tag{7}$$

$$r(\rho) = \varepsilon + (1 - \varepsilon)\tanh(\rho/\lambda) \tag{8}$$

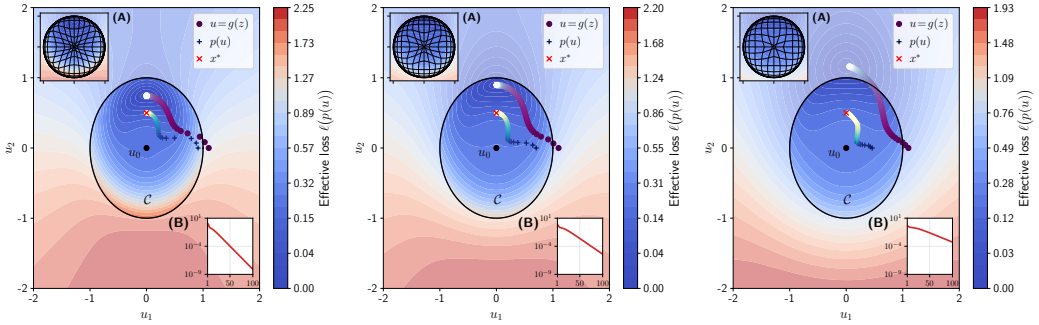
The *rational* radial contraction decays polynomially ($\mathcal{O}(\rho^{-1})$), whereas the *exponential* and *hyperbolic* forms exhibit exponential decay. Consequently, the rational form preserves larger gradient magnitudes for predictions falling far outside the feasible set ($\rho \gg 1$), which helps mitigate the vanishing gradient problem for extreme outliers.

Similar to the problem illustrated in Fig. 1, we evaluate the role of $r(\cdot)$ in terms of warping the effective loss landscape (A), influencing the training loss (B), and determining the convergence to the target x^*. Figures 3, 4, and 5 visualize the effective loss landscape for the *rational* (6), *exponential* (7), and *hyperbolic* (8) functions, respectively. Furthermore, we illustrate in Fig. 6 how the minimum scaling threshold ε modifies the landscape geometry surrounding the chosen anchor u_0.



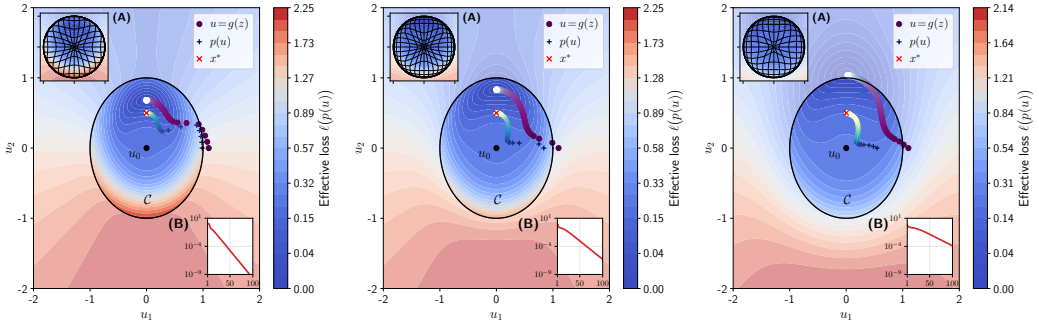
(a) $\lambda = 0.5$. (b) $\lambda = 1.0$. (c) $\lambda = 2.0$

Figure 3: Soft-radial projection $p(u)$ with *rational* radial contraction (6).



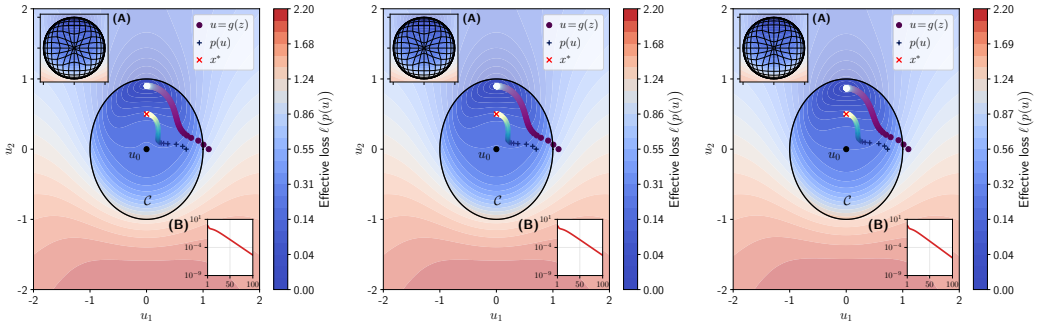
(a) $\lambda = 0.5$. (b) $\lambda = 1.0$. (c) $\lambda = 2.0$.

Figure 4: Soft-radial projection $p(u)$ with *exponential* radial contraction (7).



(a) $\lambda = 0.5$. (b) $\lambda = 1.0$. (c) $\lambda = 2.0$.

Figure 5: Soft-radial projection $p(u)$ with *hyperbolic* radial contraction (8).



(a) $\varepsilon = 0.001$. (b) $\varepsilon = 0.01$. (c) $\varepsilon = 0.1$.

Figure 6: Soft-radial projection $p(u)$ for varying minimum scaling threshold ε.

C.2 Adaptation of Competitor Methods to Capped Simplex Constraints

We explain the adaptation of the competitor methods with respect to the capped simplex introduced in the portfolio setting and formally defined as

$$\mathcal{C} \;=\; \Big\{ w \in \mathbb{R}^N \;\big|\; 0 \le w,\; \mathbb{1}^\top w = 1,\; w \le c \Big\},$$

where w is the portfolio weight and c is a cap on the weight.

HardNet. The HardNet framework [16] enforces feasibility by appending a differentiable projection layer to an unconstrained predictor $g_\theta : \mathcal{Z} \to \mathbb{R}^N$. Given an input $z \in \mathcal{Z}$ and raw output $u \coloneqq g_\theta(z)$, the method considers input-dependent affine constraints of the form

$$\underline{b}(z) \le A(z)w \le \bar{b}(z).$$

Min and Azizan [16] derive a closed-form projection for this case, denoted *HardNet-Aff*:

$$P_{\text{HN}}(u; z) \;=\; u \;+\; A_{\text{R}}^{\dagger}(z)\Big(\operatorname{ReLU}\big(\underline{b}(z) - A(z)u\big) - \operatorname{ReLU}\big(A(z)u - \bar{b}(z)\big)\Big), \tag{9}$$

where $A_{\text{R}}^{\dagger}(z) \coloneqq A(z)^\top \big(A(z)A(z)^\top\big)^{-1}$ is the Moore–Penrose *right* pseudoinverse. This closed form assumes that $A(z)$ has full row rank (Assumption 5 in [16]), describing an underdetermined system. In our portfolio setting, the feasible set is a capped simplex; stacking the equality constraint with the $2N$ bound constraints yields an *overdetermined* affine system. Consequently, the matrix $AA^\top$ is singular and the right-pseudoinverse expression in (9) is not directly applicable. We therefore use a least-squares relaxation of the HardNet correction. This involves defining a (signed) violation vector

$$v(u; z) \;\coloneqq\; \operatorname{ReLU}\big(\underline{b}(z) - Au\big) - \operatorname{ReLU}\big(Au - \bar{b}(z)\big). \tag{10}$$

Since the resulting static constraint matrix A has full column rank in our construction, we compute a least-squares correction direction via the *left* pseudoinverse

$$\Delta(u; z) \;\coloneqq\; \arg\min_{\Delta \in \mathbb{R}^N} \; \|A\Delta - v(u; z)\|^2 \;=\; A_{\text{L}}^{\dagger}\, v(u; z), \qquad A_{\text{L}}^{\dagger} \coloneqq (A^\top A)^{-1} A^\top. \tag{11}$$

We then set $\tilde{w} \coloneqq u + \Delta(u; z)$. A conceptually broader but computationally slower alternative proposed by the authors for general convex constraints is to embed the projection as a differentiable convex optimization layer [12], referred to as *HardNet-Cvx*. To remove any remaining numerical residuals during evaluation, we additionally apply an exact orthogonal projection onto the capped simplex, $w = P(\tilde{w})$, to guarantee $w \in \mathcal{C}$.

Symmetric DC3 for the capped simplex. We adapt the Deep Constraint Completion and Correction (DC3) framework [13] to $\mathcal{C}$. Standard DC3 selects $N - 1$ independent variables and defines the remaining coordinate as dependent via the equality constraint. Applied directly to the raw network output $u = g_\theta(z)$, this introduces an asymmetry: the dependent coordinate absorbs the entire sum deviation, often producing unstable magnitudes and gradients. To mitigate this, we employ a *symmetric* initialization that first distributes the sum deviation evenly by orthogonally projecting u onto the simplex hyperplane $\mathcal{H} \coloneqq \{w \in \mathbb{R}^N : \mathbf{1}^\top w = 1\}$. Concretely, we set

$$w^{(0)} \;\coloneqq\; P_{\mathcal{H}}(u) \;=\; u - \frac{1}{N}\big(\mathbb{1}^\top u - 1\big)\mathbb{1} \;\in\; \mathcal{H}. \tag{12}$$

We then take the first $N - 1$ coordinates as free variables $\xi \coloneqq (w_1^{(0)}, \ldots, w_{N-1}^{(0)})^\top \in \mathbb{R}^{N-1}$ and enforce the equality constraint through a completion map

$$w(\xi) \;\coloneqq\; \begin{bmatrix} \xi \\ 1 - \mathbb{1}^\top \xi \end{bmatrix}, \tag{13}$$

which ensures $\mathbb{1}^\top w(\xi) = 1$ for all ξ. To correct violations of the bound constraints $0 \le w \le c$, we minimize a squared hinge-residual energy

$$\mathcal{V}(\xi) \;\coloneqq\; \sum_{i=1}^{N} \Big(\operatorname{ReLU}\big(-w_i(\xi)\big)^2 + \operatorname{ReLU}\big(w_i(\xi) - c_i\big)^2 \Big), \tag{14}$$

via an unrolled gradient descent loop on the free variables

$$\xi^{(t+1)} \;=\; \xi^{(t)} - \eta\, \nabla_{\xi} \mathcal{V}\big(\xi^{(t)}\big), \qquad t = 0, \ldots, L-1. \tag{15}$$

The final output is given by $w^{(L)} = w(\xi^{(L)})$. Gradients are backpropagated through the completion map (13), properly accounting for the dependence of the dependent coordinate on the free variables. Similar to the HardNet implementation, we apply an orthogonal projection during evaluation to ensure strict feasibility.

For the standard simplex, orthogonal projection uses a sorting-based threshold algorithm; for the capped simplex, it uses bisection on the dual threshold. For network-constrained polytopes, we use Dykstra's algorithm with feasibility tolerance 10^{-6}.

C.3 Controlled Geometry Benchmarks

The application benchmarks combine several sources of difficulty: data noise, architecture choice, loss geometry, and constraint enforcement. We therefore include two controlled benchmarks whose purpose is to isolate the geometry of the projection layer itself. Both are amortized convex quadratic programs over the same type of affine polytope, $\mathcal{C} = \{x \in \mathbb{R}^n : A_{\text{eq}} x = b_{\text{eq}},\ A_{\text{ineq}} x \le b_{\text{ineq}}\}$. The feasible set is known exactly, and each sample has a known constrained optimizer. We can therefore vary one geometric property at a time while keeping the data-generating process fixed, and evaluate methods by optimality gap, training time, and inference latency.

Boundary-margin benchmark. The first benchmark varies the distance between the downstream optimizer and the boundary. For each sample, we draw a direction from an interior reference point and place the quadratic minimizer at a prescribed relative margin η from $\partial\mathcal{C}$. Smaller values of η generate instances whose optimum lies closer to a boundary face. This is the regime in which boundary-valued enforcement maps, such as hard radial maps or orthogonal projections, can lose informative gradients by mapping many infeasible predictions to the same face. The projection anchor u_0 is fixed throughout, so changes in the results reflect sensitivity to the location of the optimum rather than changes in anchor quality.

Table 4: Boundary-margin sensitivity results aggregated over five seeds (mean ± std).

η	Method	Rel. gap↓	Training [s]↓	Inference [ms]↓
0.1	O-Proj	0.07 (0.01)	689.43 (50.03)	135.40 (52.87)
0.1	DC3	0.07 (0.01)	33.35 (0.59)	25.74 (2.72)
0.1	HardNet	0.07 (0.01)	3.15 (0.10)	1.17 (0.15)
0.1	RAYEN	0.03 (0.00)	2.60 (0.47)	1.10 (0.23)
0.1	SRP	0.02 (0.00)	3.43 (0.16)	1.27 (0.32)
0.01	O-Proj	0.07 (0.01)	800.86 (40.53)	180.47 (73.27)
0.01	DC3	0.07 (0.01)	33.48 (1.60)	27.73 (2.34)
0.01	HardNet	0.07 (0.01)	2.75 (0.07)	0.99 (0.05)
0.01	RAYEN	0.03 (0.00)	2.56 (0.38)	0.93 (0.16)
0.01	SRP	0.03 (0.00)	2.67 (0.17)	0.90 (0.09)
0.0	O-Proj	0.07 (0.01)	803.97 (19.46)	224.84 (104.72)
0.0	DC3	0.07 (0.01)	33.99 (1.57)	34.54 (8.36)
0.0	HardNet	0.07 (0.01)	2.67 (0.13)	0.96 (0.11)
0.0	RAYEN	0.03 (0.00)	2.60 (0.42)	0.94 (0.21)
0.0	SRP	0.03 (0.00)	2.10 (0.15)	0.69 (0.02)

Anchor-conditioning benchmark. The second benchmark holds the feasible set, data distribution, and constrained optima fixed, and varies only the interior anchor u_0 used by radial methods. We move the anchor toward $\partial\mathcal{C}$ through an anchor-slack parameter δ, with smaller δ indicating a lower-slack anchor. Since the underlying quadratic programs do not change, variation across this table should be interpreted as conditioning induced by the projection geometry, not as a change in statistical or optimization difficulty.

Boundary optima and matched hard-radial control. To examine the effect of finite latent norm budgets when optimal solutions lie on the feasible boundary, we consider 1,024 quadratic optimization

Table 5: Anchor-sensitivity results aggregated over five seeds (mean $\pm$ std).

δ	Method	Rel. gap↓	Training [s]↓	Inference [ms]↓
0.1	O-Proj	0.07 (0.01)	561.24 (24.90)	168.92 (64.50)
0.1	DC3	0.07 (0.01)	46.58 (9.42)	34.53 (20.08)
0.1	HardNet	0.07 (0.01)	11.36 (3.12)	6.58 (4.51)
0.1	RAYEN	0.04 (0.00)	7.95 (1.11)	1.93 (1.15)
0.1	SRP	0.04 (0.00)	8.06 (2.62)	2.38 (1.31)
0.01	O-Proj	0.07 (0.01)	571.92 (17.42)	205.95 (74.30)
0.01	DC3	0.07 (0.01)	43.76 (6.59)	50.04 (24.16)
0.01	HardNet	0.07 (0.01)	2.52 (0.07)	0.77 (0.30)
0.01	RAYEN	0.05 (0.01)	3.86 (1.18)	0.95 (0.52)
0.01	SRP	0.04 (0.00)	2.89 (0.45)	0.78 (0.15)

instances over a fixed polytope in $\mathbb{R}^{12}$ with intrinsic dimension eight, four equality constraints, and 32 inequalities. Each instance has an optimum on a designated facet, with boundary-pressure parameter $\beta = 1$. SRP uses the rational contraction with $\lambda = 1$ and $\varepsilon = 0.05$. As a control, we introduce a matched hard-radial projection (Matched HR) that applies the same latent norm budget M and contracts the hard-radial output toward the anchor by $r(M^2)$. This matches the limiting interior margin of SRP while retaining the many-to-one behavior of hard radial projection. Both methods use a two-layer MLP of width 64, trained for 50 epochs over five seeds. Performance is measured by the test-set mean objective gap, normalized by the improvement from the anchor to the constrained optimum.

Table 6: Relative regret for boundary-optimum instances under different latent norm budgets M, aggregated over five seeds (mean $\pm$ std). Lower is better.

M	Matched HR	SRP
1	0.4454 (0.0035)	0.4446 (0.0034)
2	0.1979 (0.0048)	0.1962 (0.0038)
4	0.0937 (0.0057)	0.0962 (0.0044)
8	0.0632 (0.0060)	0.0711 (0.0047)
16	0.0554 (0.0060)	0.0654 (0.0046)

Table 6 shows that relative regret decreases with the latent norm budget for both methods. Matched HR achieves lower mean regret for larger budgets, illustrating that injectivity alone does not guarantee superior optimization performance when the optimum lies on the boundary. Without an explicit latent norm budget, SRP achieves a relative regret of $0.0636\ (0.0046)$, compared with $0.0529\ (0.0061)$ for RAYEN.

End-to-end training and fine-tuning. To assess the applicability of SRP to pretrained neural networks, we compare direct end-to-end training with unconstrained pretraining followed by fine-tuning on 1,024 quadratic optimization instances over a fixed polytope in $\mathbb{R}^{12}$ with intrinsic dimension eight. Direct end-to-end training optimizes the constrained model for 50 epochs. Alternatively, the base network is pretrained for 20 epochs to predict the reduced coordinates of the constrained optimizers using mean squared error, followed by 30 epochs of end-to-end fine-tuning. A third setting evaluates the pretrained network immediately after attaching the constraint layer, without fine-tuning (zero-shot). Results are aggregated over five seeds. Performance is measured by the test-set mean objective gap relative to the absolute optimal objective value, using a different normalization from the anchor-relative regret reported above.

Table 7 shows that end-to-end training and fine-tuning achieve comparable performance, with substantially larger objective gaps under zero-shot evaluation. SRP achieves lower mean gaps than RAYEN under both training strategies, whereas RAYEN performs better when the projection layer is applied without fine-tuning. These results demonstrate that SRP supports both direct end-to-end training and the adaptation of pretrained predictors.

Table 7: Relative objective gap under different training strategies, aggregated over five seeds (mean $\pm$ std).

Method	Training protocol	Relative gap↓
SRP	End-to-end (50 epochs)	$0.0264\,(0.0022)$
	Pretrained, zero-shot	$0.2373\,(0.0225)$
	Pretrained, fine-tuned (20+30 epochs)	$0.0292\,(0.0032)$
RAYEN	End-to-end (50 epochs)	$0.0312\,(0.0030)$
	Pretrained, zero-shot	$0.1762\,(0.0301)$
	Pretrained, fine-tuned (20+30 epochs)	$0.0361\,(0.0049)$

C.4 Portfolio and Ride-Sharing Hyperparameters

Table 8 summarizes the hyperparameters used in the main experiments, distinguishing between the portfolio optimization and ride-sharing dispatch task. For both domains, we perform a grid search to select the optimal configuration based on validation performance. As noted earlier, for HardNet and DC3, we apply a final orthogonal projection to the candidate solution during evaluation only. This ensures feasibility of the reported decisions; the projection is not applied during training. Reported inference times include the evaluation-time correction.

Table 8: Hyperparameter settings for portfolio optimization and ride-sharing dispatch experiments.

Parameter	**Portfolio Optimization**	**Ride-Sharing Dispatch**
Architecture & Training		
Base Architecture	LSTM	MLP
Hidden Units	$\{32, 64\}$	
Dropout Rate	$\{0.1, 0.2\}$	
Optimizer	Adam	
Learning Rate	$\{1 \times 10^{-4}, 5 \times 10^{-4}\}$	
Lookback Horizon (H)	$\{10, 30\}$	24
Batch Size	64	128
Training Epochs	50	100
Projection Layer Hyperparameters		
Softmax Temperature (τ)	$\{0.1, 0.5, 1.0, 2.0\}$	
Soft-radial λ	$\{0.5, 1.0, 2.0, 5.0, 10.0\}$	
Soft-radial ε	0.05	
Soft-radial $r(\cdot)$	{Rational, Exponential, Hyperbolic}	
HardNet LS-correction iterations	$\{1, 3, 5\}$	
DC3 Steps	$\{1, 3, 5\}$	
DC3 Learning Rate	$\{10^{-2}, 10^{-1}\}$	
DC3 Momentum	$\{0.5, 0.9\}$	

Results analysis. Table 4 shows that moving the constrained optimizer toward the boundary does not degrade the radial methods: RAYEN and SRP remain stable across all margins and achieve lower relative gaps than the projection and repair baselines. Table 5 isolates a different effect. When the radial anchor is moved closer to the boundary, we observe a mild degradation, consistent with poorer conditioning of the ray-boundary map for low-slack anchors.

C.5 Extended Analysis: Portfolio Optimization

C.5.1 Discussion: Deep Portfolio Optimization vs. Classical Approaches

The formulation presented in Section 5.2 represents a *deep portfolio optimization* problem [38], which fundamentally differs from the classical mean-variance optimization framework [39].

Predict-then-optimize vs. end-to-end learning. Classical mean-variance optimization typically operates in a "predict-then-optimize" framework: one first estimates the expected return vector $\mu \in \mathbb{R}^N$ and the covariance matrix $\Sigma \in \mathbb{S}^N_+$, and then solves a quadratic program to maximize $w^\top \mu - \zeta w^\top \Sigma w$. This two-step process suffers from two major limitations. First, it separates the prediction error from

the optimization cost; a small error in estimating Σ can lead to drastically suboptimal portfolios (a phenomenon formally known as the "error maximization" property of Markowitz [40]). Second, while standard sample estimators neglect temporal dependencies (implicitly assuming i.i.d. returns), more advanced dynamic estimators (*e.g.*, generalized autoregressive conditional heteroskedasticity (GARCH) model) often impose rigid parametric structures that fail to capture the complex, non-linear interactions between assets that recurrent neural network architectures in the form of g_θ are designed to exploit.

Parameter sensitivity and convexity. A significant hurdle in classical frameworks is the sensitivity to the risk-aversion parameter ζ, which requires frequent recalibration. Our formulation bypasses this by directly maximizing the Sharpe ratio—a scale-invariant metric that eliminates the need for a subjective risk-return trade-off coefficient. While the static Sharpe ratio maximization is a quasiconcave problem that can be reformulated as a Second-Order Cone Program (SOCP) for global convergence, parameterizing the weights via a neural network g_θ introduces non-convexity. We accept this trade-off to leverage the representational power of deep learning; the model can ingest complex state histories to generate non-linear weight mappings, far surpassing the constraints of the linear-quadratic assumptions in traditional finance.

Transaction cost as temporal regularization. Finally, we address the cost of trading. Classical frameworks often use static L_2 regularization to prevent overfitting or L_1 regularization to enforce sparsity [41]. However, these spatial constraints fail to penalize policy volatility: a network can satisfy strict sparsity constraints while still oscillating wildly between assets over time.

To enforce temporal consistency, we explicitly model transaction costs via the term $\frac{\gamma}{2}\|w_t - w_{t-1}^+\|_1$. Unlike static penalties, this targets the velocity of the portfolio turnover. Crucially, costs are computed relative to the *drifted* portfolio w_{t-1}^+, ensuring that passive price movements do not incur penalties. Let w_{t-1} be the weights at $t-1$. Over the interval $(t-1, t]$, simple asset returns y_t cause these weights to evolve naturally into the drifted weights $w_{t-1}^+ = (\mathrm{diag}(\mathbb{1} + y_t)\, w_{t-1})/((\mathbb{1} + y_t)^\top w_{t-1})$. The numerator captures asset-specific growth, while the denominator renormalizes the sum.

Remark C.1 (Cost normalization). The factor $1/2$ normalizes the L_1 distance to represent *one-way turnover* (the fraction of portfolio value actively replaced). Thus, γ represents the effective cost per unit of turnover.

This formulation has a profound effect on the learning dynamics. By including this cost directly in the objective, the gradient signal penalizes high-frequency control noise. The policy learns a valid economic trade-off: it naturally creates a *hysteresis effect*, where the portfolio weights exhibit inertia against small market fluctuations. This recovers the classical *no-trade zone* behavior predicted by Davis and Norman [42]: rebalancing is only performed if the expected Sharpe ratio gain strictly outweighs the explicit cost of the action.

C.5.2 Descriptive Statistics

In Table 9, we provide descriptive statistics for the asset universes included in our analysis. We evaluate performance across three distinct portfolios, each representing a different difficulty level and correlation structure:

- *Global:* The Global portfolio consists of 11 highly liquid exchange-traded products that provide broad exposure to global equities, fixed income, commodities, and real estate. It represents a macro-allocation problem characterized by lower average cross-asset correlation.
- *Sectors:* The Sectors dataset comprises the 11 primary Global Industry Classification Standard (GICS) sectors. It serves as a standard industry benchmark with high inter-asset correlation, effectively representing the broader US equity market.
- *Liquid:* The Liquid collection contains a selection of 50 high-volume individual stocks. This universe captures the higher volatility and idiosyncratic risk associated with individual stock picking.

Returns are derived from Yahoo Finance adjusted closing prices to account for corporate actions, such as splits and dividends, and are partitioned chronologically to strictly prevent look-ahead bias.[5] The training set spans from 2010/01/01 to 2018/12/31, covering the post-financial-crisis recovery.

[5]We note that survivorship bias is not compensated for within our analysis, and assets with missing data are omitted from the study.

The validation set, used for hyperparameter tuning, covers 2019/01/01 to 2020/12/31. The test set extends from 2021/01/01 to 2025/12/31.

Table 9 reports the annualized mean and median returns alongside annualized volatility. To fully characterize the risk profile and market dynamics, we also include the Maximum Drawdown (Max DD), the average pairwise correlation between assets ($\bar{\varrho}$), the lag-1 autocorrelation (AC(1)), and the Kolmogorov-Smirnov p-value (p_{KS}). As shown in the table, the test period exhibits significantly higher volatility and maximum drawdown compared to the training period, reflecting more turbulent market conditions. This distribution shift is quantified by p_{KS}: for the Global, Liquid, and Sectors universes, $p_{\text{KS}} < 0.05$ in the test set, rejecting the null hypothesis that the test data follows the training distribution. This confirms that our evaluation effectively tests the model's generalization to out-of-distribution market regimes.

Table 9: Summary statistics. Returns, Volatility, and Drawdown are in percentages. $\bar{\varrho}$ is average correlation. p_{KS} tests for regime shift.

Universe	Split	μ (%)	Median (%)	σ (%)	MaxDD (%)	$\bar{\varrho}$	AC(1)	p_{KS}
Global	Train	7.30	11.84	17.47	−19.37	0.35	−0.037	
Global	Val	7.46	14.22	15.68	−13.87	0.32	−0.020	0.607
Global	Test	9.30	13.60	22.34	−28.46	0.46	−0.045	0.011
Liquid	Train	19.12	14.72	24.98	−16.11	0.38	−0.021	
Liquid	Val	19.71	30.76	25.40	−18.11	0.39	−0.029	0.319
Liquid	Test	16.95	16.93	31.44	−30.56	0.40	−0.088	0.019
Sectors	Train	13.46	20.17	17.05	−18.61	0.70	−0.037	
Sectors	Val	9.75	26.72	17.15	−18.47	0.58	−0.003	0.826
Sectors	Test	13.99	22.01	24.57	−36.91	0.65	−0.107	0.007

C.5.3 Implementation Details

In the portfolio optimization setting, we augment raw asset returns with rolling volatility and rolling correlation to a cross-sectional market proxy (defined as the mean return across the asset universe). These features are computed over the lookback horizon H and normalized using training-set statistics to prevent look-ahead bias. We acknowledge that achieving state-of-the-art financial performance typically requires a broader set of predictive signals (alphas). However, our primary objective is to evaluate the stability of the learning process under the high noise-to-signal ratio characteristic of financial time series, rather than to maximize the Sharpe ratio through signal engineering.

Anchor selection. For portfolio constraints on the standard simplex, we use the barycenter $u_0 = \mathbb{1}/N$. For capped-simplex constraints with cap vector c, we use

$$(u_0)_i = \frac{c_i}{\sum_j c_j} \quad \forall i.$$

Provided that the capped simplex has nonempty relative interior, this point lies in the relative interior of the feasible set. The construction is deterministic, depends only on the constraint set, and reduces to the barycenter whenever the cap vector is constant.

Smooth transaction-cost approximation. As discussed in Sec. 5.2, the L_1 transaction-cost term in (3) is non-differentiable at zero, which can make gradient-based optimization less stable. We therefore replace the exact L_1 norm with a smooth Charbonnier/Pseudo-Huber approximation,

$$L_\delta(w_t, w_{t-1}^+) = \sum_{i=1}^{N} \left(\sqrt{\delta^2 + (w_{i,t} - w_{i,t-1}^+)^2} - \delta \right), \tag{16}$$

where $\delta > 0$ controls the smoothness of the approximation. As $\delta \to 0$, this penalty converges pointwise to $\|w_t - w_{t-1}^+\|_1$, thereby preserving the intended transaction-cost structure while providing stable gradients during training.

C.5.4 Additional Empirical Results

In Section 5, we presented results for the Sharpe ratio maximization formulation defined in (3). Table 11 reports LSTM results across all three asset universes, while Table 12 reports the corresponding Transformer results. Consistent with our findings in Section 5, our *soft-radial projection* achieves strong financial performance across different datasets and network architectures. We further compare against GLinSAT [43] and T-SKM-Net [44] on the Global dataset in Table 10.

Table 10: Additional portfolio optimization results for GLinSAT and T-SKM-Net on the *Global* dataset using an LSTM policy, aggregated over five seeds (mean ± std). Both methods enforce the capped-simplex constraints.

Universe	Set	Method	SR (Net)↑	Turnover↓	Training [s]↓	Inference [ms]↓
Global	$\mathcal{C}$	GLinSAT	0.68 (0.03)	0.02 (0.01)	1815.99 (435.36)	361.51 (90.96)
		T-SKM-Net	0.43 (0.08)	0.18 (0.01)	228.72 (0.73)	35.90 (1.25)

Table 11: Portfolio optimization results for the LSTM policy over five seeds (mean ± std). The Global results reproduce Table 1 for completeness.

Universe	Set	Method	SR (Net)↑	Turnover↓	Training [s]↓	Inference [ms]↓
Global	Δ	Softmax	0.65 (0.09)	0.11 (0.02)	11.16 (0.19)	0.34 (0.08)
Global	$\mathcal{C}$	O-Proj	0.44 (0.13)	0.16 (0.02)	46.29 (0.97)	5.81 (0.19)
		DC3	0.25 (0.18)	0.23 (0.01)	23.52 (0.41)	8.44 (0.36)
		HardNet	0.46 (0.21)	0.16 (0.04)	17.27 (0.92)	6.87 (0.10)
		RAYEN	0.60 (0.06)	0.11 (0.01)	14.09 (0.25)	0.68 (0.01)
		SRP	0.70 (0.06)	0.04 (0.02)	14.73 (0.14)	0.76 (0.01)
Liquid	Δ	Softmax	0.70 (0.13)	0.14 (0.02)	10.94 (0.22)	0.33 (0.07)
Liquid	$\mathcal{C}$	O-Proj	0.64 (0.07)	0.22 (0.01)	56.01 (1.18)	6.61 (1.17)
		DC3	0.59 (0.07)	0.23 (0.01)	27.84 (0.14)	6.62 (0.82)
		HardNet	0.59 (0.10)	0.20 (0.00)	29.49 (0.35)	7.52 (1.18)
		RAYEN	0.85 (0.05)	0.12 (0.00)	14.97 (1.50)	0.71 (0.10)
		SRP	0.87 (0.05)	0.08 (0.01)	14.56 (0.28)	0.75 (0.10)
Sectors	Δ	Softmax	0.70 (0.12)	0.12 (0.04)	11.60 (0.30)	0.33 (0.07)
Sectors	$\mathcal{C}$	O-Proj	0.82 (0.01)	0.01 (0.00)	51.53 (1.34)	6.11 (0.87)
		DC3	0.81 (0.02)	0.01 (0.00)	28.71 (0.73)	6.92 (1.13)
		HardNet	0.82 (0.01)	0.01 (0.00)	26.44 (0.39)	6.13 (0.69)
		RAYEN	0.82 (0.01)	0.01 (0.00)	14.31 (0.43)	0.66 (0.04)
		SRP	0.83 (0.00)	0.00 (0.00)	14.21 (0.18)	0.74 (0.15)

C.6 Extended Analysis: Ride-Sharing Dispatch

Our taxi dataset is derived from the publicly available records of the NYC Taxi and Limousine Commission (TLC) for the period of January to June 2024. To ensure a robust evaluation of temporal generalization, the data is split chronologically into 70% training, 15% validation, and 15% testing sets. To capture the multi-scale periodicity of urban mobility, we represent temporal states using cyclical sine-cosine encodings for both daily and weekly cycles. Unlike categorical indicators that create artificial boundaries at midnight, this continuous representation allows the policy network g_θ to learn smooth, time-dependent transition boundaries. This enables the model to proactively align fleet distribution with anticipated demand peaks, maximizing the global served rate within the constraints of the time-varying supply S_t.

Anchor selection. For the normalized taxi capped simplex $\{u \geq 0 : \mathbb{1}^\top u = 1,\ u_i \leq \kappa\ \forall i\}$, the reported settings use $N = 150$ zones and satisfy $N\kappa > 1$. Hence the barycenter $\bar{u}_0 = \mathbb{1}/N$ is strictly feasible. The anchor in allocation coordinates is $u_{0,t} = S_t \bar{u}_0$.

In Section 5 (Table 2), we presented results for an MLP model subject to the scaled capped simplex with $\kappa = 0.1$. Here, we introduce a more challenging experimental setting with $\kappa = 0.02$. This

Table 12: Portfolio optimization results for the Transformer policy over five seeds (mean ± std).

Universe	Set	Method	SR (Net)↑	Turnover↓	Training [s]↓	Inference [ms]↓
Global	Δ	Softmax	0.32 (0.18)	0.21 (0.09)	16.20 (0.10)	0.73 (0.14)
Global	$\mathcal{C}$	O-Proj	0.51 (0.20)	0.10 (0.07)	51.52 (1.35)	6.08 (0.52)
		DC3	0.52 (0.10)	0.17 (0.04)	40.54 (0.40)	8.57 (0.52)
		HardNet	0.48 (0.07)	0.04 (0.07)	34.63 (1.24)	7.25 (0.91)
		RAYEN	0.46 (0.11)	0.10 (0.02)	18.34 (0.40)	1.07 (0.01)
		SRP	0.56 (0.09)	0.15 (0.07)	18.99 (0.44)	1.09 (0.03)
Liquid	Δ	Softmax	−0.03 (0.25)	0.41 (0.14)	17.24 (0.21)	0.73 (0.04)
Liquid	$\mathcal{C}$	O-Proj	0.68 (0.14)	0.11 (0.03)	56.76 (0.48)	6.72 (0.97)
		DC3	0.64 (0.15)	0.16 (0.05)	34.17 (1.32)	7.55 (1.04)
		HardNet	0.73 (0.18)	0.06 (0.03)	35.49 (0.51)	7.55 (1.21)
		RAYEN	0.79 (0.07)	0.10 (0.03)	20.92 (0.51)	1.11 (0.08)
		SRP	0.80 (0.05)	0.10 (0.02)	20.18 (0.21)	1.14 (0.12)
Sectors	Δ	Softmax	0.67 (0.15)	0.01 (0.01)	17.86 (0.54)	0.80 (0.15)
Sectors	$\mathcal{C}$	O-Proj	0.80 (0.02)	0.01 (0.01)	57.20 (1.03)	6.60 (1.01)
		DC3	0.80 (0.01)	0.01 (0.00)	46.94 (0.85)	9.83 (1.30)
		HardNet	0.82 (0.01)	0.01 (0.00)	35.21 (0.32)	7.95 (1.15)
		RAYEN	0.80 (0.01)	0.01 (0.00)	21.11 (0.57)	1.12 (0.13)
		SRP	0.83 (0.00)	0.00 (0.00)	21.05 (0.38)	1.35 (0.24)

tighter constraint creates a significantly more difficult allocation problem, requiring the distribution of the fleet across all $N = 150$ zones with strictly limited per-zone capacity. For completeness, we also provide results for projection onto the standard simplex. While this baseline serves as a reference point consistent with Table 2, we observe a clear deterioration in performance among the constrained models under this stricter regime, characterized by a lower volume of served customers (*cf.* Table 13). Consistent with the $\kappa = 0.1$ regime, soft-radial projection, RAYEN, DC3, and HardNet all achieve the best observed served rate (0.69), while orthogonal projection attains a lower served rate. These results extend the favorable performance of soft-radial projection to more restrictive capacity constraints.

Table 13: Ride-sharing dispatch results aggregated over five seeds (mean ± std).

Set	Method	Served Rate↑	Training [s]↓	Inference [ms]↓
Δ	Softmax	0.84 (0.00)	13.05 (1.47)	0.36 (0.04)
$\kappa = 0.02$	O-Proj	0.64 (0.00)	74.92 (2.93)	9.65 (0.81)
	DC3	0.69 (0.00)	42.49 (1.35)	12.53 (1.55)
	HardNet	0.69 (0.00)	42.65 (0.55)	10.81 (1.24)
	RAYEN	0.69 (0.00)	16.33 (0.60)	0.85 (0.06)
	SRP	0.69 (0.00)	15.24 (0.37)	0.92 (0.09)

C.7 Network-Constrained Quadratic Programs

To evaluate the layer's capacity to handle structured polyhedral sets, we benchmark on two network-constrained operational tasks: Min-Cost Flow and DC Optimal Power Flow (DC-OPF). Unlike standard simplex projections, feasibility in these domains enforces strict physical conservation laws and asset capacities, yielding complex intersections of affine equalities and inequalities.

Both tasks are formulated as amortized convex quadratic programs. A neural policy observes a temporal window of dynamic cost features and predicts a candidate control vector in reduced coordinates, which the constraint layer subsequently projects into the feasible polytope. This setup strictly isolates the layer's ability to preserve informative gradients across active constraint boundaries while benchmarking the resulting optimality gap, strict feasibility, and forward-pass latency.

C.7.1 Min-Cost Flow

The min-cost flow benchmark represents a capacitated allocation problem on a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The decision variable is the vector of edge flows $x \in \mathbb{R}^m$, where $m = |\mathcal{E}|$. Feasibility requires flow conservation at every node: apart from prescribed sources and sinks, the policy cannot create or destroy mass. It must also respect lower and upper capacity bounds on every edge. Across samples, the graph, supplies, demands, and capacities remain fixed, while the quadratic and linear edge costs vary.

For each sample t, we consider the convex quadratic min-cost flow problem

$$\begin{aligned} \min_{x \in \mathbb{R}^m} \quad & \frac{1}{2} x^\top Q_t x + c_t^\top x \\ \text{s.t.} \quad & \sum_{e \in \delta^+(i)} x_e - \sum_{e \in \delta^-(i)} x_e = b_i, \quad \forall i \in \mathcal{V}, \\ & \underline{x}_e \le x_e \le \overline{x}_e, \quad \forall e \in \mathcal{E}. \end{aligned} \tag{17}$$

Here, $\delta^+(i)$ and $\delta^-(i)$ denote the sets of outgoing and incoming edges at node i. The vector $b \in \mathbb{R}^{|\mathcal{V}|}$ represents node supplies and demands and satisfies $\mathbb{1}^\top b = 0$. The bounds $\underline{x}, \overline{x} \in \mathbb{R}^m$ represent edge capacities. These conditions strictly define the feasible set $\mathcal{C}$.

Sample-dependent objective generation. Let $q^{(0)}, c^{(0)} \in \mathbb{R}^m$ denote base edge costs with $q_e^{(0)} > 0$. To avoid trivial, unstructured instances, we generate a temporally correlated cost sequence using an autoregressive process defined by a global latent factor $g_t = \rho g_{t-1} + \sigma_g \varepsilon_t$ and edge-specific factors $h_{t,e} = \rho h_{t-1,e} + \sigma_h \eta_{t,e}$ for $e \in \mathcal{E}$, where ε_t and $\eta_{t,e}$ are drawn from standard normal distributions. Here, $\rho \in [0, 1)$ controls temporal persistence, while σ_g and σ_h set the volatility of the global and component-specific shocks. The sample-dependent cost coefficients are generated via multiplicative log-normal perturbations, $q_{t,e} = q_e^{(0)} \exp(g_t + h_{t,e})$ and $c_{t,e} = c_e^{(0)} \exp(g_t + h_{t,e})$, with the quadratic term set as $Q_t = \mathrm{diag}(q_t)$. This dynamic construction prevents the network from memorizing a static optimal solution, formulating the task as a true amortized prediction problem.

Observed features and prediction task. Because costs span multiple orders of magnitude and the min-cost flow construction uses positive edge costs, the policy network observes log-transformed quadratic and linear cost coefficients. Given a small numerical stability constant $\epsilon > 0$, we define the stacked feature vector $\phi_t = \big(\log(q_t + \epsilon\mathbb{1}),\ \log(c_t + \epsilon\mathbb{1})\big) \in \mathbb{R}^{2m}$. A neural network maps a lookback window of these features to a candidate flow in reduced coordinates, which the projection layer subsequently maps into $\mathcal{C}$.

Table 14: Min-cost-flow results on synthetic transport graphs with $|\mathcal{V}| \in \{8, 12\}$ nodes, aggregated over five seeds (mean $\pm$ std).

$\lvert\mathcal{V}\rvert$	Method	Rel. gap↓	Training [s]↓	Inference [ms]↓
8	O-Proj	0.10 (0.03)	789.49 (44.37)	456.27 (125.75)
8	RAYEN	1.51 (0.09)	1.72 (0.13)	1.09 (0.14)
8	SRP	1.52 (0.11)	1.52 (0.28)	0.96 (0.19)
12	O-Proj	0.06 (0.04)	1354.76 (65.17)	695.99 (171.10)
12	RAYEN	1.11 (0.09)	1.85 (0.09)	1.08 (0.14)
12	SRP	1.08 (0.09)	2.03 (0.14)	1.09 (0.12)

C.7.2 DC Optimal Power Flow

The DC-OPF benchmark evaluates amortized dispatch using the linearized DC power-flow approximation on standard PGLib test cases [45]. The decision variable $x \in \mathbb{R}^n$ represents the vector of generator outputs. A feasible dispatch must exactly match total demand, respect generation operating ranges, and maintain induced line flows below their thermal limits. These residuals have direct operational meaning: mismatches violate power balance, while line-limit violations correspond to overloaded transmission lines.

Fixed-demand variant. The first variant keeps the PGLib network, base load vector, and all operational limits fixed across samples. The quadratic and linear generation-cost coefficients vary

over time (using the same autoregressive construction as in min-cost flow, with scenario-specific parameter values). For each sample t, we solve

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & \frac{1}{2} x^\top Q_t x + c_t^\top x \\ \text{s.t.} \quad & \mathbb{1}^\top x = D, \\ & \underline{x} \le x \le \overline{x}, \\ & -\overline{f} \le F(Cx - d) \le \overline{f}. \end{aligned} \tag{18}$$

Here, $D \in \mathbb{R}$ denotes total demand, $C \in \mathbb{R}^{b \times n}$ maps generation to nodal injections, $d \in \mathbb{R}^b$ is the demand vector, $F \in \mathbb{R}^{\ell \times b}$ maps nodal injections to line flows, and $\overline{f} \in \mathbb{R}^\ell$ represents line capacities.

Table 15: DC-OPF results aggregated over five seeds (mean ± std).

Scenario	Method	Rel. gap↓	Training [s]↓	Inference [ms]↓
14 Bus	O-Proj	0.00 (0.00)	12.88 (1.80)	13.60 (3.40)
14 Bus	RAYEN	0.00 (0.00)	4.62 (0.84)	3.19 (1.45)
14 Bus	SRP	0.00 (0.00)	4.90 (0.96)	3.07 (1.29)
118 Bus	O-Proj	0.01 (0.00)	7473.10 (106.65)	2821.26 (64.62)
118 Bus	RAYEN	0.03 (0.00)	1.94 (0.08)	0.68 (0.06)
118 Bus	SRP	0.03 (0.00)	3.67 (1.18)	0.85 (0.09)

Dynamic-demand variant. To test stability under a dynamic feasible polytope, we allow the total load level to vary over time. The base demand vector is scaled as $d_t = s_t d$, where the load scale s_t follows a clipped log-autoregressive process $\log s_t = \rho \log s_{t-1} + \sigma \xi_t$, bounded within the interval $[\underline{s}, \bar{s}]$. Here, ρ controls persistence in aggregate demand and σ controls the magnitude of load fluctuations. If an extreme sampled scale does not admit a strict interior point, we linearly revert it toward the nominal scale $s = 1$ until strict feasibility is recovered. Consequently, the feasible set $\mathcal{C}_t$ becomes sample-dependent in both the power-balance and line-flow constraints.

The policy observes log-transformed quadratic costs and signed-log-transformed linear costs, augmented with $\log s_t$ in the dynamic-demand setting. We compute a strictly feasible particular point $x_p(s_t)$ to parameterize the equality set as $x = x_p(s_t) + B\xi$, where the columns of $B \in \mathbb{R}^{n \times (n-1)}$ form a basis for the null space of the power-balance row $\mathbb{1}^\top$ and $\xi \in \mathbb{R}^{n-1}$ denotes reduced coordinates. For radial methods, this point also serves as the ray anchor; for projection and repair methods, it serves only as the affine offset. To avoid solving an anchor problem at every training step, these strict particular points are precomputed over a uniform grid (4 points) of load scales and linearly interpolated during training.

Table 16 reports the dynamic-demand IEEE 14 Bus case for RAYEN and SRP using a 4-point anchor grid. We omit orthogonal projection in this dynamic variant because each changing feasible polytope would require solving a separate projection problem at inference time, making the comparison dominated by solver overhead rather than by the geometry of the learned feasibility layer. Both methods remain feasible across the test set and achieve comparable objective gaps, while retaining substantially lower latency than the orthogonal-projection baseline in the fixed-demand setting.

Table 16: Dynamic-demand DC-OPF results aggregated over five seeds (mean ± std).

Scenario	Method	Rel. gap↓	Training [s]↓	Inference [ms]↓	Feas. Rate	Anchor Precompute [ms]↓
14 Bus	RAYEN	0.01 (0.00)	2.50 (1.78)	1.88 (1.61)	1.00 (0.00)	5.79 (0.51)
14 Bus	SRP	0.00 (0.00)	3.38 (2.63)	2.48 (1.80)	1.00 (0.00)	7.87 (4.84)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 6.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please refer to Sections 2–4, as well as the Appendix B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Details in Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The source code is accessible at `https://anonymous.4open.science/r/soft-radial-projection-e2e-neurips/`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Refer to Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work addresses the fundamental problem of modeling optimization problems subject to constraints.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not pose a risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The data and frameworks used in this paper are cited; see Section 5 and Appendix C.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide an anonymized repository containing the code used to validate our results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The research does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The methods of this research do not involve LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.