Not All Objectives Are Born Equal: Priority-Constrained Descent for Hierarchical Multi-Objective Optimization

Anonymous authors
Paper under double-blind review

Abstract

Deep learning problems rarely involve objectives that are equal in importance. A primary objective defines the goal, whilst secondary objectives, such as sparsity, compression, or robustness constrain the solution. While existing multi-objective methods have proven effective in practice, they have a clear symmetry problem and neglect the inherent objective hierarchy built into these objective spaces. We introduce Priority-Constrained Descent (PCD), a gradient-based optimization framework designed to explicitly exploit hierarchical objective structures. PCD preserves the direction of primary descent whilst allowing for the minimal distortion necessary to guarantee progress on secondary objectives, controlled by a single $\tau \in [0, 1]$ that dictates the strength of the distortion. The resulting formulation is invariant to objective scaling and admits exact closed-form solutions for problems with two and three objectives. We evaluate PCD within structured network compression settings, unstructured sparsity and low-rankness, and across a variety of synthetic experiments, showing Pareto dominance and better per-objective performance with secondary progress guarantees over existing methods, further exhibiting the interpretable trade-off that τ provides.

1 Introduction

Optimization in modern deep learning is fundamentally hierarchical. While models are typically evaluated based on their primary objective performance (such as accuracy in the case of classification, mean squared error for regression, or fidelity for generative modeling), solutions often need to satisfy a number of *secondary* conditions to be usable in practice. In the case of pruning, the primary task loss is paired with a sparsity-inducing regularizer that heavily constrains the final parameterization (He & Xiao, 2023). Low-rankness pairs the primary objective with a rank-constraining regularizer to force the network parameters into a low-dimensional subspace (Idelbayev & Carreira-Perpinán, 2020). Quantization-aware training anchors the task against a fidelity term penalizing drift from discrete values (Abushahla et al., 2025). Similarly, continual learning balances fitness on new data against a replay objective preserving past knowledge (Wang et al., 2024), while federated learning manages local accuracy alongside communication constraints (Zhang et al., 2021). In all these examples, the structure of the optimization problem is the same: the relationship between objectives is asymmetric by design. One loss represents the core capability, while the others dictate the boundaries within which that capability must be learned. Nonetheless, standard multi-objective optimization (MOO) frameworks ignore this structural truth.

The conventional paradigm for handling multiple objectives is to collapse all objectives into a single weighted summation $\mathcal{L} = \sum_i w_i \mathcal{L}_i$ (Marler & Arora, 2010). Here, importance is encoded implicitly through the weights w_i, but this requires fragile hyperparameter sweeps and struggles to access non-convex regions of the Pareto front. Gradient-based methods (Désidéri, 2009; Yu et al., 2020; Liu et al., 2021a; 2023) avoid these pitfalls but share the issue of treating objectives symmetrically. Consequently, they target Pareto stationarity, where no objective can be improved without harming others. For hierarchical problems, we argue that this is too restrictive of a criterion. It can lead to what we term a *conflict equilibrium*, a sub-optimal state where progress on the primary objective is halted because its gradient is counterbalanced by the push-back of constraints with equal authority. This is a failure mode and not a feasible solution for practitioners.

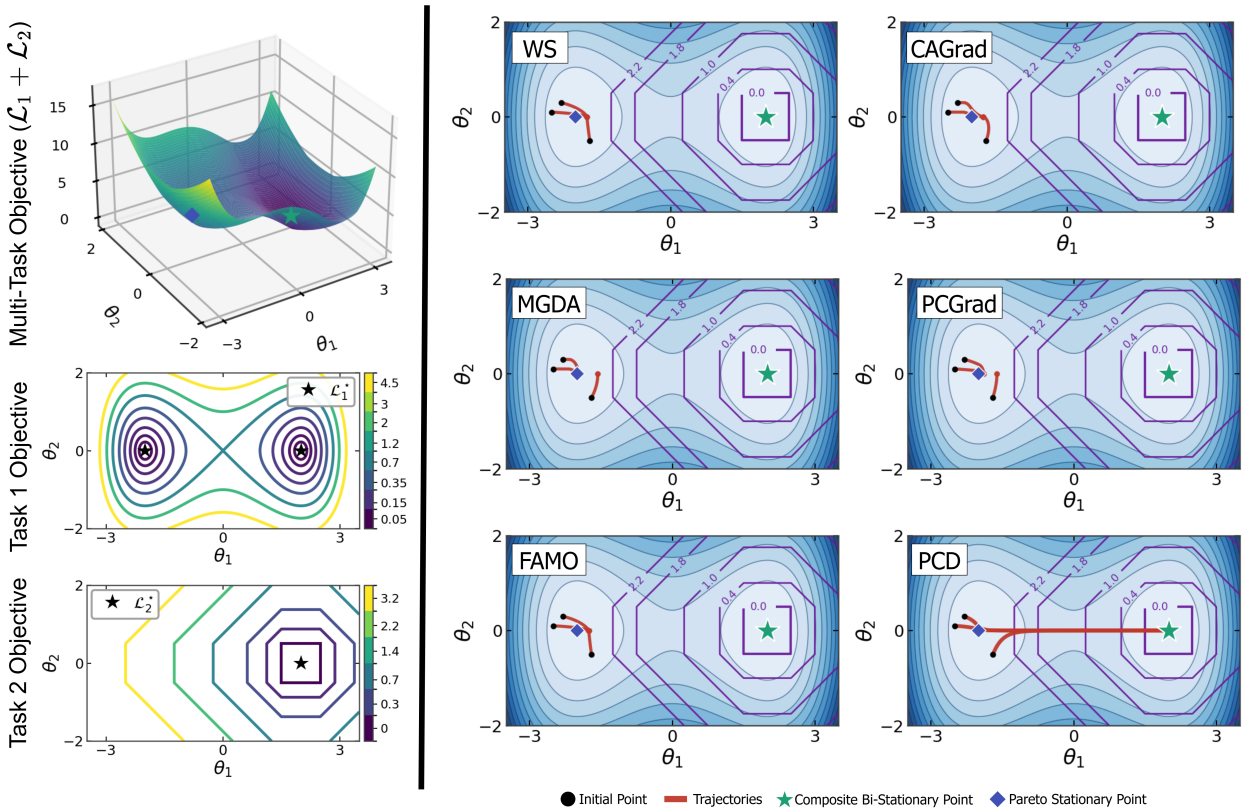


Figure 1: PCD on a 2D toy problem with two local minima of $\mathcal{L}_1$, only one of which is also a minimum of $\mathcal{L}_2$. From three initializations, weighted sum, MGDA, PCGrad, CAGrad, and FAMO all terminate at a Pareto-stationary point at which $\mathcal{L}_2$ remains non-zero (the conflict equilibrium of Sec. 3). From all three initializations, PCD crosses the saddle barrier in parameter space and reaches the composite bi-stationary point where $\nabla\mathcal{L}_1 = \nabla\mathcal{L}_2 = 0$, which is a minimizer for both $\mathcal{L}_1$ and $\mathcal{L}_2$.

What hierarchical optimization problems call for is an optimizer that imposes priority at the gradient level, avoiding the issue of halting prematurely by overly enforcing secondary objectives. To this end, we introduce Priority-Constrained Descent (PCD), a framework designed to follow the primary gradient as closely as possible, deflecting only as much as necessary to keep secondary objectives on track. At each step, PCD computes the minimum Euclidean deviation in parameter space that guarantees a fraction of descent on secondary objectives. A single scalar tolerance $\tau \in [0, 1]$ controls this guarantee. The resulting update anchors strictly to the primary task, adding only a non-negative combination of the secondary gradients whose constraints are currently active. As a result, PCD can never be trapped in conflict equilibria. Its updates vanish only at *composite multi-stationary* (CMS) points, a strictly stronger condition than Pareto stationarity where every objective is individually stationary.

This geometric advantage is immediately visible in a 2D landscape (Fig. 1). When faced with a saddle barrier separating a local and a global minimum, all symmetric baselines stall at a Pareto-stationary conflict equilibrium where $\mathcal{L}_2$ remains non-zero. PCD is the only method that pushes through the secondary constraints to reach the composite bi-stationary point where $\nabla\mathcal{L}_1 = \nabla\mathcal{L}_2 = 0$, successfully minimizing both objectives.

We summarize the core contributions of this study as follows:

- Priority-Constrained Descent: A gradient-based optimizer that solves a small convex quadratic program (QP) to minimally distort the primary gradient, enabling progress for the primary objective whilst *guaranteeing* progress on secondaries. We derive exact closed-form solutions for 2 and 3 ob-

jectives, prove invariance to per-objective scaling, and show the update reduces to standard gradient descent when secondary constraints are naturally satisfied.

- Structured Network Compression: We show the dominance of PCD over existing MOO frameworks across a variety of architectures and datasets for structured pruning tasks, showing minimal drop in accuracy compared to unpruned baselines. On CIFAR-100 trained on the ResNet-34 architecture, PCD maintains a 70.9% accuracy at 90% sparsity (compared to an unpruned initial accuracy of 73.3%), where baselines collapse.

- Unstructured Pruning and Low-Rankness: PCD dominates the tri-objective space for experiments with both ℓ_1 and nuclear norms, maintaining accuracies of 71% for CIFAR-100 trained on ResNet-34 with an unstructured sparsity of 92.6% and an effective rank of 24.5, whilst baselines once again collapse in terms of accuracy and over-compress.

- Synthetic Verification: In known objective spaces, we confirm the advantages of PCD, showing scale invariance, feasibility, and scale up to larger number of objectives, displaying its ability to find CMS points without running into issues with infeasibility.

2 Related Works

Modern gradient-based MOO rests on a target established by Gale et al. (1951). A point θ^* is *Pareto stationary* if the origin lies in the convex hull of the objective gradients, meaning no common descent direction exists. Traditional MOO frameworks aim at reaching a Pareto-stationary point, with more nuanced techniques allowing for some degree of preference over the Pareto front itself. In the pursuit of Pareto stationarity, the field has branched into two distinct strategies: scalarization and direct gradient manipulation.

Classical weighted-sum scalarization (Marler & Arora, 2010) collapses the objective space prior to optimization. While any stationary point with strictly positive weights (assigned to each objective) is technically Pareto stationary, this approach cannot reach non-convex regions of the Pareto front (Miettinen, 1999) and requires brittle hyperparameter tuning. A recent study reports that a well-tuned weighted sum frequently matches the performance of more elaborate alternatives, suggesting that the gains attributed to complex MOO algorithms are often artifacts of the training process rather than fundamental algorithmic superiority (Xin et al., 2022). Regardless, to bypass these pathologies, Fliege & Svaiter (2000) proposed a second strategy: discarding scalar weights entirely to compute the steepest Pareto-descent direction directly from the gradients. Nearly every modern gradient-based MOO method inherits this direct-direction strategy. However, while they successfully escape scale sensitivity and access broader regions of the Pareto front, they preserve a structural assumption that contradicts the reality of hierarchical deep learning: that objectives are symmetric and inherently interchangeable.

In Pursuit of Symmetric Compromise. The direct-direction lineage largely stems from MGDA (Désidéri, 2009), which operationalizes Pareto stationarity by identifying the minimum-norm element of the gradients' convex hull. Sener & Koltun (2018) scaled MGDA to deep learning via a Frank–Wolfe solver (Jaggi, 2013), while MoCo (Fernando et al., 2023; 2024) mitigated the gradient bias inherent to stochastic MGDA variants, recovering provable convergence in non-convex settings. In parallel, a separate family of methods takes the gradients as given and attempts to condition them before they are combined. Magnitude-based approaches balance objectives by scaling losses via learned homoscedastic noise (Kendall et al., 2018), driving gradient norms to a common rate (Chen et al., 2018), ensuring equal gradient projections (Liu et al., 2021b), or tracking loss histories to balance step decreases (Liu et al., 2023). Geometry-based approaches attempt to resolve gradient misalignment by projecting conflicting task gradients onto each other's normal planes (Yu et al., 2020), or by finding an optimal direction within an averaged-gradient ball (Liu et al., 2021a; Senushkin et al., 2023).

However, even when methods attempt to prioritize certain outcomes, they do so symmetrically. Preference-steering methods like EPO (Mahapatra & Rajan, 2021), PMGDA (Zhang et al., 2025), and hypernetwork adaptations (Navon et al., 2020) target specific Pareto-optimal points via controlled ascent or test-time

network mapping. Yet, the preference vector enters these formulations as a symmetric annotation on the objectives, not as a structural distinction. Across all these methods, whether modifying directions, magnitudes, or geometries, the fundamental goal is compromise. As noted earlier, enforcing this symmetric compromise in hierarchical settings is precisely what traps optimizers in sub-optimal conflict equilibria.

Breaking The Symmetry. The shift away from symmetric MOO has begun to surface in recent game-theoretic and constrained-optimization literature. Nash MTL (Navon et al., 2022) applies the Nash Bargaining solution to produce proportionally fair update directions. Building on this, AuxiNash (Shamsian et al., 2023) marks a critical conceptual pivot: it extends Nash-MTL to asymmetric settings by learning task-specific bargaining power based on each auxiliary's contribution to the primary objective. It explicitly names the limitation that symmetric methods treat structurally unequal objectives as peers. Similarly, constrained-optimization frameworks like Cooper (Gallego-Posada et al., 2025) establish hierarchy through Lagrangian multiplier dynamics, updating learned dual variables alongside primal parameters.

While AuxiNash and Cooper recognize the asymmetric reality of modern deep learning, they enforce this hierarchy through learned bargaining weights and Lagrangian dynamics, respectively. PCD discards the symmetry assumption at a much lower level. Rather than relying on learned coefficients, PCD embeds hierarchy directly into the geometry of the update. By calculating the closed-form Euclidean projection of the primary gradient onto the secondary-progress polyhedron, PCD yields a strict, per-step descent guarantee on secondary objectives. It fundamentally rejects the symmetric compromise of Pareto stationarity, driving the optimization instead toward the strictly stronger endpoint of composite multi-stationarity.

3 Preliminaries and Definitions

3.1 Problem Setup

We consider optimization problems over a parameter space $\boldsymbol{\theta} \in \mathbb{R}^n$ with $K \geq 2$ objectives arranged in an asymmetric hierarchy:

$$\underbrace{\mathcal{L}_1(\boldsymbol{\theta})}_{\text{primary}}, \quad \underbrace{\mathcal{L}_2(\boldsymbol{\theta}), \ldots, \mathcal{L}_K(\boldsymbol{\theta})}_{\text{secondary}}. \tag{1}$$

We use $[K] = \{1, \ldots, K\}$ and $[K]_{-1} = \{2, \ldots, K\}$ for index sets, $\|\cdot\|$ for the Euclidean norm, $\partial\mathcal{L}$ for the convex subdifferential, $\operatorname{conv}\{\cdot\}$ for the convex hull, and $\operatorname{cone}\{\cdot\}$ for the conic hull.

Assumption 3.1 (Objective regularity)**.**

(i) Primary: $\mathcal{L}_1 : \mathbb{R}^n \to \mathbb{R}$ *is continuously differentiable with gradient* $\mathbf{g}_1 = \nabla\mathcal{L}_1(\boldsymbol{\theta})$.

(ii) Secondary: *For each* $j \in [K]_{-1}$, $\mathcal{L}_j : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *is proper, convex, and lower semi-continuous, with subgradient* $\mathbf{g}_j \in \partial\mathcal{L}_j(\boldsymbol{\theta})$ *available at every* $\boldsymbol{\theta} \in \operatorname{dom}(\mathcal{L}_j)$.

This assumption encompasses a broad class of convex objectives frequently encountered in optimization and machine learning, including both smooth and non-smooth formulations. More generally, any proper, convex, lower semi-continuous objective with accessible subgradients falls within the framework. In particular, the secondary objectives may represent structural regularizers, auxiliary convex penalties, or constraint-like terms commonly used in modern machine learning, such as the ℓ_1 norm, Group Lasso, nuclear norm, total variation, and hinge-type losses. The convexity assumption ensures a well-defined subgradient geometry and preserves the convex feasibility structure underlying the PCD projection step.

3.2 Stationarity Concepts

Definition 3.2 (Pareto stationarity)**.** *A point* $\boldsymbol{\theta}$ *is* Pareto stationary *if there exists no common ascent direction* $\mathbf{d} \in \mathbb{R}^n$ *satisfying* $\mathbf{g}_i^\top \mathbf{d} > 0$ *for all* $i \in [K]$ *simultaneously, where* $\mathbf{g}_1 = \nabla\mathcal{L}_1(\boldsymbol{\theta})$ *and* $\mathbf{g}_i \in \partial\mathcal{L}_i(\boldsymbol{\theta})$ *for* $i \geq 2$. *Equivalently,*

$$\mathbf{0} \in \operatorname{conv}\{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_K\}. \tag{2}$$

Definition 3.3 (Composite multi-stationarity)**.** *A point $\boldsymbol{\theta}$ is* composite multi-stationary *(CMS) with respect to the hierarchy $(\mathcal{L}_1; \mathcal{L}_2, \ldots, \mathcal{L}_K)$ if*

$$\nabla \mathcal{L}_1(\boldsymbol{\theta}) = \mathbf{0} \quad \textit{and} \quad \mathbf{0} \in \partial \mathcal{L}_j(\boldsymbol{\theta}) \quad \textit{for all } j \in [K]_{-1}. \tag{3}$$

For $K = 2$, we use composite bi-stationarity *(CBS).*

CMS asks that every objective be individually stationary at $\boldsymbol{\theta}$. Pareto stationarity, by contrast, admits *conflict equilibria* at which competing gradients cancel within the convex hull (Eq. (2)) even though no individual $\mathbf{g}_i$ vanishes. Every CMS point is Pareto stationary, but not conversely: in dimension one, $\mathcal{L}_1(\theta) = \frac{1}{2}\theta^2$ and $\mathcal{L}_2(\theta) = \frac{1}{2}(\theta - 1)^2$ at $\theta = \frac{1}{2}$ have gradients $\frac{1}{2}$ and $-\frac{1}{2}$ that cancel in convex combination, so the point is Pareto stationary, but $\nabla \mathcal{L}_1(\frac{1}{2}) = \frac{1}{2} \neq 0$, so it is not CMS. Conflict equilibria of this kind are exactly the points at which symmetric MOO methods such as MGDA (Désidéri, 2009) can halt without making any single objective stationary, and exactly the regime PCD's QP is designed to rule out by construction (Sec. 4.8).

4 Priority-Constrained Descent

Based on the principle of anchoring to the primary gradient and admitting corrections along the secondaries to ensure progress, we now present the algorithmic framework for PCD. This framing of "corrections" should be carefully defined: in Euclidean distance, it is simply the resultant direction that lies as close as possible to the primary gradient vector. Coupled with the role of the tolerance parameter $\tau \in [0, 1]$, we formalize the PCD update rule in this section. The development proceeds as follows: Sec. 4.1 states the QP and identifies its solution as the Euclidean projection of the primary gradient onto a polyhedron of secondary-progress half-spaces. Sec. 4.2 extracts the KKT identity that makes the asymmetric structure of the update explicit, and Sec. 4.3 draws the per-step secondary-progress guarantee directly from it. Sec. 4.4 then introduces the gradient normalization that gives τ a scale-invariant interpretation across objectives of vastly different magnitude. Sec. 4.5 specializes to two objectives, where the QP admits a closed-form solution and τ's geometric meaning becomes quantitative through an analytical threshold at which primary descent breaks down, and Sec. 4.6 states the full algorithm. Sec. 4.7 delimits the scope of the guarantees, and Sec. 4.8 closes with the endpoint characterization that strictly rules out conflict equilibria. Comparison with existing gradient-based MOO methods is deferred to Sec. 4.9.

4.1 The PCD Quadratic Program

Definition 4.1 (PCD quadratic program)**.** *Given normalized gradients $\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_K$ and a scalar tolerance $\tau \in [0, 1]$, the PCD update direction is*

$$\tilde{\mathbf{d}}^* \;=\; \arg\min_{\tilde{\mathbf{d}} \in \mathbb{R}^n} \left\| \tilde{\mathbf{d}} - \tilde{\mathbf{g}}_1 \right\|^2 \quad s.t. \quad \tilde{\mathbf{g}}_j^\top \tilde{\mathbf{d}} \;\geq\; \tau \left\| \tilde{\mathbf{g}}_j \right\|^2 \;\; \forall\, j \in [K]_{-1}. \tag{4}$$

Among all directions guaranteeing at least a τ-fraction of maximum normalized progress on every secondary, $\tilde{\mathbf{d}}^*$ is the one closest in Euclidean distance to the primary gradient $\tilde{\mathbf{g}}_1$. The secondary objectives thus act as *half-space constraints* on the primary update, not as symmetric competitors. The feasible set

$$F_\tau(\boldsymbol{\theta}) \;=\; \bigcap_{j=2}^{K} H_j, \qquad H_j = \{\tilde{\mathbf{d}} \in \mathbb{R}^n : \tilde{\mathbf{g}}_j^\top \tilde{\mathbf{d}} \;\geq\; \tau \left\| \tilde{\mathbf{g}}_j \right\|^2\}, \tag{5}$$

is a closed convex polyhedron, possibly empty. For $K = 2$ with $\tilde{\mathbf{g}}_2 \neq \mathbf{0}$, $F_\tau(\boldsymbol{\theta})$ is nonempty for every $\tau \in [0, 1]$. For $K > 2$, infeasibility can occur when the secondary-progress constraints are mutually incompatible. Geometrically, when $\mathbf{0} \in \operatorname{cone}\{\tilde{\mathbf{g}}_2, \ldots, \tilde{\mathbf{g}}_K\} - \{\mathbf{0}\}$. This is an intrinsic limitation of imposing simultaneous directional-progress requirements rather than a failure of the solver. The full feasibility analysis via Farkas' lemma (Dinh & Jeyakumar, 2014), together with a sufficient pairwise-angle condition for $K > 2$, is given in App. B.1.

4.2 Geometry and KKT Structure

Eq. (4) is a convex QP with affine inequality constraints. Its solution admits a geometric characterization that locates $\tilde{\mathbf{d}}^*$ in space.

Theorem 4.2 (Projection and KKT characterization)**.** *Suppose $F_\tau(\boldsymbol{\theta}) \neq \varnothing$. Then $\tilde{\mathbf{d}}^*$ is the Euclidean projection of the primary gradient onto the secondary-progress feasible set,*

$$\tilde{\mathbf{d}}^* \;=\; \mathrm{Proj}_{F_\tau(\boldsymbol{\theta})}(\tilde{\mathbf{g}}_1). \tag{6}$$

Equivalently, there exist multipliers $\mu_j^* \geq 0$ ($j \in [K]_{-1}$) such that

$$\tilde{\mathbf{d}}^* \;=\; \tilde{\mathbf{g}}_1 \;+\; \sum_{j=2}^{K} \mu_j^* \, \tilde{\mathbf{g}}_j, \tag{7}$$

with primal feasibility $\tilde{\mathbf{g}}_j^\top \tilde{\mathbf{d}}^* \geq \tau \|\tilde{\mathbf{g}}_j\|^2$ and complementary slackness $\mu_j^*\big(\tilde{\mathbf{g}}_j^\top \tilde{\mathbf{d}}^* - \tau \|\tilde{\mathbf{g}}_j\|^2\big) = 0$ for all $j \in [K]_{-1}$.

Proof sketch. The projection statement is immediate: Eq. (4) minimizes squared Euclidean distance from $\tilde{\mathbf{g}}_1$ over the closed convex set $F_\tau(\boldsymbol{\theta})$, which is by definition the projection. The KKT form follows from the Lagrangian

$$\mathcal{L}(\tilde{\mathbf{d}}, \boldsymbol{\mu}) \;=\; \tfrac{1}{2}\left\|\tilde{\mathbf{d}} - \tilde{\mathbf{g}}_1\right\|^2 \;-\; \sum_{j=2}^{K} \mu_j\big(\tilde{\mathbf{g}}_j^\top \tilde{\mathbf{d}} - \tau \|\tilde{\mathbf{g}}_j\|^2\big), \qquad \mu_j \geq 0,$$

whose stationarity in $\tilde{\mathbf{d}}$ gives $\tilde{\mathbf{d}} - \tilde{\mathbf{g}}_1 - \sum_j \mu_j \tilde{\mathbf{g}}_j = \mathbf{0}$, hence Eq. (7). Primal and dual feasibility and complementary slackness are the standard KKT conditions for inequality-constrained convex QPs with affine constraints; no further constraint qualification is required. The active-set determination of $\{\mu_j^*\}$ is detailed in App. B.2. □

The Geometric Picture. The projection identity of Eq. (6) makes precise the "minimum Euclidean distortion from the primary gradient" framing carried throughout this work. The primary gradient $\tilde{\mathbf{g}}_1$ sits somewhere in $\mathbb{R}^n$. The feasible polyhedron $F_\tau(\boldsymbol{\theta})$ is the intersection of the $K-1$ half-spaces of admissible directions, one per secondary, with boundaries determined by τ. PCD locates $\tilde{\mathbf{d}}^*$ as the unique point of $F_\tau(\boldsymbol{\theta})$ closest to $\tilde{\mathbf{g}}_1$ in Euclidean distance. Two regimes are immediate from this picture. If $\tilde{\mathbf{g}}_1 \in F_\tau(\boldsymbol{\theta})$ already, no projection is needed and $\tilde{\mathbf{d}}^*$ coincides with $\tilde{\mathbf{g}}_1$ exactly. We record this as:

Corollary 4.3 (Inactive regime)**.** *If $\tilde{\mathbf{g}}_j^\top \tilde{\mathbf{g}}_1 \geq \tau \|\tilde{\mathbf{g}}_j\|^2$ for every $j \in [K]_{-1}$, then $\tilde{\mathbf{d}}^* = \tilde{\mathbf{g}}_1$, and PCD reduces to gradient descent on $\mathcal{L}_1$.*

The secondaries do not perturb the primary at all when they do not have to: PCD reverts to standard gradient descent on $\mathcal{L}_1$ precisely when the primary direction itself already provides sufficient progress on every secondary. Otherwise $\tilde{\mathbf{g}}_1$ lies outside $F_\tau(\boldsymbol{\theta})$ and the projection lands on the boundary of the polyhedron at the closest boundary point, on the face determined by the secondaries whose constraint is currently binding: the *active set* $A(\boldsymbol{\theta}) \subseteq [K]_{-1}$. As τ grows, the half-spaces H_j recede from the origin in the direction of each $\tilde{\mathbf{g}}_j$, the polyhedron $F_\tau(\boldsymbol{\theta})$ shrinks, and $\tilde{\mathbf{d}}^*$ is pushed further from $\tilde{\mathbf{g}}_1$ along faces parametrized by the active set. This is a monotone deformation whose two-objective realization is shown in Fig. 2.

This characterization makes the asymmetric structure of PCD's update explicit. Symmetric MOO methods produce updates of the form $\sum_i \alpha_i \mathbf{g}_i$, in which the primary and secondary coefficients are determined by the same rule: convex-hull weights for MGDA, fixed weights for weighted sum, projection coefficients for PCGrad and CAGrad. In PCD, the primary's coefficient is structurally pinned to one and the secondaries' coefficients are non-negative, possibly zero, and determined by which constraints currently bind. The two roles (anchor and correction) are visible in the construction itself, rather than recovered from it. The full contrast against symmetric methods is detailed in Sec. 4.9.

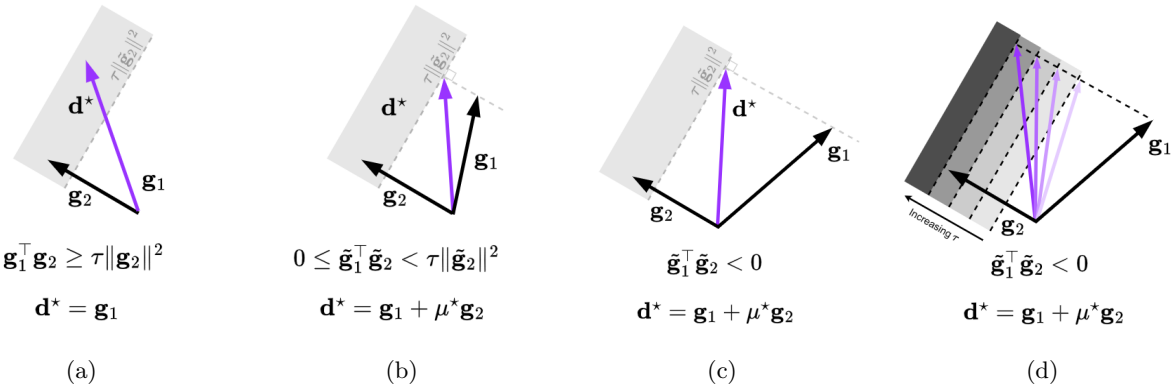


Figure 2: The four canonical regimes of PCD for $K = 2$, as predicted by Cor. 4.6. **(a) Non-conflicting, sufficient progress**: the constraint is inactive and $\mathbf{d}^\star = \mathbf{g}_1$. **(b) Non-conflicting, insufficient progress**: the constraint is active and $\mathbf{d}^\star$ projects onto its boundary, with $\mathbf{d}^\star = \mathbf{g}_1 + \mu^\star\mathbf{g}_2$. **(c) Conflicting**: the constraint is active and the same projection formula applies, with $\mu^\star$ strictly larger than in **(b)** due to the additional negative term. **(d)** The effect of increasing τ at fixed gradient configuration: the feasible set shrinks (Sec. 4.5, τ-monotonicity), and $\mathbf{d}^\star$ moves further from $\mathbf{g}_1$.

4.3 Secondary-Progress Guarantee

The projection characterization yields the central per-step guarantee directly. At each iterate, PCD produces an update direction $\mathbf{d}^*(\boldsymbol{\theta})$ and applies

$$\boldsymbol{\theta}^+ \;=\; \boldsymbol{\theta} \;-\; \eta\,\mathbf{d}^*(\boldsymbol{\theta}). \tag{8}$$

Under this sign convention, the first-order change in any objective along the update is $\mathcal{L}_j(\boldsymbol{\theta}^+) = \mathcal{L}_j(\boldsymbol{\theta}) - \eta\,\mathbf{g}_j^\top\mathbf{d}^*(\boldsymbol{\theta}) + O(\eta^2)$, so $\mathbf{g}_j^\top\mathbf{d}^* > 0$ corresponds to first-order *descent* of $\mathcal{L}_j$ for $j \geq 2$. We treat $\mathbf{d}$ throughout as a *gradient-like* direction whose role is to make $\mathbf{g}_j^\top\mathbf{d}$ positive. This is the opposite of the descent-direction convention used classically, and we make it explicit because the following inequality depends on it.

Proposition 4.4 (Per-step secondary progress)**.** *Suppose $F_\tau(\boldsymbol{\theta}) \neq \varnothing$ and $\tau > 0$. The PCD direction satisfies*

$$\tilde{\mathbf{g}}_j^\top\tilde{\mathbf{d}}^* \;\geq\; \tau\,\|\tilde{\mathbf{g}}_j\|^2 \quad \textit{for all } j \in [K]_{-1}. \tag{9}$$

Consequently, under the idealized iteration $\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta\tilde{\mathbf{d}}^*$ with $\mathcal{L}_j$ Lipschitz-smooth (or with a Lipschitz continuous subgradient selection),

$$\mathcal{L}_j(\boldsymbol{\theta}^+) \;\leq\; \mathcal{L}_j(\boldsymbol{\theta}) \;-\; \eta\,s_j^{-1}\tau\,\|\tilde{\mathbf{g}}_j\|^2 \;+\; O(\eta^2), \tag{10}$$

so for $\tilde{\mathbf{g}}_j \neq \mathbf{0}$ and sufficiently small η, $\mathcal{L}_j$ strictly decreases at this step.

Proof. Eq. (9) is the primal feasibility condition of Eq. (4). For Eq. (10), the first-order Taylor expansion gives $\mathcal{L}_j(\boldsymbol{\theta}^+) = \mathcal{L}_j(\boldsymbol{\theta}) - \eta\mathbf{g}_j^\top\tilde{\mathbf{d}}^* + O(\eta^2)$, and $\mathbf{g}_j^\top\tilde{\mathbf{d}}^* = s_j^{-1}\tilde{\mathbf{g}}_j^\top\tilde{\mathbf{d}}^* \geq s_j^{-1}\tau\,\|\tilde{\mathbf{g}}_j\|^2$. □

Proposition 4.4 is the practical content of PCD: every step guarantees *tolerance-controlled* first-order descent on every non-stationary secondary. This is structurally unlike PCGrad (which carries no per-objective progress guarantee in either the conflicting or non-conflicting regime), unlike MGDA (which guarantees descent on the minimum-norm convex hull element, an aggregate quantity), and unlike weighted sum (which guarantees descent only on the scalarized loss). All three are detailed in Sec. 4.9.

4.4 Gradient Normalization

Gradient magnitudes from different objectives may differ by orders of magnitude and vary throughout training. To ensure that the tolerance parameter τ has a consistent, scale-invariant interpretation, PCD normalizes all gradients via exponential moving averages (EMA).

Definition 4.5 (EMA normalization)**.** *Fix $\beta \in (0,1)$ and $\varepsilon > 0$. For each $i \in [K]$ and iteration t:*

$$v_{i,t} = \beta\, v_{i,t-1} + (1-\beta)\, \|\mathbf{g}_{i,t}\|^2, \quad \hat{v}_{i,t} = \tfrac{v_{i,t}}{1-\beta^t}, \quad s_{i,t} = \tfrac{1}{\sqrt{\hat{v}_{i,t}+\varepsilon}}, \quad \tilde{\mathbf{g}}_{i,t} = s_{i,t}\, \mathbf{g}_{i,t}, \tag{11}$$

with $v_{i,0} = 0$.

In steady state ($t \gg \dfrac{1}{1-\beta}$), $\hat{v}_{i,t} \approx \mathbb{E}[\|\mathbf{g}_i\|^2]$, so $\|\tilde{\mathbf{g}}_i\| \approx 1$. Hence a constraint of the form $\tilde{\mathbf{g}}_j^\top \tilde{\mathbf{d}} \geq \tau \|\tilde{\mathbf{g}}_j\|^2$ demands a τ-fraction of the maximum normalized secondary progress, with τ directly interpretable on $[0,1]$ regardless of the raw magnitudes of $\mathcal{L}_2, \ldots, \mathcal{L}_K$.

Scale invariance. Under per-objective rescaling $L_i \to c_i L_i$ ($c_i > 0$), the EMA buffers transform as $\hat{v}_{i,t} \to c_i^2 \hat{v}_{i,t}$, so $s_{i,t} \to s_{i,t}/c_i$ in the limit $\varepsilon \to 0$ (or once $\hat{v}_{i,t} \gg \varepsilon/c_i^2$), and the normalized gradients $\tilde{\mathbf{g}}_i = s_{i,t}\mathbf{g}_i$ are unchanged. Consequently, the QP in Eq. (4) and its solution $\tilde{\mathbf{d}}^*$ are invariant in the same limit. By contrast, $\mathbf{d}_{\mathrm{WS}} = \sum_i w_i \mathbf{g}_i$ transforms to $\sum_i w_i c_i \mathbf{g}_i$, so its optimal weight ratio depends on the relative magnitudes of the raw objectives. Scale invariance is a structural advantage of PCD whenever the raw L_i differ by orders of magnitude (typical when a task loss is paired with a structural regularizer). The exact finite-ε residual and an ε-free normalization for which invariance is exact are in App. B.6.

4.5 Closed Form Solutions and the Role of τ

For $K = 2$, Eq. (4) is a projection onto a single half-space, and Theorem 4.2 reduces to a closed-form update that exposes the geometric meaning of τ.

Corollary 4.6 ($K = 2$ closed form)**.** *For $K = 2$ with $\tilde{\mathbf{g}}_2 \neq \mathbf{0}$,*

$$\tilde{\mathbf{d}}^* = \begin{cases} \tilde{\mathbf{g}}_1 & \text{if } \tilde{\mathbf{g}}_2^\top \tilde{\mathbf{g}}_1 \geq \tau \|\tilde{\mathbf{g}}_2\|^2 \quad \text{(constraint inactive)}, \\ \tilde{\mathbf{g}}_1 + \left(\tau - \dfrac{\tilde{\mathbf{g}}_2^\top \tilde{\mathbf{g}}_1}{\|\tilde{\mathbf{g}}_2\|^2}\right)\tilde{\mathbf{g}}_2 & \text{otherwise (constraint active).} \end{cases} \tag{12}$$

In the active regime, PCD adds *exactly* the component along $\tilde{\mathbf{g}}_2$ needed to satisfy the constraint at equality, and no more. Geometrically, $\tilde{\mathbf{d}}^*$ lands on the boundary $\{\tilde{\mathbf{d}} : \tilde{\mathbf{g}}_2^\top \tilde{\mathbf{d}} = \tau \|\tilde{\mathbf{g}}_2\|^2\}$ at the point closest to $\tilde{\mathbf{g}}_1$. The four canonical regimes of Eq. (12) as $\tilde{\mathbf{g}}_1^\top \tilde{\mathbf{g}}_2$ and τ vary are illustrated in Fig. 2.

For $K = 3$, the active-regime multipliers also admit a closed form via Cramer's rule (App. B.3). For general K, the active set is determined by a standard primal active-set method (App. B.2), and $\tilde{\mathbf{d}}^*$ remains the orthogonal projection of $\tilde{\mathbf{g}}_1$ onto $F_\tau(\boldsymbol{\theta})$.

Primary preservation. The two-objective formula admits a clean analytical characterization of when PCD continues to make first-order descent on the primary, and when it does not. Assume $\|\tilde{\mathbf{g}}_1\| = \|\tilde{\mathbf{g}}_2\| = 1$ (the EMA steady state) and write $c = \tilde{\mathbf{g}}_1^\top \tilde{\mathbf{g}}_2$. In the active regime, Eq. (12) gives $\tilde{\mathbf{d}}^* = \tilde{\mathbf{g}}_1 + (\tau - c)\tilde{\mathbf{g}}_2$.

Proposition 4.7 (Primary preservation, $K = 2$)**.** *Under the active regime of Cor. 4.6 with $\|\tilde{\mathbf{g}}_1\| = \|\tilde{\mathbf{g}}_2\| = 1$ and $c = \tilde{\mathbf{g}}_1^\top \tilde{\mathbf{g}}_2$,*

$$\tilde{\mathbf{g}}_1^\top \tilde{\mathbf{d}}^* = 1 + \tau c - c^2. \tag{13}$$

Consequently the PCD update preserves first-order primary descent ($\tilde{\mathbf{g}}_1^\top \tilde{\mathbf{d}}^* > 0$) iff $1 + \tau c - c^2 > 0$. In the non-conflicting case $c \geq 0$ this holds for every $\tau \in [0,1]$. In the conflicting case $c < 0$ it holds iff

$$\tau < \frac{1-c^2}{-c}. \tag{14}$$

Proof. Substituting $\tilde{\mathbf{d}}^* = \tilde{\mathbf{g}}_1 + (\tau - c)\tilde{\mathbf{g}}_2$ into $\tilde{\mathbf{g}}_1^\top \tilde{\mathbf{d}}^*$, using $\|\tilde{\mathbf{g}}_1\|^2 = 1$ and $\tilde{\mathbf{g}}_1^\top \tilde{\mathbf{g}}_2 = c$, gives $\tilde{\mathbf{g}}_1^\top \tilde{\mathbf{d}}^* = 1 + (\tau - c)c = 1 + \tau c - c^2$. The conditions follow from $1 + \tau c - c^2 > 0$. □

Proposition 4.7 is the analytical fingerprint of PCD's behavior. Small τ leaves $\tilde{\mathbf{g}}_1^\top \tilde{\mathbf{d}}^*$ close to 1, keeping the update nearly aligned with the primary gradient. Larger τ deflects $\tilde{\mathbf{d}}^*$ further from $\tilde{\mathbf{g}}_1$, and in sufficiently severe conflict (c close to -1, with τ exceeding the threshold in Eq. (14)), $\tilde{\mathbf{g}}_1^\top \tilde{\mathbf{d}}^*$ can fall to zero or below. Geometrically, τ acts as a *secondary-pressure parameter*: it controls how much of the secondary direction is forced into the update, and accordingly how much of the primary direction is sacrificed in severe conflict. The threshold Eq. (14) predicts the breakdown observed empirically as τ is increased past a problem-dependent operating boundary.

4.6 Algorithm

Algorithm 1 summarizes a single PCD step. The QP of Eq. (4) admits a closed-form solution for $K = 2$ (Cor. 4.6), and is solved by a small primal active-set routine for $K \geq 3$ (App. B.2). Per-step cost is dominated by the K backpropagations needed to obtain $\mathbf{g}_1, \ldots, \mathbf{g}_K$. The QP itself is $K \times K$ and adds negligible overhead.

Algorithm 1: Priority-Constrained Descent (PCD), scalar τ, general K.

Input: Initial $\boldsymbol{\theta}_0$; scalar tolerance $\tau \in [0, 1]$; EMA decay β, stabilizer ε; step size η; optimizer $\mathcal{O}$.

Initialize $v_{i,0} \leftarrow 0$ for $i \in [K]$
for $t = 1, 2, \ldots, T$ **do**
  Compute $\mathbf{g}_1 \leftarrow \nabla L_1(\boldsymbol{\theta})$, $\mathbf{g}_j \leftarrow \mathrm{subgrad}(L_j, \boldsymbol{\theta})$ for $j \geq 2$
  for $i = 1, \ldots, K$ **do**
    Update $v_{i,t}, \hat{v}_{i,t}, s_{i,t}$ via Eq. (11); set $\tilde{\mathbf{g}}_i \leftarrow s_{i,t}\mathbf{g}_i$
  end
  if $\tilde{\mathbf{g}}_j^\top \tilde{\mathbf{g}}_1 \geq \tau \|\tilde{\mathbf{g}}_j\|^2$ *for all* $j \in [K]_{-1}$ **then**
    $\mathbf{d}^* \leftarrow \tilde{\mathbf{g}}_1$ `// constraint inactive`
  else
    Solve Eq. (4) via Cor. 4.6 (K=2) or active-set method (App. B.2) $\rightarrow \tilde{\mathbf{d}}^*$
  end
  $\mathbf{d}^* \leftarrow \tilde{\mathbf{d}}^* \cdot \|\mathbf{g}_1\| / \|\tilde{\mathbf{d}}^*\|$ `// magnitude rescaling, Sec. 4.7`
  $\boldsymbol{\theta} \leftarrow \mathcal{O}.\mathrm{step}(\boldsymbol{\theta}, \mathbf{d}^*, \eta)$
end
return $\boldsymbol{\theta}_T$

4.7 Scope of Guarantees

Our theoretical claims attach to the *normalized PCD direction* $\tilde{\mathbf{d}}^*$ and to its idealized first-order step $\boldsymbol{\theta}^+ = \boldsymbol{\theta} - \eta\tilde{\mathbf{d}}^*$. They are *local* and *directional*: Theorem 4.2 characterizes what direction the QP computes, and Proposition 4.4 characterizes the first-order progress this direction induces on each secondary.

The practical update in Algorithm 1 also rescales $\tilde{\mathbf{d}}^*$ to match the raw primary-gradient magnitude, $\mathbf{d}^* = \tilde{\mathbf{d}}^* \cdot \|\mathbf{g}_1\| / \|\tilde{\mathbf{d}}^*\|$, before passing it to the optimizer $\mathcal{O}$. Positive scalar rescaling preserves direction, hence preserves the signs of $\tilde{\mathbf{g}}_j^\top \mathbf{d}^*$ and the directional descent properties of Prop. 4.4. The magnitude rescaling exists for compatibility with standard optimizer step-size scheduling and is treated throughout as an implementation choice. We discuss its interaction with the convergence picture in App. C.3.

This separation between $\tilde{\mathbf{d}}^*$ (the analyzed object) and $\mathbf{d}^*$ followed by an optimizer step (the deployed object) is the same separation used in PCGrad, CAGrad, MGDA, FAMO, and other gradient-manipulation methods. As in those methods, our guarantees characterize the geometric direction-construction step. The empirical evaluation in Sec. 5 reports PCD as deployed.

4.8 Relation to Pareto Stationarity

PCD's fixed-point structure provides a sharp endpoint characterization that distinguishes it from symmetric MOO methods. The key observation is as follows: for symmetric methods, a Pareto-stationary point can occur through *gradient cancellation*, $\mathbf{0} \in \operatorname{conv}\{\mathbf{g}_1, \ldots, \mathbf{g}_K\}$, even when no individual $\mathbf{g}_i$ vanishes. PCD's QP rules this out.

Theorem 4.8 (Fixed-point characterization)**.** *Let $\tau > 0$ and suppose $F_\tau(\boldsymbol{\theta}) \neq \varnothing$. Then*

$$\tilde{\mathbf{d}}^*(\boldsymbol{\theta}) = \mathbf{0} \iff \tilde{\mathbf{g}}_i(\boldsymbol{\theta}) = \mathbf{0} \text{ for all } i \in [K]. \tag{15}$$

Proof sketch. ($\Leftarrow$) If every $\tilde{\mathbf{g}}_i = \mathbf{0}$ then $\tilde{\mathbf{d}} = \mathbf{0}$ is feasible and attains the objective minimum $\|\mathbf{0} - \mathbf{0}\|^2 = 0$. ($\Rightarrow$) Primal feasibility with $\tilde{\mathbf{d}}^* = \mathbf{0}$ gives $0 \geq \tau \|\tilde{\mathbf{g}}_j\|^2$, and since $\tau > 0$ this forces $\tilde{\mathbf{g}}_j = \mathbf{0}$ for every $j \geq 2$. KKT stationarity then forces $\tilde{\mathbf{g}}_1 = \mathbf{0}$ as well. The full proof, including the precise dependence on Theorem 4.2, is in App. C.1. □

The right-hand side of Eq. (15) is composite multi-stationarity (CMS; Def. 3.3), which by definition is strictly stronger than Pareto stationarity, ruling out the cancellation equilibria at which symmetric methods can halt. Theorem 4.8 is best read as an *endpoint characterization* of the deterministic, normalized QP: it identifies the only points at which the idealized PCD direction can vanish, but is not a stochastic convergence claim. Whether the iterates of Algorithm 1 approach such an endpoint depends on optimizer, step size, batch noise, and feasibility along the path, none of which are subsumed by the QP characterization. A complementary endpoint result at $\tau = 0$ recovers Pareto stationarity rather than CMS (App. C.2).

Operating curves in τ. As τ varies in $[0, 1]$ the feasible set $F_\tau(\boldsymbol{\theta})$ varies monotonically. On asymmetric problems where joint stationarity is structurally possible (for instance, regularizer-style secondaries that vanish at minima of the primary in the over-parameterized regime), this monotone family yields a continuous operating curve interpolating between the Pareto-stationary endpoint ($\tau = 0$) and CMS endpoints ($\tau > 0$). On genuine trade-off problems with no common stationary point, no CMS point exists and PCD with $\tau > 0$ has no fixed point. Iterates need not converge in this regime, and the framework is best understood as targeting the regularizer-style hierarchy of Eq. (1).

4.9 Comparison with Existing Methods

Fig. 3 contrasts PCD against established gradient-based MOO methods in the standard conflicting ($\mathbf{g}_1^\top \mathbf{g}_2 < 0$) and non-conflicting ($\mathbf{g}_1^\top \mathbf{g}_2 \geq 0$) regimes. The contrast is structural: existing methods combine objective gradients symmetrically, producing updates in which $\mathbf{g}_1$ and $\mathbf{g}_2$ are interchangeable. Weighted sum (Marler & Arora, 2010) forms $w_1\mathbf{g}_1 + w_2\mathbf{g}_2$ at fixed weights, MGDA (Désidéri, 2009) returns the minimum-norm convex combination $\alpha^*\mathbf{g}_1 + (1-\alpha^*)\mathbf{g}_2$, PCGrad (Yu et al., 2020) projects out conflicting components pairwise and averages, reducing to $\mathbf{g}_1 + \mathbf{g}_2$ when the gradients do not conflict, and CAGrad (Liu et al., 2021a) anchors at the average $\mathbf{g}_0 = \frac{1}{2}(\mathbf{g}_1 + \mathbf{g}_2)$ and improves the worst case within a trust region. Swapping the labels of $\mathbf{g}_1$ and $\mathbf{g}_2$ leaves every one of these updates unchanged.

PCD breaks this symmetry by construction. The KKT identity of Eq. (7) pins to the primary gradient and admits the secondaries only through non-negative multipliers that vanish on inactive constraints. The QP objective $\|\tilde{\mathbf{d}} - \tilde{\mathbf{g}}_1\|^2$ and the constraints involving only $\tilde{\mathbf{g}}_j$ for $j \geq 2$ encode the asymmetry directly, not through preference weights. Two consequences follow. First, PCD exhibits *three* update regimes rather than two: constraint inactive ($\tilde{\mathbf{d}}^* = \tilde{\mathbf{g}}_1$), constraint active without conflict, and constraint active with conflict. This is determined by where $\tilde{\mathbf{g}}_2^\top \tilde{\mathbf{g}}_1$ sits relative to $\tau \|\tilde{\mathbf{g}}_2\|^2$ (Fig. 2, Cor. 4.6). Second, the MGDA failure at $\mathbf{g}_1 = \mathbf{0}$, where $\mathbf{0} \in \operatorname{conv}\{\mathbf{0}, \mathbf{g}_2, \ldots, \mathbf{g}_K\}$ collapses the combination to $\mathbf{0}$ regardless of the remaining $K - 1$ secondaries (App. B.7, Prop. B.10), does not occur for PCD with $\tau > 0$: the QP continues to drive every non-stationary secondary toward stationarity (Lemma B.7).

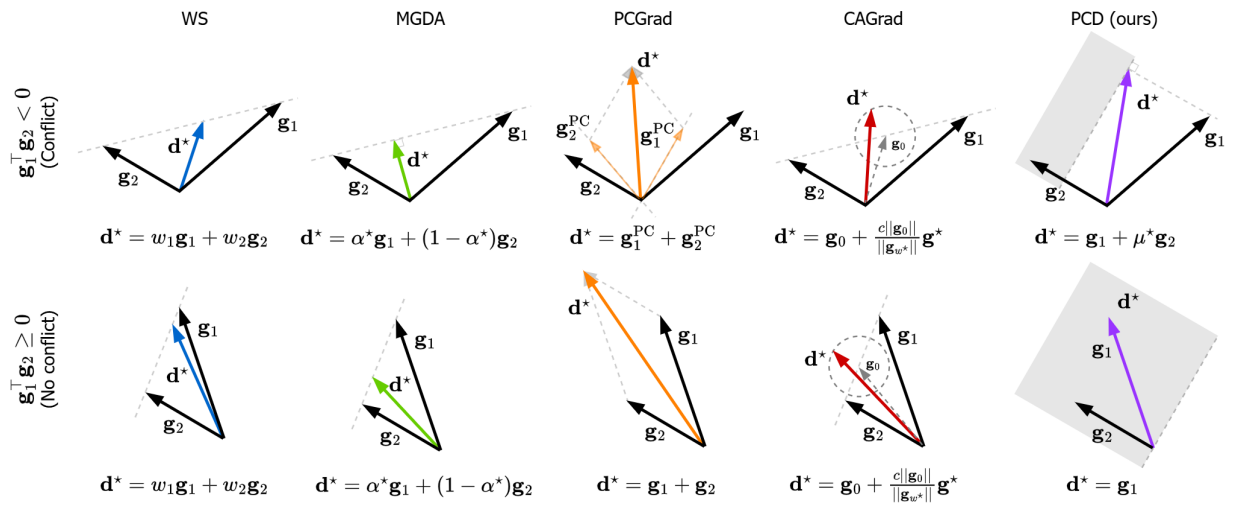


Figure 3: Comparison of PCD against established gradient-based multi-objective methods. **Top Row**: Conflicting case, $\mathbf{g}_1^\top \mathbf{g}_2 < 0$. **Bottom Row**: Non-conflicting case, $\mathbf{g}_1^\top \mathbf{g}_2 \geq 0$. For PCD in the non-conflicting case, $\tau = 0$ is used to align conflict definitions.

5 Experimental Results

In this section, we conduct experiments over two application domains and one set of synthetic experiments to demonstrate the superiority of PCD in asymmetric applications. In particular, we conduct experiments involving structured pruning and compression of neural networks (Sec. 5.1) with $K = 2$ objectives, unstructured sparse and low-rank networks (Sec. 5.2) with $K = 3$ objectives, and a set of synthetic experiments to validate PCD's advantages (Sec. 5.3).

5.1 Structured Pruning

Structured neural-network compression is a natural instance of asymmetric multi-objective optimization. We set the primary objective $\mathcal{L}_1$ to be the standard cross-entropy loss and the secondary objective $\mathcal{L}_2$ to be the Group Lasso penalty:

$$\mathcal{L}_2(\theta) \;=\; \sum_g \|\theta_g\|_2, \tag{16}$$

where the sum runs over groups g defined at the output-channel level of each layer. The formulation is asymmetric by construction: task accuracy is the deliverable, while structural sparsity is a secondary objective whose pressure is controlled by τ. After training, groups whose norm $\|\theta_g\|_2$ falls below a threshold are removed and the resulting network is evaluated directly without fine-tuning. To produce results at a specified parameter-reduction target $r \in \{80\%, 85\%, 90\%, 95\%, 99\%\}$, the threshold is selected post hoc, per (method, architecture, dataset, target), as the smallest value for which the fraction of removed parameters reaches r.

Experimental Protocol. We evaluate on four architectures spanning a range of structural properties: DenseNet-121 (Huang et al., 2017), ResNet-34 (He et al., 2016), an Inception-style model (Szegedy et al., 2015), and MobileNetV2 (Howard et al., 2017; Sandler et al., 2018), each trained on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009). All models are trained from scratch with Adam (Kingma & Ba, 2014), learning rate 10^{-3}, batch size 128, cosine annealing, and 300 epochs. Each configuration is run for 5 independent seeds, with the reported numbers being means. PCD is compared against six multi-objective baselines: PCGrad, MGDA, FAMO, CAGrad, weighted sum (WS), and AuxiNash. For baselines with a tunable hyperparameter (the weight w for WS and the conflict coefficient c for CAGrad, and p for AuxiNash), we sweep over $\{0.1, 0.2, \ldots, 0.9\}$ and report the best operating point per configuration. MGDA, PCGrad, and

FAMO have no such hyperparameter and are run at their default settings. For PCD we use a denser τ sweep $\in \{0.01, 0.02, \ldots, 0.09, 0.1, 0.2, \ldots, 0.9, 1\}$.

Fig. 4 reports test accuracy as a function of parameter reduction across all eight configurations. PCD (purple, ⋆) sits above every baseline across all eight panels and at every compression target. The gap is widest on the harder benchmarks: on ResNet-34/CIFAR-100, PCD retains 73.0% accuracy at 80–85% parameter reduction, while MGDA (1.2%), FAMO (2.7%), PCGrad (7.4%), and CAGrad (7.3%) collapse to near-random performance across all compression targets. AuxiNash is the strongest baseline at 44.4%, 28.6% worse than PCD. The DenseNet-121/CIFAR-10 results illustrate PCD's operating range: 92.9% accuracy at 95% parameter reduction and 89.2% at 99% reduction. These are accuracy levels no baseline approaches at any compression target. Seed-by-seed values for PCD and for every baseline are tightly clustered around the reported means. Tables 1 and 2 give the full numerical comparison for these two configurations.

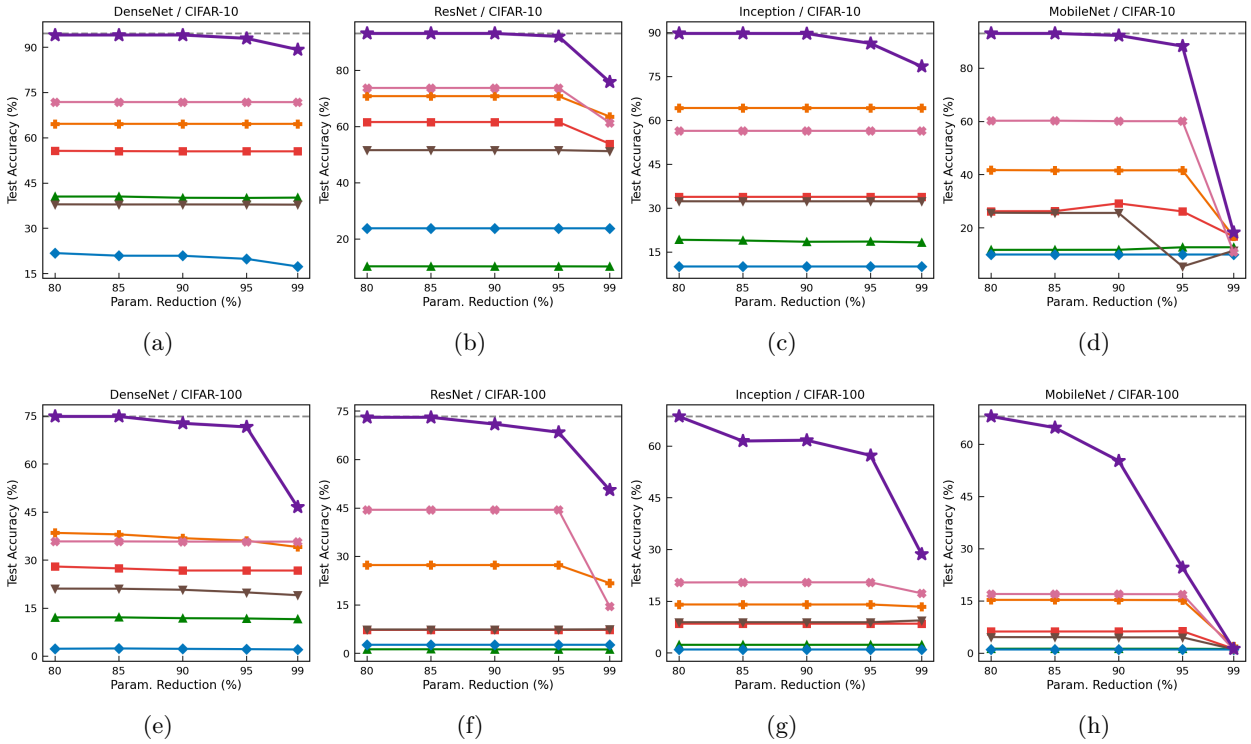


Figure 4: Sparsity results as a function of pruning across DenseNet-121, ResNet-34, Inception, and MobileNet-v2, trained on CIFAR-10 (top row) and CIFAR-100 (bottom row). Across all architecture / dataset configurations and at all pruning levels, PCD (purple) outperforms all multi-objective baselines, including weighted sum (orange), CAGrad (red), PCGrad (brown), MGDA (green), FAMO (blue), and AuxiNash (pink).

Efficiency and Pareto frontiers. Tables 1 and 2 also report the hardware impact of PCD's compression on the two representative architectures, measured on CPU inference. On DenseNet-121/CIFAR-10, $\tau = 0.01$ yields a 5.7× FLOPs reduction, 3.7× latency reduction, and 11.5× file-size reduction while matching unpruned accuracy (94.0% vs. 94.5%). Increasing to $\tau = 0.05$ reaches 43.2× FLOPs reduction and 97.2× file-size reduction at 89.2% accuracy. ResNet-34/CIFAR-100 shows the same pattern: small accuracy costs traded for order-of-magnitude resource savings.

Fig. 5 shows accuracy–efficiency Pareto frontiers against all baselines for ResNet-34/CIFAR-100. PCD's τ sweep traces a curve across the (accuracy, resource) plane on each of the three resource axes (latency, file size, FLOPs), while each baseline contributes a single operating point per (compression target) row of Tab. 2. We plot the best operating point reached over its hyperparameter sweep for the swept baselines, and the default-setting operating point for the others. On every axis and at every compression level shown, baseline

Table 1: Test accuracy (%) and deployment efficiency for all methods on DenseNet-121 / CIFAR-10 (unpruned: 94.5%, 1.82 G FLOPs, 62.4 ms latency, 27.2 MB). FLOPs, latency, and model-size reductions are relative to the unpruned model and depend only on the number of remaining parameters at each target; all methods achieve the same efficiency gains (↑ same). Accuracy values are means over 5 seeds.

Target	Metric	PCD (Ours)	WS	CAGrad	MGDA	PCGrad	FAMO	AuxiNash
Unpruned	Acc.	**94.5**	94.5	94.5	94.5	94.5	94.5	94.5
	FLOPs	1.82 G						
	Lat.	62.4 ms				↑ same		
	Size	27.2 MB						
80%	Acc.	**94.0**	64.6	55.6	40.5	37.9	21.8	71.8
	FLOPs	0.48 G (3.8×)						
	Lat.	24.0 ms (2.6×)				↑ same		
	Size	4.9 MB (5.6×)						
85%	Acc.	**94.0**	64.6	55.5	40.5	37.9	20.9	71.8
	FLOPs	0.38 G (4.8×)						
	Lat.	20.8 ms (3.0×)				↑ same		
	Size	3.7 MB (7.4×)						
90%	Acc.	**94.0**	64.6	55.5	40.1	37.9	20.9	71.8
	FLOPs	0.32 G (5.7×)						
	Lat.	16.9 ms (3.7×)				↑ same		
	Size	2.4 MB (11.5×)						
95%	Acc.	**92.9**	64.6	55.5	40.0	37.9	19.8	71.8
	FLOPs	0.12 G (15.7×)						
	Lat.	9.2 ms (6.8×)				↑ same		
	Size	0.88 MB (30.8×)						
99%	Acc.	**89.2**	64.6	55.5	40.1	37.8	17.3	71.8
	FLOPs	0.042 G (43.2×)						
	Lat.	5.0 ms (12.5×)				↑ same		
	Size	0.28 MB (97.2×)						

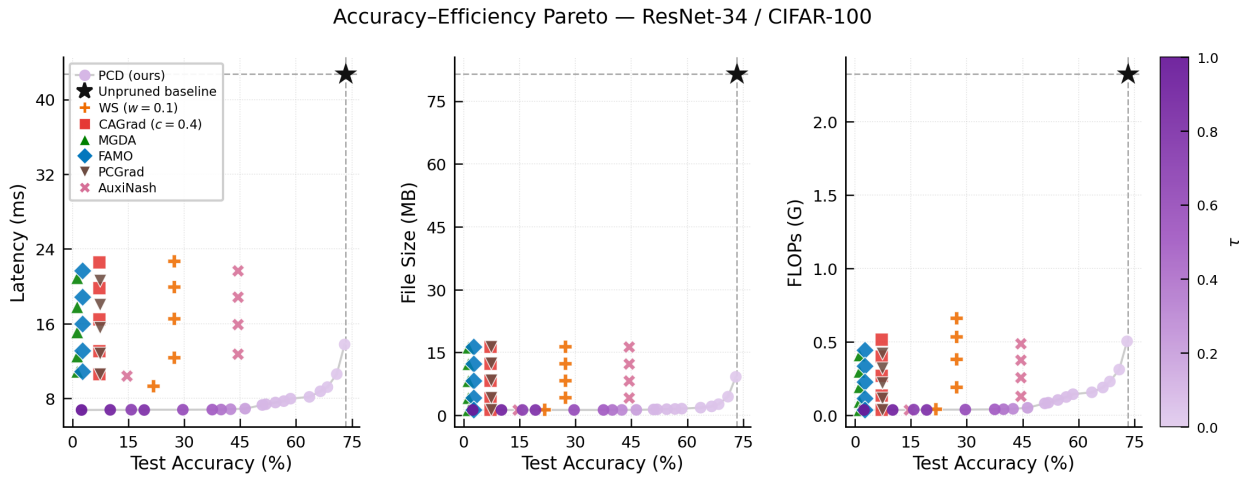


Figure 5: Accuracy–efficiency Pareto frontiers for PCD on ResNet-34/CIFAR-100 across inference latency (ms), model file size (MB), and FLOPs (G). Unpruned baseline (⋆) is shown for reference.

operating points lie above the PCD curve. The comparison is asymmetric in structure (a swept curve against a small set of points), but it is the natural comparison given that the swept baselines' best operating points already saturate their accuracy ceiling: increasing w for WS or c for CAGrad beyond the chosen value does

Table 2: Test accuracy (%) and deployment efficiency for all methods on ResNet-34 / CIFAR-100 (unpruned: 73.3%, 2.33 G FLOPs, 42.7 ms latency, 81.5 MB). FLOPs, latency, and model-size reductions are relative to the unpruned model and are shared across all methods at a given target (↑ same). The ≈99% row reflects the maximum achievable reduction (≈98.5%) for ResNet-34 under this pruning scheme. Accuracy values are means over 5 seeds.

Target	Metric	PCD (Ours)	WS	CAGrad	MGDA	PCGrad	FAMO	AuxiNash
Unpruned	Acc.	**73.3**	73.3	73.3	73.3	73.3	73.3	73.3
	FLOPs	2.33 G						
	Lat.	42.7 ms				↑ same		
	Size	81.5 MB						
80%	Acc.	**73.0**	27.3	7.3	1.2	7.4	2.7	44.4
	FLOPs	0.58 G (4.0×)						
	Lat.	18.6 ms (2.3×)				↑ same		
	Size	12.7 MB (6.4×)						
85%	Acc.	**73.0**	27.3	7.3	1.2	7.4	2.7	44.4
	FLOPs	0.44 G (5.3×)						
	Lat.	15.8 ms (2.7×)				↑ same		
	Size	9.5 MB (8.6×)						
90%	Acc.	**70.9**	27.3	7.3	1.2	7.4	2.7	44.4
	FLOPs	0.31 G (7.5×)						
	Lat.	10.7 ms (4.0×)				↑ same		
	Size	4.4 MB (18.4×)						
95%	Acc.	**68.4**	27.3	7.3	1.2	7.4	2.7	44.4
	FLOPs	0.23 G (10.1×)						
	Lat.	9.3 ms (4.6×)				↑ same		
	Size	2.6 MB (30.8×)						
≈99%	Acc.	**50.6**	21.7	7.3	1.2	7.4	2.7	14.5
	FLOPs	0.037 G (62.9×)						
	Lat.	6.8 ms (6.3×)				↑ same		
	Size	1.3 MB (64.7×)						

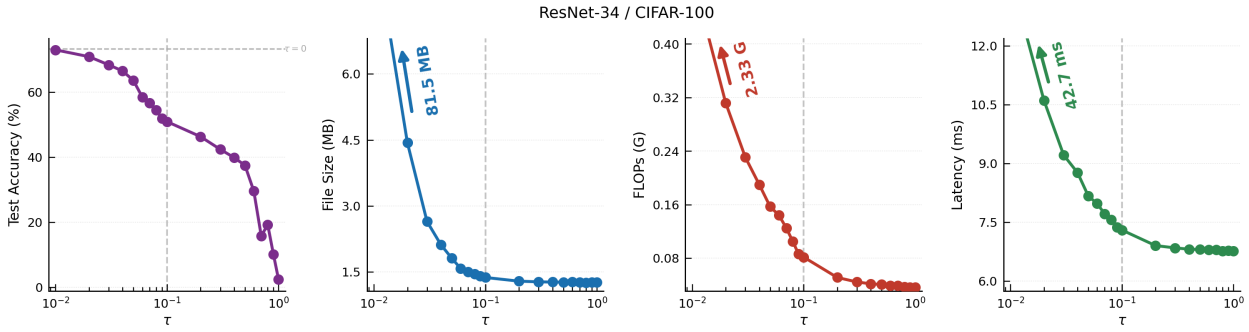


Figure 6: Accuracy, file size (MB), FLOPs (G) and latency (ms) as a function of the τ parameter for ResNet-34 trained on the CIFAR-100 dataset. Arrows pointing up show values for the respective efficiency metrics at $\tau = 0$ (unpruned baseline).

not yield further accuracy improvement. The takeaway is operational rather than topological: not only can τ be tuned to land at a chosen point along PCD's frontier, but PCD's operating points are simultaneously better than the baselines' best operating points across all three resource axes.

A practical consequence is that τ acts as a single, interpretable compression dial. Rather than searching over weight ratios (WS), conflict coefficients (CAGrad), or separate regularization schedules (MGDA / PCGrad /

FAMO), a practitioner can sweep $\tau \in (0, 1]$ to trace a continuous Pareto curve at training time. Small τ (light purple in Fig. 5) yields light compression with near-baseline accuracy, whereas large τ (dark blue) yields aggressive compression. The curve is smooth and monotone, with no discontinuous jumps or instabilities as τ increases. This is a property that simplifies deployment-time selection based on a target resource budget. As Fig. 6 shows for ResNet-34/CIFAR-100, the largest resource gains come from $\tau \in (0, 0.1]$, with curves plateauing beyond this. Accuracy drops are also gentlest within this range. For $\tau > 0.1$ on this configuration, accuracy degrades sharply. The same qualitative pattern holds across the other seven (architecture, dataset) combinations. Per-configuration τ sweeps are reported in App. E.

5.2 Unstructured Sparsity and Low-Rankness

The structured-pruning study of Sec. 5.1 pairs the task loss with a single secondary. We now move to a genuine three-objective setting ($K = 3$), in which the primary cross-entropy loss $\mathcal{L}_1$ is constrained *simultaneously* by two structurally distinct regularizers. The first is an ℓ_1 penalty, which drives unstructured weight sparsity,

$$\mathcal{L}_2(\theta) \;=\; \|\theta\|_1 \;=\; \sum_i |\theta_i|, \tag{17}$$

and the second is a nuclear-norm penalty, which drives each layer toward low rank,

$$\mathcal{L}_3(\theta) \;=\; \sum_\ell \|W_\ell\|_* \;=\; \sum_\ell \sum_k \sigma_k(W_\ell), \tag{18}$$

where W_ℓ is the weight matrix of layer ℓ and $\sigma_k(\cdot)$ its k-th singular value. As in the structured case the hierarchy is asymmetric by construction: task accuracy is the deliverable, while sparsity and low-rankness are secondary structural objectives whose pressure is modulated by τ.

We note that the two secondaries are *not* aligned. The ℓ_1 penalty acts entry-wise, zeroing individual weights wherever they are small and without regard to their position, whereas the nuclear norm acts on the spectrum, collapsing entire singular directions and thereby favoring weight matrices whose rows and columns are linearly dependent. Although in heavily over-compressed regimes the two objectives can appear correlated, their useful operating ranges impose competing pressures on shared parameters.

Experimental Protocol. We evaluate on ResNet-34 (He et al., 2016) and an Inception-style model (Szegedy et al., 2015), each trained on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), following the protocol of Sec. 5.1. All models are trained from scratch with Adam, learning rate 10^{-3}, batch size 128, cosine annealing, and 300 epochs. Each configuration is run for 3 independent seeds, with the reported numbers being means. PCD is compared against the same six multi-objective baselines: PCGrad, MGDA, FAMO, CAGrad, weighted sum (WS), and AuxiNash. For the baselines with a tunable hyperparameter (the weight w for WS, the conflict coefficient c for CAGrad, and p for AuxiNash) we sweep $\{0.1, 0.2, \ldots, 0.9\}$ and report the best operating point per configuration based on accuracy. MGDA, PCGrad, and FAMO are run at their default settings. After training, weight sparsity is read off by thresholding small-magnitude weights and low-rankness from the singular spectrum of each layer, both post hoc and without fine-tuning.

We summarize each method in two complementary ways. First, on the raw and directly interpretable axes of test accuracy, parameter sparsity, and effective rank, we trace PCD's operating point as a function of τ and overlay each baseline's best-accuracy operating point as a horizontal reference (Fig. 7). Second, to compare the full achievable sets in a single scalar, we report the dominated hypervolume. As the three objective losses span different numerical ranges, we first place them on a common footing by normalizing each one per configuration via $x \mapsto (\log_{10} x - \log_{10} x_{\min})/(\log_{10} x_{\max} - \log_{10} x_{\min})$, mapping every axis to $[0, 1]$ with lower values better and the joint ideal at the origin. Given a method's set of operating points and a fixed reference point, the hypervolume is the volume of this normalized space that the set dominates. It rewards a method both for approaching the joint ideal on all three axes at once and for covering more of the trade-off surface, so a larger hypervolume indicates that a method's achievable set is jointly superior rather than merely better on one axis in isolation. We summarize low-rankness by the *effective rank* (Roy & Vetterli, 2007): the exponential of the entropy of each layer's normalized singular values, averaged over layers (which is threshold-free and, unlike a hard rank count, is not inflated by simply driving a layer's weights toward zero).

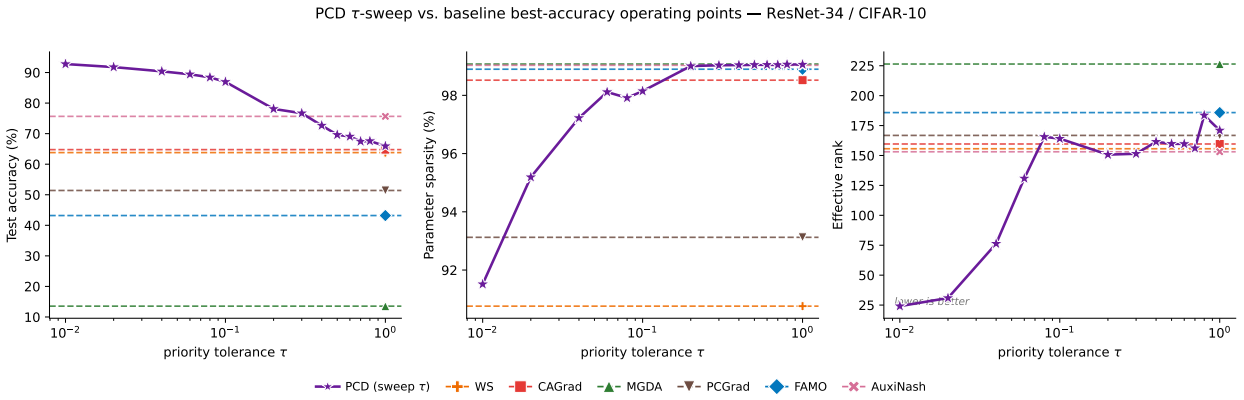


Figure 7: PCD's τ-sweep on the three deployment-relevant axes for ResNet-34/CIFAR-10: test accuracy, parameter sparsity, and effective rank (lower is better). PCD ($\star$) is swept over $\tau \in (0, 1]$, each baseline is a horizontal line at its highest-accuracy operating point (the points in Table 3), so a vertical slice at any τ compares PCD against every baseline on all three axes at once. Over a wide range of τ, PCD's accuracy exceeds every baseline and its effective rank lies well below every baseline. Increasing τ trades these for parameter sparsity, which reaches the baseline cluster near $\tau \approx 0.2$.

Table 3: Operating points on the $K = 3$ (cross-entropy / ℓ_1 / nuclear norm) task, across all four architecture/-dataset configurations. For each baseline we report the operating point of highest mean test accuracy. For PCD we report a single fixed tolerance ($\tau = 0.02$), a representative low-tolerance setting that trades a small amount of accuracy for stronger compression, rather than its accuracy-optimal point. At the selected point we report accuracy, parameter sparsity, and *effective rank*. Accuracy is the main "deliverable", whilst sparsity and effective rank characterize the compression attained there. Across every configuration PCD attains the highest accuracy by a wide margin. The baselines reach higher nominal sparsity only at a large accuracy cost. PCD recovers near-unpruned accuracy (unpruned: 94.1/71.9/90.7/67.8% for the four configurations, in row order). Means over 3 seeds.

Configuration	Metric	PCD (Ours)	WS	CAGrad	MGDA	PCGrad	FAMO	AuxiNash
ResNet-34 CIFAR-10	Acc. (%)	**91.8**	63.8	64.8	13.5	51.4	43.2	75.7
	Sparsity (%)	95.2	90.8	98.5	99.1	93.1	98.9	99.0
	Eff. rank ↓	30.9	155.6	159.6	226.4	166.8	185.8	153.1
ResNet-34 CIFAR-100	Acc. (%)	**71.0**	34.2	36.1	1.1	30.0	20.3	48.7
	Sparsity (%)	92.6	90.6	98.9	98.9	92.2	98.4	98.8
	Eff. rank ↓	24.5	137.9	186.3	227.6	160.0	190.5	184.8
Inception CIFAR-10	Acc. (%)	**89.2**	52.1	38.2	23.1	33.8	10.0	60.3
	Sparsity (%)	86.1	90.8	99.1	99.1	91.6	99.2	99.0
	Eff. rank ↓	31.3	39.7	51.2	76.4	55.8	100.3	18.0
Inception CIFAR-100	Acc. (%)	**67.8**	29.8	14.1	4.3	12.4	1.0	31.9
	Sparsity (%)	76.1	86.6	94.3	94.5	86.8	94.5	93.1
	Eff. rank ↓	27.7	28.1	59.0	70.3	52.4	100.3	12.1

Operating curves. Fig. 7 reports the result for ResNet-34/CIFAR-10. Each baseline collapses to a single operating point regardless of its hyperparameter, so we draw it as a horizontal line at its best-accuracy configuration (the values in Table 3). PCD, in contrast, is swept over τ and traces a continuous curve on each axis. A vertical slice at any τ therefore compares PCD against every baseline on accuracy, sparsity, and effective rank at once.

The comparison is decisive on two axes and competitive on the third. Over the entire low-to-moderate range of τ, PCD's test accuracy exceeds that of every baseline, remaining near the dense-network ceiling at small τ and only dropping below the strongest baseline once τ is pushed into the aggressively compressed

regime. On the rank axis, PCD's effective rank lies far below every baseline at small τ: it discovers genuinely low-dimensional structure, rather than the degenerate near-zero spectra that inflate the baselines' effective rank. Parameter sparsity is the lone axis on which the baselines hold an initial edge. PCD starts less sparse and climbs with τ, reaching the baseline cluster near $\tau \approx 0.2$. The baselines attain that sparsity only by collapsing the network, paying for it with the destroyed accuracy and inflated effective rank visible in the adjacent panels, whereas PCD reaches comparable sparsity while preserving both accuracy and low-rank structure.

Table 3 quantifies these results. Across all configurations PCD retains the highest accuracy relative to the original dense networks, whereas the other methods over-compress, trading accuracy for gains on the ℓ_1 and nuclear norms. PCD maintains competitive effective rank, with the only exceptions on the Inception-style architecture: AuxiNash reaches lower effective rank on both Inception configurations, despite trailing PCD by 30–40% in accuracy. The single competitive baseline overall is AuxiNash, which nonetheless attains a substantially smaller dominated hypervolume than PCD (Fig. 8). It is very easy to see here, however, that having independent τ values for each objective can, in principle, improve the performance here even further, which we defer to future work.

Hypervolume. Fig. 8 aggregates these comparisons into the dominated hypervolume of each method, per configuration. Across configurations PCD dominates a substantially larger region of the joint objective space than any baseline, and the margin is stable across seeds. The hypervolume confirms that the per-axis separation in Fig. 7 is not an artifact of any single axis or operating point, but reflects a genuinely dominant operating set: one closer to the joint optimum on cross-entropy, sparsity, and low-rankness at the same time.

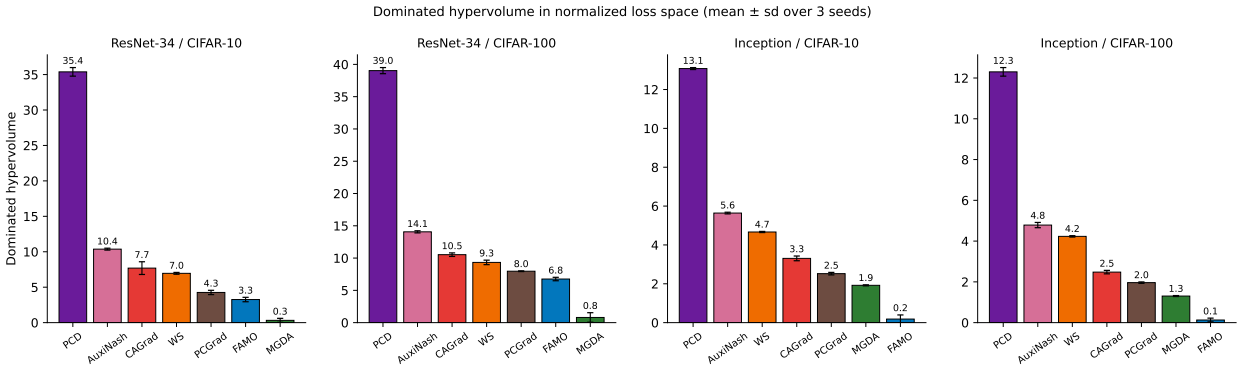


Figure 8: Dominated hypervolume in the normalized $K = 3$ objective space, one panel per architecture/dataset configuration. For each configuration the three losses are mapped to goodness coordinates $g = 1 - \tilde{L} \in [0,1]^3$ (per-configuration $\log_{10}$ min–max normalization of each loss across all methods, ideal at $(1,1,1)$), and we report the volume of $[0,1]^3$ dominated by each method's full set of operating points relative to the origin reference $(0,0,0)$. Values are scaled by 100 for readability and shown as mean ± sd over 3 seeds. PCD attains the largest hypervolume in every configuration, consistently performing 2.3–3.4× the strongest baseline (AuxiNash), reflecting that sweeping τ traces a controllable front that reaches the accuracy ceiling, whereas the symmetric baselines cluster in the over-compressed corner.

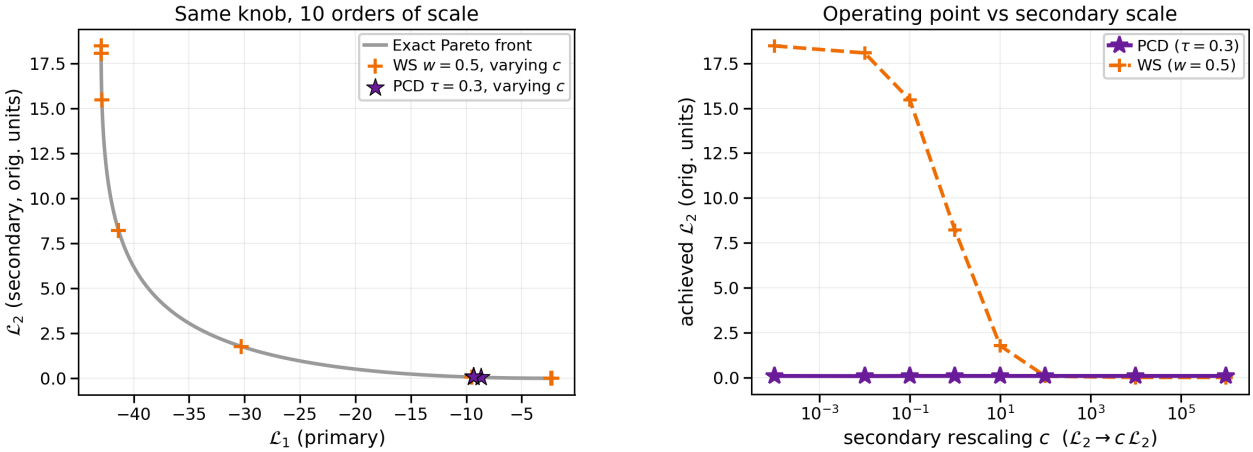
5.3 Synthetic Experiments

We complement the deep-network results with controlled problems that expose quantities the network experiments cannot measure directly: the scale-dependence of a method's operating point, the feasibility and active-set structure of the QP as objectives accumulate, and the exact stationarity reached at convergence. Throughout, the primary $\mathcal{L}_1(\theta) = \frac{1}{2}\theta^\top H\theta - c^\top\theta$ is a convex quadratic in $\mathbb{R}^{50}$ with logarithmically spaced eigenvalues and condition number $\kappa(H) = 10^2$, $c \sim \mathcal{N}(0, I)$, mimicking the long-tailed spectra of network Hessians (Sagun et al., 2017; Ghorbani et al., 2019). PCD uses the EMA normalization of Definition 4.5 with $\beta = 0.999$, $\varepsilon = 10^{-8}$. We compare against weighted sum (WS), MGDA, PCGrad, CAGrad, and AuxiNash. The first two experiments verify properties of the PCD construction itself: the scale invariance of its

tolerance and the feasibility of its QP as objectives accumulate on convex problems where they are cleanly measurable. The third isolates the property that genuinely requires non-convexity: the strictly-stronger fixed-point guarantee of Theorem 4.8, and how the resulting separation from symmetric methods scales with the number of objectives.

5.3.1 Scale invariance of τ

The meaning of τ rests on PCD's gradient normalization. As a result, we test whether a fixed τ yields a fixed operating point as a secondary is rescaled. Pairing the primary with a single convex secondary and holding $\tau = 0.3$ (PCD) and $w = 0.5$ (WS), we rescale $\mathcal{L}_2 \mapsto c\,\mathcal{L}_2$ over $c \in [10^{-4}, 10^{6}]$ and record the achieved operating point, in the *original* units of $\mathcal{L}_2$, on the exact Pareto front (Fig. 9). PCD's operating point is invariant across all ten orders of magnitude (total spread 0.0075 in $\mathcal{L}_2$), confirming the asymptotic scale invariance of Theorem B.8. WS instead traverses the entire front. Its achieved secondary ranges from 18.5 at $c = 10^{-4}$ to effectively 0 at $c = 10^{6}$, so the same nominal weight encodes a completely different priority depending on the arbitrary scale of the penalty, exactly the failure of Corollary B.9. The EMA normalization is what makes τ transferrable across objectives whose raw magnitudes differ by orders of magnitude, the regime typical of a task loss paired with secondary constraints.



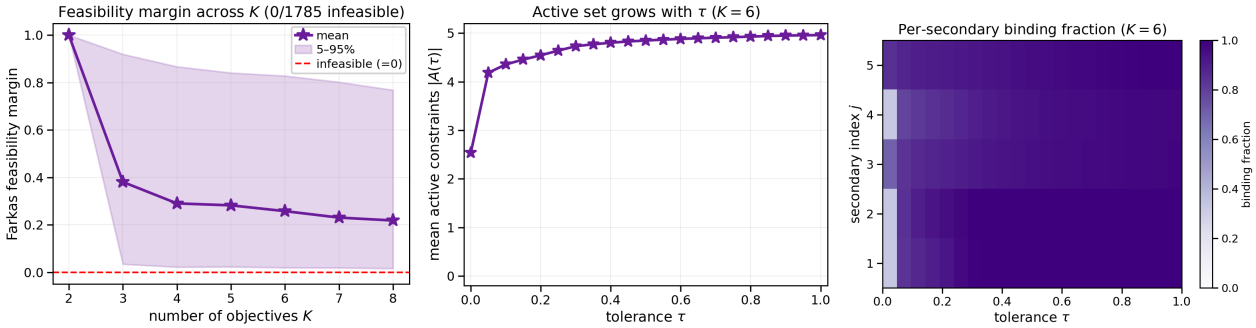
(a) Operating points on the front

(b) Achieved secondary vs. scale

Figure 9: **Scale invariance of τ.** Rescaling the secondary by $c \in [10^{-4}, 10^{6}]$ at fixed knobs. (a) PCD ($\tau = 0.3$) stays at one point on the front (spread 0.0075) while WS ($w = 0.5$) is dragged across it. (b) The achieved $\mathcal{L}_2$ in original units is flat for PCD and sweeps the full range for WS. PCD's knob has a scale-free meaning (Theorem B.8), while WS's weight does not (Corollary B.9).

5.3.2 Feasibility and active set with more objectives

The one phenomenon specific to $K > 2$ is that the secondary-progress polyhedron $\mathcal{F}_\tau(\theta)$ can in principle be empty (Sec. 4.1), which we probe directly. Three quantities certify that the K-objective QP is well-posed: whether a feasible update exists at all (Fig. 10a), how many secondary constraints the solution enforces (Fig. 10b), and whether enforcement is spread across all secondaries or concentrated on a few (Fig. 10c). A constraint is *binding* (active) when the solution meets it with equality, $\tilde{g}_j^\top \tilde{d}^* = \tau \|\tilde{g}_j\|^2$ (equivalently $\mu_j > 0$), so that secondary actively shapes the update, otherwise it is slack ($\mu_j = 0$) and drops out of $\tilde{d}^* = \tilde{g}_1 + \sum_{j \geq 2} \mu_j \tilde{g}_j$. For $K \in \{2, \ldots, 8\}$ we draw random normalized secondary-gradient configurations and evaluate the Farkas feasibility margin of App. B.1 (positive iff $\mathcal{F}_\tau \neq \varnothing$). Fig. 10(a) shows the margin shrinking with K. More simultaneous directional constraints leave less slack, but remain strictly positive throughout, with 0 infeasible events in 1785 checks, empirically consistent with infeasibility being a Lebesgue measure-zero event in the high-dimensional regime (Corollary B.3(iii)). Figs. 10(b) and (c) characterize the solution at $K = 6$: the active set grows monotonically with τ, from 2.54 active constraints at $\tau = 0$ to 4.96 of 5 (since we have one primary and 5 secondaries for $K = 6$) as $\tau \to 1$. The receding half-spaces tighten the polyhedron



(a) Feasibility margin vs. K (b) Active set vs. τ (K=6) (c) Per-secondary binding (K=6)

Figure 10: **Feasibility and active-set behavior at $K > 2$.** (a) The Farkas feasibility margin decreases with K but stays strictly positive; 0 of 1785 configurations were infeasible (Corollary B.3(iii)). (b) The active set grows monotonically with τ (2.54→4.96 of 5). (c) Each secondary binds with high frequency for $\tau \gtrsim 0.05$.

(Sec. 4.2), and every secondary binds with high frequency once $\tau \gtrsim 0.05$. Fig. 10(c) shows this binding distributed roughly uniformly across the five secondaries rather than concentrated on a subset: each is a genuinely active constraint that the projection meets at equality, not one incidentally over-satisfied while a few dominate, so the QP exercises all $K-1$ constraints rather than collapsing to a low-effective-K problem. The K-objective QP is thus well-posed in practice and behaves as the geometry of Sec. 4 predicts.

5.3.3 Conflict equilibria and CMS escape as the number of objectives grows

The final experiment isolates the property that separates PCD's fixed-point set from that of symmetric methods (Theorem 4.8) and measures how that separation scales with K. We construct a *non-convex* primary with a double-well along each of $K-1$ random orthonormal directions u_j (plus a convex bowl in the orthogonal complement), together with $K-1$ secondaries, the j-th pulling $\langle u_j, \theta\rangle \to +1$. The secondaries are jointly satisfiable, so a composite multi-stationary point (at which every objective is individually stationary) exists at $\theta^\star = \sum_j u_j$. This is what makes the problem an *escape* problem rather than a trade-off. All methods are initialized in a wrong basin ($\langle u_j, \theta\rangle = -0.6$ for every j), where the canceled primary gradient is largest. Concretely, writing $s_j = \langle u_j, \theta\rangle$, the primary is

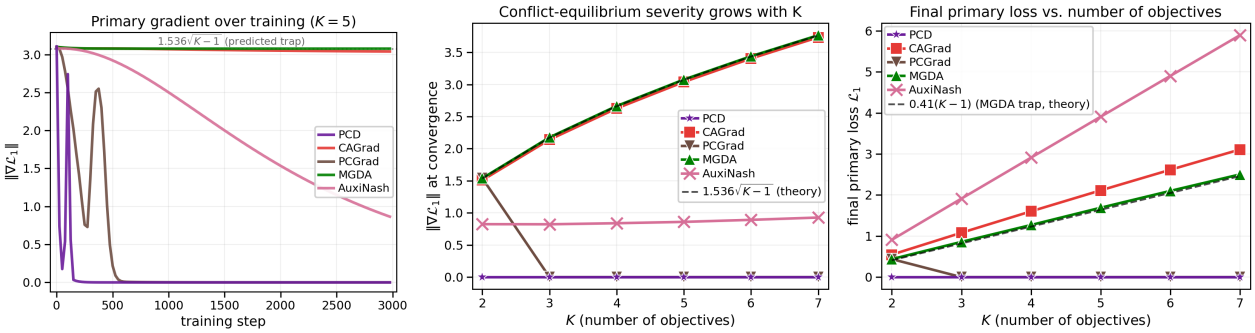
$$\mathcal{L}_1(\theta) = \sum_{j=1}^{K-1} \left(s_j^2 - 1\right)^2 + \tfrac{1}{2}\|P_\perp \theta\|^2,$$

where $P_\perp$ projects onto the orthogonal complement of $\mathrm{span}\{u_j\}$. Each well $W(s) = (s^2-1)^2$ has minima at $s = \pm 1$ and a barrier at $s = 0$, and the j-th secondary pulls $s_j \to +1$, so the joint optimum $\theta^\star = \sum_j u_j$ lies across the barrier from the initialization. At the shared start $s_j = -0.6$ the primary-gradient component along each u_j has magnitude $|W'(-0.6)| = |4(-0.6)\big((-0.6)^2 - 1\big)| = 1.536$, near the steepest point of the barrier; since the u_j are orthonormal, these $K-1$ components add in quadrature, giving $\|\nabla \mathcal{L}_1\| = 1.536\sqrt{K-1}$ (the dashed line in Fig. 11b). We sweep $K \in \{2, \ldots, 7\}$ over 5 seeds.

Combination methods are pinned at the equilibrium. A conflict equilibrium is a point at which $0 \in \mathrm{conv}\{g_1, \ldots, g_K\}$ while no objective is individually stationary. This halting is not specific to our setting: any method whose update lies in $\mathrm{conv}\{g_i\}$ (MGDA exactly, and CAGrad and WS up to their trust region and weighting) returns 0 precisely when $0 \in \mathrm{conv}\{g_i\}$, which is the defining condition of a conflict equilibrium. MGDA returns the minimum-norm element of $\mathrm{conv}\{g_i\}$, CAGrad anchors on their average within a trust region, and WS forms a fixed combination. What *is* specific to our construction is the rate at which the uncanceled primary gradient grows with K: Fig. 11(a) and (b) show MGDA and CAGrad halting exactly at the equilibrium, the canceled primary gradient tracking the $1.536\sqrt{K-1}$ law derived above (dashed line) across the entire range.

The primary loss they leave unconverged grows linearly in K (Fig. 11(c)). WS is omitted from the sweep because its behavior is weight-dependent and does not reduce to a single curve in K. For any primary-respecting weight it traps with the same combination-method geometry.

PCD reaches CMS for every K. PCD's update is $\tilde{g}_1 + \sum_{j\geq 2} \mu_j \tilde{g}_j$ with the primary coefficient pinned at one and $\mu_j \geq 0$ (Theorem 4.2). It is structurally *not* a convex combination of the gradients, and Theorem 4.8 guarantees it vanishes only when every gradient does. Across all K, PCD crosses every barrier and converges to the CMS point for every $\tau > 0$: its primary gradient at convergence is at machine precision ($\sim 10^{-15}$) and its primary loss is $\sim 10^{-29}$, against the unbounded growth of the combination methods. The learning curve at $K = 5$ (Fig. 11(a)) makes the distinction visible. PCD drives the primary gradient to zero within a few hundred steps while MGDA and CAGrad stay pinned at 3.07.



(a) $\|\nabla\mathcal{L}_1\|$ over training (K=5) (b) $\|\nabla\mathcal{L}_1\|$ at convergence vs. K (c) Primary loss left vs. K

Figure 11: **Conflict-equilibrium escape as K grows** (5 seeds). (a) At K=5, PCD drives the primary gradient to zero while MGDA/CAGrad remain pinned at $1.536\sqrt{K-1}$, PCGrad crosses after an initial stall and AuxiNash drifts slowly. (b) The canceled primary gradient of the combination methods (MGDA, CAGrad) tracks $1.536\sqrt{K-1}$ (dashed) across all K, while PCD sits at zero. (c) The unconverged primary loss grows linearly in K for the trapped methods. PCGrad ($K \geq 3$) and PCD reach CMS; AuxiNash leaves the most loss. See text for the combination-vs-surgery distinction and the separability caveat.

Two qualifications accompany this result. First, PCGrad and AuxiNash are not combination methods and are not bound to the equilibrium: PCGrad traps at $K = 2$ but, for $K \geq 3$, its pairwise gradient projection lets the orthogonal single-direction secondaries pull through, so it too reaches CMS in this construction. AuxiNash, which learns to down-weight the conflicting auxiliaries, neither sits at the equilibrium nor reaches CMS but drifts into the high-loss barrier region, leaving the largest primary loss of any method. Second, the construction is separable across the u_j. This is what makes the $\sqrt{K-1}$ law exactly derivable, and we present it as such. The conflict-equilibrium structure nonetheless requires all K gradients to balance with equal weight. The $\sqrt{K-1}$ law thus characterizes the combination methods it is derived for, and PCGrad's escape is specific to this clean separable geometry. In the non-separable compression landscapes of Sec. 5.1, PCGrad is dominated at every compression target. The conclusion is the one the theory predicts: PCD's only fixed point is CMS, so it is the only method that reaches the joint optimum *by construction* rather than incidentally, and it does so uniformly as the number of competing objectives grows.

6 Conclusion

Multi-objective optimization in deep learning has long been limited in its applications by an artificial symmetry. By treating primary tasks and secondary constraints as equals, standard MOO methods routinely trap optimization trajectories in conflict equilibria. *Priority-Constrained Descent* (PCD) dismantles this symmetry. By anchoring updates strictly to the primary gradient and admitting only the minimal Euclidean distortion necessary to maintain secondary progress, PCD guarantees that constraints are satisfied without sacrificing the primary deliverable. Governed by a single, interpretable scalar τ, the framework is scale-invariant, admits exact closed-form solutions for practical objective counts, and vanishes only at composite multi-stationary points. As demonstrated across neural network compression and synthetic ex-

periments, PCD provides exactly what hierarchical pipelines demand: a stable, controllable trade-off curve that consistently dominates symmetric baselines.

The theoretical guarantees of PCD rely on secondary objectives being proper, convex, and lower semi-continuous. Consequently, inherently non-convex constraints such as adversarial losses or hard combinatorial sparsity targets fall outside our formal geometric proofs, which apply locally to the normalized first-order step. Furthermore, while the composite multi-stationarity theorem characterizes the exact fixed-point structure of the deterministic update, the underlying quadratic program could theoretically become infeasible for $K > 2$ if the secondary constraints are strictly and mutually incompatible. Although demonstrated in Sec. 5.3 to be improbable, we note this as a possible limitation regardless. While our synthetic experiments scale up to $K = 8$ to rigorously stress-test the framework against traditional MOO baselines, we note that typical asymmetric optimization applications rarely necessitate such a large number of distinct objectives in practice.

PCD establishes a foundation for hierarchical optimization, opening several immediate avenues for extension. The most direct is the development of *adaptive τ schedules*, which would dynamically anneal constraint strength based on the training trajectory or current feasibility (as outlined in App. B.1). Beyond the strict primary–secondary dichotomy, extending the formulation to *multi-level priority hierarchies* where an arbitrary number of objectives are strictly ordered with distinct, cascading tolerances, offers a natural mathematical next step. More broadly, the per-objective tolerances of Remark B.1 trace a $(K-1)$-dimensional operating manifold within the tolerance hypercube $[0,1]^{K-1}$, of which scalar τ is the diagonal slice and cascading hierarchies an ordered sub-family. The open challenge is recovering this expressiveness by setting the τ_j adaptively or via a learned criterion, rather than by exponential grid search over the cube, so that PCD's single-knob interpretability is extended rather than spent. On the application side, deploying PCD under strict, hardware-aware constraints, or adapting it for structured generative modeling where primary output fidelity must survive heavy structural regularizers, represent highly promising frontiers for the framework.

References

Hamza A Abushahla, Dara Varam, Ariel Justine N Panopio, and Mohamed I AlHajri. Neural network quantization for microcontrollers: A comprehensive survey of methods, platforms, and applications. *arXiv preprint arXiv:2508.15008*, 2025.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803. PMLR, 2018.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda). 2009.

N Dinh and V Jeyakumar. Farkas' lemma: three decades of generalizations for mathematical optimization. *Top*, 22(1):1–22, 2014.

Heshan Fernando, Lisha Chen, Songtao Lu, Pin-Yu Chen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, Gaowen Liu, Meng Wang, and Tianyi Chen. Variance reduction can improve trade-off in multi-objective learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6975–6979. IEEE, 2024.

Heshan Devaka Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent approach. In *The Eleventh International Conference On Learning Representations*, 2023.

Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical methods of operations research*, 51(3):479–494, 2000.

David Gale, Harold W. Kuhn, and Albert W. Tucker. Linear programming and the theory of games. *Activity analysis of production and allocation*, 13:317–335, 1951.

Jose Gallego-Posada, Juan Ramirez, Meraj Hashemizadeh, and Simon Lacoste-Julien. Cooper: A library for constrained optimization in deep learning. *arXiv preprint arXiv:2504.01212*, 2025.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241. PMLR, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(5):2900–2919, 2023.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Yerlan Idelbayev and Miguel A Carreira-Perpinán. Low-rank compression of neural nets: Learning the rank of each layer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8049–8059, 2020.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v28/jaggi13.html`.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021a.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. Famo: Fast adaptive multitask optimization. *Advances in Neural Information Processing Systems*, 36:57226–57243, 2023.

Liyang Liu, Yi Li, Zhanghui Kuang, Jing-Hao Xue, Yimin Chen, Wenming Yang, Qingmin Liao, and Wayne Zhang. Towards impartial multi-task learning. In *International conference on learning representations*, 2021b.

Debabrata Mahapatra and Vaibhav Rajan. Exact pareto optimal search for multi-task learning and multi-criteria decision-making. *arXiv preprint arXiv:2108.00597*, 2021.

R Timothy Marler and Jasbir S Arora. The weighted sum method for multi-objective optimization: new insights. *Structural and multidisciplinary optimization*, 41(6):853–862, 2010.

Kaisa Miettinen. *Nonlinear multiobjective optimization*, volume 12. Springer Science & Business Media, 1999.

Aviv Navon, Aviv Shamsian, Gal Chechik, and Ethan Fetaya. Learning the pareto front with hypernetworks. *arXiv preprint arXiv:2010.04104*, 2020.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning*, pp. 16428–16446. PMLR, 2022.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018.

Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20083–20093, 2023.

Aviv Shamsian, Aviv Navon, Neta Glazer, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Auxiliary learning as an asymmetric bargaining game. In *International Conference on Machine Learning*, pp. 30689–30705. PMLR, 2023.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5362–5383, 2024.

Derrick Xin, Behrooz Ghorbani, Justin Gilmer, Ankush Garg, and Orhan Firat. Do current multi-task optimization methods in deep learning even help? *Advances in neural information processing systems*, 35: 13597–13609, 2022.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020.

Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. A survey on federated learning. *Knowledge-Based Systems*, 216:106775, 2021.

Xiaoyuan Zhang, Xi Lin, and Qingfu Zhang. Pmgda: A preference-based multiple gradient descent algorithm. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.

A Roadmap of results

Fig. 12 provides a clickable map of the paper's theoretical structure: every node is a hyperlink to the corresponding statement, so the figure doubles as a table of contents for the technical material below.

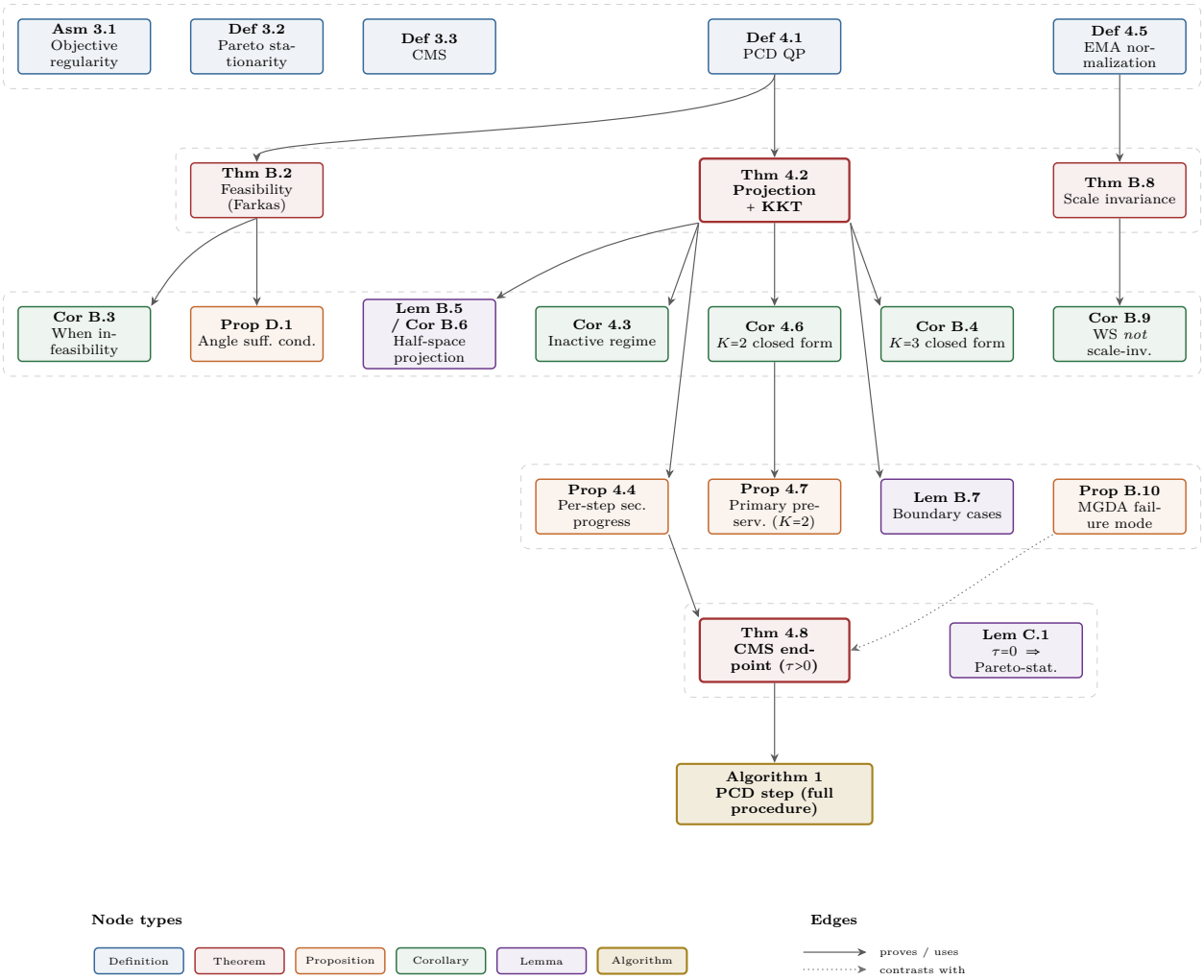


Figure 12: Theory navigation map for PCD. Foundations (definitions and the regularity assumption) sit at the top. The three core theorems (Thm B.2: feasibility via Farkas, Thm 4.2: projection + KKT characterization, Thm B.8: scale invariance) occupy the second row. Their direct corollaries, the per-step guarantees, the asymptotic endpoint (Thm 4.8), and the τ=0 corner case (Lem C.1) form the lower tiers, culminating in Algorithm 1. The dotted edge from Prop B.10 marks a contrast: under matched magnitudes MGDA's mixing coefficient collapses, whereas PCD avoids this failure. Every node is a hyperlink to the corresponding statement in the paper.

B Full Details and Proofs for Section 4

Remark B.1 (Per-objective tolerances)**.** *The main text and the appendix use a single scalar tolerance $\tau \in [0,1]$ that applies uniformly to every secondary objective. The framework admits a trivial generalization to per-objective $\tau_j \in [0,1]$, in which the constraint of Eq. (4) becomes $\tilde{\mathbf{g}}_j^\top \tilde{\mathbf{d}} \geq \tau_j \|\tilde{\mathbf{g}}_j\|^2$. All proofs and the algorithm carry through verbatim under this substitution. We work with scalar τ throughout because all experiments and ablations in this paper use a single shared tolerance; per-objective τ_j adds expressive flexibility without changing the underlying analysis.*

B.1 Feasibility Analysis and Adaptive Tolerance Scaling

$K = 2$ **feasibility.** For $K = 2$ with $\tilde{\mathbf{g}}_2 \neq \mathbf{0}$, the point $\tilde{\mathbf{d}} = \tau\tilde{\mathbf{g}}_2$ satisfies $\tilde{\mathbf{g}}_2^\top\tilde{\mathbf{d}} = \tau\|\tilde{\mathbf{g}}_2\|^2$, so $F(\tau) = H_2 \neq \varnothing$ for all $\tau \in [0,1]$.

General K via Farkas' lemma. Define $\widetilde{G} \in \mathbb{R}^{(K-1)\times n}$ with rows $\tilde{\mathbf{g}}_2^\top, \ldots, \tilde{\mathbf{g}}_K^\top$ and $\mathbf{b} \in \mathbb{R}^{K-1}$ with $b_j = \tau\|\tilde{\mathbf{g}}_j\|^2$. By Farkas' lemma, $F(\tau) = \varnothing$ iff $\exists\, \mathbf{y} \geq \mathbf{0}$, $\mathbf{y} \neq \mathbf{0}$, with $\widetilde{G}^\top\mathbf{y} = \mathbf{0}$ and $\mathbf{b}^\top\mathbf{y} > 0$.

Theorem B.2 (Feasibility characterization)**.** *With $\tau > 0$,*

$$F(\tau) = \varnothing \quad \Longleftrightarrow \quad \mathbf{0} \in \operatorname{cone}\{\tilde{\mathbf{g}}_j : j \in [K]_{-1},\, \tilde{\mathbf{g}}_j \neq \mathbf{0}\} - \{\mathbf{0}\}.$$

Proof. $\widetilde{G}^\top\mathbf{y} = \mathbf{0}$ is exactly $\sum_{j\geq 2} y_j\tilde{\mathbf{g}}_j = \mathbf{0}$ (positive dependence). With $\tau > 0$ and $y_j \geq 0$, the condition $\mathbf{b}^\top\mathbf{y} = \sum_j y_j\tau\|\tilde{\mathbf{g}}_j\|^2 > 0$ is automatically satisfied whenever any $y_j > 0$ corresponds to $\tilde{\mathbf{g}}_j \neq \mathbf{0}$. □

Corollary B.3 (When infeasibility occurs)**.**
- *(i) $K = 2$: a single nonzero ray cannot positively combine to zero, so $F(\tau) \neq \varnothing$.*
- *(ii) $K = 3$: $y_2\tilde{\mathbf{g}}_2 + y_3\tilde{\mathbf{g}}_3 = \mathbf{0}$ with $y_2, y_3 \geq 0$ not both zero requires $\tilde{\mathbf{g}}_3 = -\lambda\tilde{\mathbf{g}}_2$ ($\lambda > 0$), i.e. exact anti-parallelism.*
- *(iii) General K, $n \gg K$: $\mathbf{0} \in \operatorname{cone}\{\tilde{\mathbf{g}}_2, \ldots, \tilde{\mathbf{g}}_K\}$ defines a variety of positive codimension in $(\mathbb{R}^n)^{K-1}$. Infeasibility is Lebesgue measure-zero.*

B.2 KKT Proof and Active-Set Solver

Proof of Theorem 4.2. The Lagrangian of Eq. (4) is

$$\mathcal{L}(\tilde{\mathbf{d}}, \boldsymbol{\mu}) = \tfrac{1}{2}\left\|\tilde{\mathbf{d}} - \tilde{\mathbf{g}}_1\right\|^2 - \sum_{j=2}^{K} \mu_j(\tilde{\mathbf{g}}_j^\top\tilde{\mathbf{d}} - \tau\|\tilde{\mathbf{g}}_j\|^2), \quad \mu_j \geq 0.$$

Setting $\nabla_{\tilde{\mathbf{d}}}\mathcal{L} = \mathbf{0}$ gives $\tilde{\mathbf{d}} - \tilde{\mathbf{g}}_1 - \sum_j \mu_j\tilde{\mathbf{g}}_j = \mathbf{0}$, i.e. stationarity (i). Primal feasibility (ii), dual feasibility (iii), and complementary slackness (iv) are the standard KKT conditions for inequality-constrained convex QPs with affine constraints. For such problems KKT is necessary and sufficient at every primal optimum without further constraint qualification. □

Active-set linear system. For active set $\mathcal{A} \subseteq [K]_{-1}$, let $G_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|}$ be the Gram matrix with $(G_{\mathcal{A}})_{jk} = \tilde{\mathbf{g}}_j^\top\tilde{\mathbf{g}}_k$, and $\mathbf{b}_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ with $(\mathbf{b}_{\mathcal{A}})_j = \tau\|\tilde{\mathbf{g}}_j\|^2 - \tilde{\mathbf{g}}_j^\top\tilde{\mathbf{g}}_1$. Substituting $\tilde{\mathbf{d}}^* = \tilde{\mathbf{g}}_1 + \sum_{k\in\mathcal{A}} \mu_k^*\tilde{\mathbf{g}}_k$ into the active equality constraints $\tilde{\mathbf{g}}_j^\top\tilde{\mathbf{d}}^* = \tau\|\tilde{\mathbf{g}}_j\|^2$, $j \in \mathcal{A}$:

$$G_{\mathcal{A}}\,\boldsymbol{\mu}_{\mathcal{A}} = \mathbf{b}_{\mathcal{A}}. \tag{19}$$

$G_{\mathcal{A}}$ is positive definite whenever $\{\tilde{\mathbf{g}}_j\}_{j\in\mathcal{A}}$ are linearly independent (generic in $\mathbb{R}^n$ for $|\mathcal{A}| \leq n$).

Algorithm. Standard primal active-set: initialize $\mathcal{A} = \varnothing$, $\tilde{\mathbf{d}}^* = \tilde{\mathbf{g}}_1$. Repeatedly add the most-violated constraint, solve Eq. (19), drop any constraint with $\mu_j < 0$, and re-solve. The procedure terminates in at most $\binom{K-1}{|\mathcal{A}|}$ iterations. Per-step cost is $O(K^2n)$, dominated by Gram-matrix construction. This matches MGDA and is negligible relative to the $O(Kn)$ cost of the K backpropagations.

B.3 K=2 and K=3 Closed Forms

Proof of Cor. 4.6. For $K = 2$ the active set is either $\varnothing$ (if $\tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_1 \geq \tau\|\tilde{\mathbf{g}}_2\|^2$, KKT gives $\tilde{\mathbf{d}}^* = \tilde{\mathbf{g}}_1$) or $\{2\}$ (otherwise). In the latter case, Eq. (19) reduces to the scalar equation $\|\tilde{\mathbf{g}}_2\|^2\mu_2^* = \tau\|\tilde{\mathbf{g}}_2\|^2 - \tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_1$, giving $\mu_2^* = \tau - \tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_1/\|\tilde{\mathbf{g}}_2\|^2$. □

Corollary B.4 (Closed form for $K = 3$, both constraints active)**.** *For $K = 3$ with $\mathcal{A} = \{2, 3\}$, the 2×2 Gram system gives*

$$\mu_2^* = \frac{\|\tilde{\mathbf{g}}_3\|^2(\tau\|\tilde{\mathbf{g}}_2\|^2 - \tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_1) - (\tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_3)(\tau\|\tilde{\mathbf{g}}_3\|^2 - \tilde{\mathbf{g}}_3^\top\tilde{\mathbf{g}}_1)}{\Delta},$$
$$\mu_3^* = \frac{\|\tilde{\mathbf{g}}_2\|^2(\tau\|\tilde{\mathbf{g}}_3\|^2 - \tilde{\mathbf{g}}_3^\top\tilde{\mathbf{g}}_1) - (\tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_3)(\tau\|\tilde{\mathbf{g}}_2\|^2 - \tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_1)}{\Delta},$$

with $\Delta = \|\tilde{\mathbf{g}}_2\|^2 \|\tilde{\mathbf{g}}_3\|^2 - (\tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_3)^2 = \|\tilde{\mathbf{g}}_2\|^2 \|\tilde{\mathbf{g}}_3\|^2 \sin^2\theta_{23}$. Uniquely solvable iff $\tilde{\mathbf{g}}_2, \tilde{\mathbf{g}}_3$ are not collinear.

Proof. Cramer's rule applied to $G_{\mathcal{A}}\boldsymbol{\mu}_{\mathcal{A}} = \mathbf{b}_{\mathcal{A}}$ with $G_{\mathcal{A}} = \begin{pmatrix} \|\tilde{\mathbf{g}}_2\|^2 & \tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_3 \\ \tilde{\mathbf{g}}_2^\top\tilde{\mathbf{g}}_3 & \|\tilde{\mathbf{g}}_3\|^2 \end{pmatrix}$. □

B.4 Geometric Interpretation: Half-Space Projection

Lemma B.5 (Projection onto a half-space)**.** *For $H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top\mathbf{x} \geq c\}$ with $\mathbf{a} \neq \mathbf{0}$, the orthogonal projection of $\mathbf{p}$ onto H is*

$$\mathrm{proj}_H(\mathbf{p}) = \begin{cases} \mathbf{p} & \text{if } \mathbf{a}^\top\mathbf{p} \geq c, \\ \mathbf{p} + \dfrac{c - \mathbf{a}^\top\mathbf{p}}{\|\mathbf{a}\|^2}\mathbf{a} & \text{otherwise.} \end{cases}$$

Proof. If $\mathbf{p} \in H$, the closest point is $\mathbf{p}$ itself. If $\mathbf{p} \notin H$, the projection lies on the boundary hyperplane $\{\mathbf{x} : \mathbf{a}^\top\mathbf{x} = c\}$ at $\mathbf{p} + \frac{c - \mathbf{a}^\top\mathbf{p}}{\|\mathbf{a}\|^2}\mathbf{a}$. □

Corollary B.6 ($K = 2$ as half-space projection)**.** *$\tilde{\mathbf{d}}^* = \mathrm{proj}_{H_2}(\tilde{\mathbf{g}}_1)$, the orthogonal projection of $\tilde{\mathbf{g}}_1$ onto $H_2 = \{\tilde{\mathbf{d}} : \tilde{\mathbf{g}}_2^\top\tilde{\mathbf{d}} \geq \tau\|\tilde{\mathbf{g}}_2\|^2\}$.*

For general K, $\tilde{\mathbf{d}}^* = \mathrm{proj}_{F(\tau)}(\tilde{\mathbf{g}}_1)$, the orthogonal projection of $\tilde{\mathbf{g}}_1$ onto the convex polyhedron $F(\tau)$. The active set $\mathcal{A}$ identifies the faces of $F(\tau)$ containing the projection.

B.5 Boundary Cases

Lemma B.7 (Boundary behaviors)**.**

(i) All secondary gradients zero ($\tilde{\mathbf{g}}_j = \mathbf{0}$ for all $j \geq 2$): all constraints reduce to $0 \geq 0$, the QP is unconstrained, and $\tilde{\mathbf{d}}^* = \tilde{\mathbf{g}}_1$ (gradient descent on $\mathcal{L}_1$).

(ii) Primary gradient zero, some $\tau > 0$, some $\tilde{\mathbf{g}}_j \neq \mathbf{0}$ ($\tilde{\mathbf{g}}_1 = \mathbf{0}$): the QP becomes $\min\|\tilde{\mathbf{d}}\|^2$ subject to the secondary constraints. The solution $\tilde{\mathbf{d}}^* = \sum_{j\in\mathcal{A}}\mu_j^*\tilde{\mathbf{g}}_j$ is a non-negative combination of secondary gradients—PCD continues to make progress on secondary objectives even though the primary is stationary.

(iii) All gradients zero: $\tilde{\mathbf{d}}^* = \mathbf{0}$ (fixed point).

Proof. (i) Trivial. (ii) With $\tilde{\mathbf{g}}_1 = \mathbf{0}$, KKT stationarity gives $\tilde{\mathbf{d}}^* = \sum_j \mu_j^*\tilde{\mathbf{g}}_j$. The objective $\|\tilde{\mathbf{d}} - \mathbf{0}\|^2 = \|\tilde{\mathbf{d}}\|^2$ is minimized at the closest feasible point to the origin. Since $\tau > 0$ and $\tilde{\mathbf{g}}_j \neq \mathbf{0}$ for some j, the constraint $\tilde{\mathbf{g}}_j^\top\tilde{\mathbf{d}} \geq \tau\|\tilde{\mathbf{g}}_j\|^2 > 0$ forces $\tilde{\mathbf{d}}^* \neq \mathbf{0}$. (iii) Immediate. □

Lemma B.7(ii) is the geometric mechanism by which PCD avoids conflict equilibria: at near-primary-stationary points, MGDA outputs $\mathbf{d}_{\mathrm{MGDA}} \approx \mathbf{0}$ (since $\mathbf{0} \in \mathrm{conv}\{\mathbf{0}, \mathbf{g}_2, \ldots, \mathbf{g}_K\}$), halting all progress, while PCD continues to drive secondary objectives toward stationarity.

B.6 Scale Invariance Proof

We formally state the scale-invariance result whose paragraph-level summary appears in Sec. 4.5.

Theorem B.8 (Asymptotic scale invariance)**.** *Let $c_1, \ldots, c_K > 0$ and define $\mathcal{L}_i'(\boldsymbol{\theta}) = c_i\mathcal{L}_i(\boldsymbol{\theta})$. As $\varepsilon \to 0$ (or once $\hat{v}_{i,t} \gg \varepsilon/c_i^2$ for all i), the PCD direction computed from $(\mathcal{L}_1', \ldots, \mathcal{L}_K')$ coincides with that computed from $(\mathcal{L}_1, \ldots, \mathcal{L}_K)$. With the alternative normalization $s_{i,t} = (\hat{v}_{i,t}(1 + \varepsilon))^{-1/2}$, scale invariance is exact for all $\varepsilon \geq 0$.*

Proof. Under $\mathcal{L}_i \to c_i\mathcal{L}_i$, $\mathbf{g}_i \to c_i\mathbf{g}_i$. The EMA buffer transforms as $v_{i,t} \to c_i^2 v_{i,t}$, so $\hat{v}_{i,t} \to c_i^2\hat{v}_{i,t}$ and

$$s_{i,t} = (\hat{v}_{i,t} + \varepsilon)^{-1/2} \;\to\; (c_i^2\hat{v}_{i,t} + \varepsilon)^{-1/2}.$$

For $\varepsilon \ll c_i^2\hat{v}_{i,t}$ (which holds once the EMA stabilizes), $s_{i,t} \approx 1/(c_i\sqrt{\hat{v}_{i,t}}) = s_{i,t}^{\text{orig}}/c_i$. Therefore

$$\tilde{\mathbf{g}}_i' = s_{i,t}'\,\mathbf{g}_i' \;\approx\; \tfrac{s_{i,t}^{\text{orig}}}{c_i}\cdot c_i\mathbf{g}_i \;=\; s_{i,t}^{\text{orig}}\mathbf{g}_i \;=\; \tilde{\mathbf{g}}_i.$$

Since $\tilde{\mathbf{g}}_i' = \tilde{\mathbf{g}}_i$ for all i, the QP in Eq. (4) is identical, and $\tilde{\mathbf{d}}^*$ is unchanged. For the alternative normalization, the factor $(1+\varepsilon)$ is invariant under rescaling, so $s_{i,t}'/s_{i,t} = 1/c_i$ exactly, and $\tilde{\mathbf{g}}_i' = \tilde{\mathbf{g}}_i$ holds for every $\varepsilon \geq 0$. □

Corollary B.9 (Weighted-sum methods are not scale invariant)**.** $\mathbf{d}_{\text{WS}} = \sum_i w_i\mathbf{g}_i$ *transforms to* $\sum_i w_i c_i \mathbf{g}_i$*, which generally differs from* $\mathbf{d}_{\text{WS}}$*. The WS direction depends on the relative magnitudes of the objectives—an arbitrary, task-dependent quantity.*

B.7 Per-Step Progress and MGDA Failure

Proof of Prop. 4.4. Immediate from primal feasibility (Theorem 4.2(ii)) with $\tau > 0$ and $\|\tilde{\mathbf{g}}_j\| > 0$. □

Proposition B.10 (MGDA failure mode)**.** *For* $K \geq 2$*,* $\mathbf{d}_{\text{MGDA}} = \arg\min_{\mathbf{d}\in\text{conv}\{\mathbf{g}_1,\ldots,\mathbf{g}_K\}} \|\mathbf{d}\|^2$*. If* $\mathbf{g}_1 = \mathbf{0}$*, then* $\mathbf{d}_{\text{MGDA}} = \mathbf{0}$*, regardless of* $\mathbf{g}_2, \ldots, \mathbf{g}_K$*.*

Proof. $\mathbf{0} = \mathbf{g}_1 \in \text{conv}\{\mathbf{g}_1, \ldots, \mathbf{g}_K\}$, and $\|\mathbf{0}\|^2 = 0$ is the global minimum. Hence $\mathbf{d}_{\text{MGDA}} = \mathbf{0}$. □

This makes precise the structural advantage of PCD over MGDA: MGDA halts whenever *any* single gradient vanishes (even if $K-1$ secondary objectives remain non-stationary), whereas PCD with $\tau > 0$ continues to drive every non-stationary secondary toward stationarity (Lemma B.7(ii)).

C Additional Convergence Properties

C.1 Proof of Theorem 4.8

We restate Theorem 4.8 for convenience and give the full proof.

Theorem (Theorem 4.8, restated)**.** *Let* $\tau > 0$ *and suppose* $F_\tau(\boldsymbol{\theta}) \neq \varnothing$*. Then*

$$\tilde{\mathbf{d}}^*(\boldsymbol{\theta}) \;=\; \mathbf{0} \quad\Longleftrightarrow\quad \tilde{\mathbf{g}}_i(\boldsymbol{\theta}) \;=\; \mathbf{0} \;\; \textit{for all } i \in [K].$$

Proof. ($\Leftarrow$) If $\tilde{\mathbf{g}}_i = \mathbf{0}$ for all $i \in [K]$, then $\tilde{\mathbf{d}} = \mathbf{0}$ achieves $\|\mathbf{0} - \mathbf{0}\|^2 = 0$ (the global minimum of the QP objective) and trivially satisfies every constraint $\mathbf{0}^\top\mathbf{0} = 0 \geq 0$. Hence $\tilde{\mathbf{d}}^* = \mathbf{0}$.

($\Rightarrow$) Suppose $\tilde{\mathbf{d}}^* = \mathbf{0}$. By primal feasibility (Theorem 4.2),

$$\tilde{\mathbf{g}}_j^\top\tilde{\mathbf{d}}^* \;\geq\; \tau\,\|\tilde{\mathbf{g}}_j\|^2 \qquad \forall\, j \in [K]_{-1}.$$

Substituting $\tilde{\mathbf{d}}^* = \mathbf{0}$ gives $0 \geq \tau\,\|\tilde{\mathbf{g}}_j\|^2$. Since $\tau > 0$ and $\|\tilde{\mathbf{g}}_j\|^2 \geq 0$, this forces $\tilde{\mathbf{g}}_j = \mathbf{0}$ for all $j \in [K]_{-1}$. By KKT stationarity (Theorem 4.2, Eq. (7)),

$$\mathbf{0} \;=\; \tilde{\mathbf{d}}^* \;=\; \tilde{\mathbf{g}}_1 + \sum_{j=2}^{K}\mu_j^*\,\tilde{\mathbf{g}}_j \;=\; \tilde{\mathbf{g}}_1 + \sum_{j=2}^{K}\mu_j^*\,\mathbf{0} \;=\; \tilde{\mathbf{g}}_1,$$

so $\tilde{\mathbf{g}}_1 = \mathbf{0}$. Since $\tilde{\mathbf{g}}_i = s_i\mathbf{g}_i$ with $s_i > 0$, this yields $\nabla\mathcal{L}_1(\boldsymbol{\theta}) = \mathbf{0}$ and $\mathbf{0} \in \partial\mathcal{L}_j(\boldsymbol{\theta})$ for $j \in [K]_{-1}$, i.e. $\boldsymbol{\theta}$ is composite multi-stationary. □

The proof uses only primal feasibility with $\tau > 0$ and KKT stationarity. It does not depend on K, on n, or on algebraic structure of the multipliers or Gram matrix, which is what makes the result hold uniformly across problem sizes.

C.2 Fixed Points with $\tau = 0$

Lemma C.1. *If $\tau = 0$, then $\tilde{\mathbf{d}}^* = \mathbf{0}$ if and only if $-\tilde{\mathbf{g}}_1 \in \operatorname{cone}\{\tilde{\mathbf{g}}_2, \dots, \tilde{\mathbf{g}}_K\}$. Equivalently, $\tilde{\mathbf{d}}^* = \mathbf{0}$ iff $\mathbf{0} \in \operatorname{conv}\{\tilde{\mathbf{g}}_1, \tilde{\mathbf{g}}_2, \dots, \tilde{\mathbf{g}}_K\}$* with strictly positive weight on $\tilde{\mathbf{g}}_1$. *In particular, $\tilde{\mathbf{d}}^* = \mathbf{0} \Rightarrow \boldsymbol{\theta}$ is Pareto-stationary in normalized-gradient space.*

Remark C.2 (On the strict-positivity condition)**.** *The condition $\mathbf{0} \in \operatorname{conv}\{\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_K\}$ alone (Pareto stationarity) is* not *equivalent to $\tilde{\mathbf{d}}^* = \mathbf{0}$. Counter-example: in $\mathbb{R}^2$ with $\tilde{\mathbf{g}}_1 = (1,0)$, $\tilde{\mathbf{g}}_2 = (0,1)$, $\tilde{\mathbf{g}}_3 = (0,-1)$, the QP forces $d_2 = 0$ and yields $\tilde{\mathbf{d}}^* = (1,0) = \tilde{\mathbf{g}}_1 \neq \mathbf{0}$, while $\mathbf{0} = 0 \cdot \tilde{\mathbf{g}}_1 + \frac{1}{2}\tilde{\mathbf{g}}_2 + \frac{1}{2}\tilde{\mathbf{g}}_3$ (Pareto-stationary, but with zero weight on $\tilde{\mathbf{g}}_1$). Lemma C.1 uses only the implication $\tilde{\mathbf{d}}^* = \mathbf{0} \Rightarrow$ Pareto-stationary, which holds without further qualification.*

Proof. The QP becomes $\min \left\|\tilde{\mathbf{d}} - \tilde{\mathbf{g}}_1\right\|^2$ s.t. $\tilde{\mathbf{g}}_j^\top \tilde{\mathbf{d}} \geq 0$ for $j \geq 2$.

($\Leftarrow$) If $-\tilde{\mathbf{g}}_1 = \sum_{j\geq 2} \mu_j \tilde{\mathbf{g}}_j$ with $\mu_j \geq 0$, set $\tilde{\mathbf{d}} = \mathbf{0}$. Primal feasibility: $\tilde{\mathbf{g}}_j^\top \mathbf{0} = 0 \geq 0$. Dual feasibility: $\mu_j \geq 0$. Complementary slackness: $\mu_j \cdot 0 = 0$. Stationarity: $\mathbf{0} - \tilde{\mathbf{g}}_1 - \sum_j \mu_j \tilde{\mathbf{g}}_j = -\tilde{\mathbf{g}}_1 - (-\tilde{\mathbf{g}}_1) = \mathbf{0}$. Hence $\tilde{\mathbf{d}}^* = \mathbf{0}$ by KKT.

($\Rightarrow$) If $\tilde{\mathbf{d}}^* = \mathbf{0}$, KKT stationarity gives $\mathbf{0} = \tilde{\mathbf{g}}_1 + \sum_j \mu_j^* \tilde{\mathbf{g}}_j$ with $\mu_j^* \geq 0$, i.e. $-\tilde{\mathbf{g}}_1 = \sum_j \mu_j^* \tilde{\mathbf{g}}_j \in \operatorname{cone}\{\tilde{\mathbf{g}}_j\}_{j\geq 2}$.

Equivalence with $\mathbf{0} \in \operatorname{conv}$ (with $\alpha_1 > 0$) and Pareto stationarity. Given $-\tilde{\mathbf{g}}_1 = \sum_j \mu_j^* \tilde{\mathbf{g}}_j$ with $\mu_j^* \geq 0$, dividing by $1 + \sum_j \mu_j^* > 0$ gives

$$\mathbf{0} = \tfrac{1}{1+\sum_j \mu_j^*}\, \tilde{\mathbf{g}}_1 + \sum_j \tfrac{\mu_j^*}{1+\sum_j \mu_j^*}\, \tilde{\mathbf{g}}_j \in \operatorname{conv}\{\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_K\},$$

with weight $\alpha_1 = 1/(1 + \sum \mu_j^*) > 0$ on $\tilde{\mathbf{g}}_1$, hence Pareto-stationary. □

This recovers the MGDA-style guarantee at $\tau = 0$ and contrasts sharply with Theorem 4.8: positive tolerances collapse the fixed-point set from Pareto-stat to CMS only.

C.3 Sufficient-Decrease Sketch

Theorem 4.8 characterizes the endpoints of the deterministic, normalized QP but does not by itself certify that the iterates of Algorithm 1 approach such an endpoint. A standard sufficient-decrease argument applied to the primary loss yields the following partial convergence statement, sketched here for completeness; we do not claim it as a contribution of this paper.

Suppose $\mathcal{L}_1$ is ℓ_1-smooth and bounded below, the QP is feasible at every iterate, the iterate sequence is bounded, and there exists $\mu_0 > 0$ such that the cosine alignment $\widetilde{\alpha}_t := \tilde{\mathbf{g}}_{1,t}^\top \tilde{\mathbf{d}}_t^* / (\|\tilde{\mathbf{g}}_{1,t}\| \left\|\tilde{\mathbf{d}}_t^*\right\|) \geq \mu_0$ for all t. Then for constant step size $\eta \leq \mu_0/\ell_1$, the descent lemma applied to $\mathcal{L}_1$ together with the rescaling identity $\|\mathbf{d}_t^*\| = \|\mathbf{g}_{1,t}\|$ yields $\mathcal{L}_1(\boldsymbol{\theta}_{t+1}) \leq \mathcal{L}_1(\boldsymbol{\theta}_t) - \frac{\eta\mu_0}{2} \|\mathbf{g}_{1,t}\|^2$, and telescoping over $t < T$ gives $\min_{t<T} \|\mathbf{g}_{1,t}\|^2 = O(1/T)$. Two caveats are worth noting. First, the rate certifies primary stationarity only: by Theorem 4.8, CMS corresponds to $\tilde{\mathbf{d}}^*(\boldsymbol{\theta}) = \mathbf{0}$, while the rescaling halts the algorithm whenever $\mathbf{d}^*(\boldsymbol{\theta}) = \tilde{\mathbf{d}}^* \cdot \|\mathbf{g}_1\| / \left\|\tilde{\mathbf{d}}^*\right\| = \mathbf{0}$, which admits both the CMS branch and a primary-stationary non-CMS branch in which $\mathbf{g}_1 = \mathbf{0}$ but $\tilde{\mathbf{d}}^* \neq \mathbf{0}$. The asymmetric/regularizer regime of Eq. (1), in which secondaries are jointly stationary at primary minima, makes the second branch non-generic. Second, the alignment hypothesis $\widetilde{\alpha}_t \geq \mu_0$ holds with $\mu_0 = 1$ in the conflict-free regime but can fail in severe conflict, so the result is conditional rather than unconditional. Strengthening to a single-limit-point statement via a Kurdyka–Łojasiewicz argument is non-trivial: the natural Lyapunov candidate $\Phi = \sum_i \mathcal{L}_i$ has critical points at weighted-sum stationarity rather than at CMS, so a merit function whose critical points coincide with CMS would need to be constructed first. We leave this to future work.

D Feasibility at $K > 2$

The theory of Sec. 4 applies uniformly for all $K \geq 2$. The one phenomenon specific to $K > 2$ is that the feasible set $F(\tau)$ can become empty when the secondary gradients are mutually incompatible. The Farkas

characterization of Theorem B.2 gives the tight condition; the following angle-based sufficient bound is useful when only pairwise cosines are available.

Proposition D.1 (Sufficient condition for feasibility)**.** *Let* $\theta_{\max} = \max_{j \neq k \in [K]_{-1}} \theta_{jk}$ *be the maximum pairwise angle among the (nonzero, normalized) secondary gradients, with the convention* $\arccos(1/0) = 0$ *for* $K = 3$*. If*

$$\theta_{\max} \;<\; \pi - \arccos\!\left(\tfrac{1}{K-2}\right),$$

then $\mathbf{0} \notin \operatorname{cone}\{\tilde{\mathbf{g}}_2, \ldots, \tilde{\mathbf{g}}_K\}$*, hence* $F(\tau) \neq \varnothing$ *for all* $\tau \in [0, 1]$*.*

Proof. Suppose for contradiction $\mathbf{0} \in \operatorname{cone}\{\tilde{\mathbf{g}}_j\}_{j \geq 2}$ with each $\tilde{\mathbf{g}}_j$ a unit vector. Then $\mathbf{0} = \sum_{j=2}^{K} \alpha_j \tilde{\mathbf{g}}_j$ for some $\alpha_j \geq 0$, not all zero. Set $S = \sum_j \alpha_j > 0$. Computing $\left\|\sum_j \alpha_j \tilde{\mathbf{g}}_j\right\|^2 = 0$ and expanding,

$$\sum_j \alpha_j^2 \;=\; -2 \sum_{j<k} \alpha_j \alpha_k \cos\theta_{jk}.$$

By hypothesis $\cos\theta_{jk} > -1/(K-2)$ for all $j < k$, so $-\cos\theta_{jk} < 1/(K-2)$ and

$$\sum_j \alpha_j^2 \;<\; \tfrac{2}{K-2} \sum_{j<k} \alpha_j \alpha_k \;=\; \tfrac{1}{K-2}\Big(S^2 - \sum_j \alpha_j^2\Big),$$

i.e. $(K-1)\sum_j \alpha_j^2 < S^2$. But Cauchy–Schwarz on $S = \sum_j 1 \cdot \alpha_j$ gives $S^2 \leq (K-1)\sum_j \alpha_j^2$ — contradiction. Feasibility then follows from the Farkas characterization (Theorem B.2). □

For $K = 3$, the bound reduces to $\theta_{\max} < \pi$: feasibility holds unless the two secondaries are exactly anti-parallel, recovering Cor. B.3(ii). For $K \geq 4$, the bound is tight at the equilateral configuration. The condition is sufficient but not necessary — for instance, with $K = 4$ and three planar secondaries at angles $0°, 100°, 170°$, $\theta_{\max} = 170° > 120°$ but all three lie in a common open half-plane, so $\mathbf{0} \notin$ cone. The tight characterization remains the Farkas alternative.

E Expanded Results: Structured Pruning

We report full per-configuration results omitted from the main body, complementing Tables 1 and 2. Table 4 reports test accuracy at every sparsity level for all six architecture/dataset configurations and all baselines. Because the pruning target uniquely determines deployment cost, Tab. 5 reports the unpruned baselines together with the achieved efficiency at the 80% and 99% operating points; intermediate points are visualized in the Pareto frontiers (Fig. 13), and τ-sweeps appear in Fig. 14.

E.1 Ablation: Gradient Normalization in PCD

We ablate the normalization scheme applied to the gradient pair $(\mathbf{g}_1, \mathbf{g}_2)$ before the PCD quadratic program. Five conditions are compared: no normalization ($s_i = 1$), instantaneous unit-norm normalization ($s_i = 1/\|g_i\|$), and EMA normalization with $\beta \in \{0.9, 0.99, 0.999\}$ (effective gradient-step memory windows of approximately 10, 100, and 1,000 steps, respectively). All conditions use DenseNet-121 on CIFAR-10, $\tau = 0.01$, $\lambda = 10^{-3}$, and the same $\|\mathbf{g}_1\|$ output rescaling throughout.

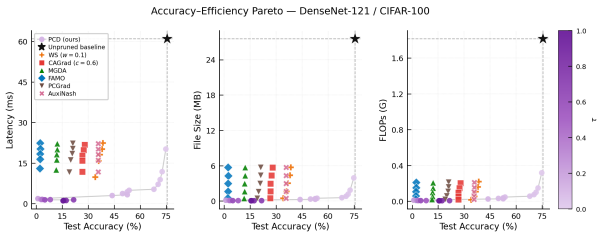
We observe that no normalization produces the worst outcome. The raw gradient magnitudes are severely mismatched ($\|\mathbf{g}_2\| \approx 100.9$ vs. $\|\mathbf{g}_1\| \approx 15.2$), causing the constraint threshold $\tau\|\mathbf{g}_2\|^2 \approx 101.8$ to dwarf the inner product $\mathbf{g}_1^\top \mathbf{g}_2 \approx 0$, so the constraint slack is ≈ -101.8 throughout yet the actual correction to the descent direction is negligible. Standard normalization operates as expected: the normalized constraint slack is exactly $-\tau$ at every epoch, and CE efficiency is 1.000 throughout, confirming the constraint costs essentially nothing in CE alignment. However, group sparsity in this case drops in comparison with EMA-normalized alternatives. EMA normalization achieves the best accuracy across all three β values (94.05–94.24%) while matching standard normalization on sparsity. EMA keeps the scale ratio $s_{\mathrm{CE}}/s_{\mathrm{GL}}$ moderate throughout, so τ retains its intended meaning as a fractional progress requirement. The three EMA variants are nearly indistinguishable (final metrics differ by $< 0.2\%$ in accuracy, $< 0.5\%$ in sparsity), suggesting that window size is not a critical factor at τ=0.01. We adopt β=0.999 as the default: it is the most conservative choice and matches the established convention in adaptive gradient methods.

Configuration	Target	PCD (Ours)	WS	CAGrad	MGDA	PCGrad	FAMO	AuxiNash
DenseNet-121 CIFAR-100	Unpruned	**75.2**	**75.2**	**75.2**	**75.2**	**75.2**	**75.2**	**75.2**
	80%	**74.8**	38.5	28.0	12.1	21.1	2.3	35.8
	85%	**74.8**	38.0	27.4	12.1	21.0	2.4	35.8
	90%	**72.7**	36.8	26.7	11.8	20.7	2.3	35.7
	95%	**71.6**	36.0	26.7	11.7	19.9	2.2	35.7
	99%	**46.5**	34.1	26.7	11.5	19.0	2.1	35.7
ResNet-34 CIFAR-10	Unpruned	**94.4**	**94.4**	**94.4**	**94.4**	**94.4**	**94.4**	**94.4**
	80%	**93.0**	70.8	61.6	10.3	51.6	23.8	73.7
	85%	**93.0**	70.8	61.6	10.3	51.6	23.8	73.7
	90%	**93.0**	70.8	61.6	10.3	51.6	23.8	73.7
	95%	**92.0**	70.8	61.6	10.3	51.6	23.8	73.7
	99%	**75.8**	63.5	53.8	10.2	51.2	23.8	61.3
Inception CIFAR-10	Unpruned	**91.0**	**91.0**	**91.0**	**91.0**	**91.0**	**91.0**	**91.0**
	80%	**89.6**	64.2	33.8	19.1	32.3	10.0	56.3
	85%	**89.6**	64.2	33.8	18.9	32.3	10.0	56.3
	90%	**89.6**	64.2	33.8	18.5	32.3	10.0	56.3
	95%	**86.2**	64.2	33.8	18.5	32.3	10.0	56.3
	99%	**78.4**	64.2	33.8	18.2	32.3	10.0	56.3
Inception CIFAR-100	Unpruned	**68.5**	**68.5**	**68.5**	**68.5**	**68.5**	**68.5**	**68.5**
	80%	**68.5**	14.0	8.5	2.4	8.9	1.0	20.4
	85%	**61.4**	14.0	8.5	2.4	8.9	1.0	20.5
	90%	**61.6**	14.0	8.5	2.4	8.9	1.0	20.5
	95%	**57.2**	14.0	8.5	2.4	8.9	1.0	20.5
	99%	**28.6**	13.4	8.5	2.4	9.4	1.0	17.3
MobileNetV2 CIFAR-10	Unpruned	**93.1**	**93.1**	**93.1**	**93.1**	**93.1**	**93.1**	**93.1**
	80%	**93.0**	41.7	26.2	11.7	25.7	10.0	60.2
	85%	**93.0**	41.6	26.2	11.7	25.6	10.0	60.2
	90%	**92.2**	41.6	29.2	11.7	25.6	10.0	60.0
	95%	**88.2**	41.6	26.1	12.7	5.5	10.0	60.0
	99%	**18.3**	16.7	16.7	12.7	11.4	10.0	10.9
MobileNetV2 CIFAR-100	Unpruned	**72.0**	**72.0**	**72.0**	**72.0**	**72.0**	**72.0**	**72.0**
	80%	**68.0**	15.3	6.2	1.2	4.6	1.0	17.0
	85%	**64.8**	15.3	6.2	1.2	4.5	1.0	16.9
	90%	**55.3**	15.2	6.2	1.2	4.5	1.0	16.9
	95%	**24.6**	15.2	6.3	1.2	4.5	1.0	16.9
	99%	1.2	**1.9**	1.2	1.1	1.1	1.0	1.0

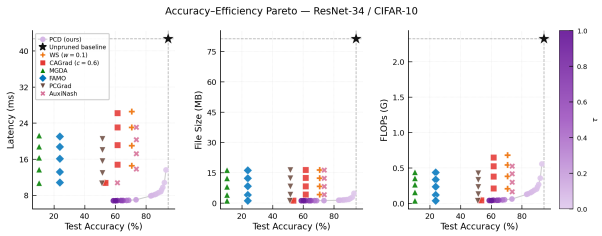
Table 4: Test accuracy (%) on the structured-pruning task across all six architecture/dataset configurations, all baselines, and all sparsity targets. Means over 5 seeds, best values bolded. AuxiNash shares the same unpruned base model as the other methods, so its Unpruned cell reports the shared base accuracy. Deployment costs at the same target sparsity are method-independent and reported separately in Tab. 5.

Configuration	Unpruned baseline			@ 80% sparsity			@ 99% sparsity		
	FLOPs (G)	Lat. (ms)	Size (MB)	FLOPs (G)	Lat. (ms)	Size (MB)	FLOPs (G)	Lat. (ms)	Size (MB)
DenseNet-121 / C-100	1.82	61.0	27.6	0.26 (7.0×)	26.5 (2.3×)	6.13 (4.5×)	0.013 (140.9×)	8.5 (7.2×)	0.47 (58.6×)
ResNet-34 / C-10	2.33	42.7	81.3	0.67 (3.5×)	25.1 (1.7×)	16.59 (4.9×)	0.067 (34.6×)	11.0 (3.9×)	1.34 (60.8×)
Inception / C-10	0.53	13.0	5.4	0.063 (8.4×)	5.9 (2.2×)	1.06 (5.1×)	0.006 (88.6×)	3.5 (3.7×)	0.10 (53.9×)
Inception / C-100	0.53	13.0	5.6	0.082 (6.5×)	6.8 (1.9×)	1.08 (5.2×)	0.002 (265.9×)	2.9 (4.5×)	0.089 (62.8×)
MobileNetV2 / C-10	0.19	9.3	8.8	0.083 (2.3×)	6.6 (1.4×)	1.54 (5.7×)	0.005 (37.8×)	3.1 (3.0×)	0.18 (48.9×)
MobileNetV2 / C-100	0.19	9.3	9.2	0.050 (3.8×)	4.9 (1.9×)	1.61 (5.7×)	0.005 (37.9×)	3.0 (3.1×)	0.24 (38.4×)

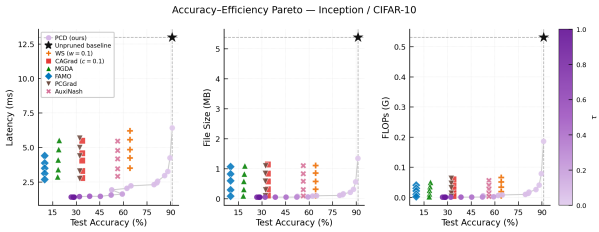
Table 5: Deployment efficiency at the boundary operating points. "Unpruned baseline" lists raw FLOPs, latency, and file size. The 80% and 99% sparsity columns list the absolute value at that operating point with the reduction factor relative to the baseline in parentheses. Identical for any method achieving the same target sparsity. Intermediate 85–95% points are visualized in Fig. 13.



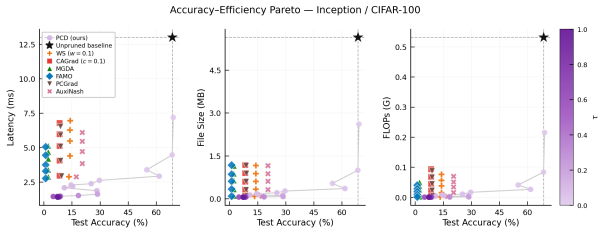
(a) DenseNet-121 / CIFAR-100



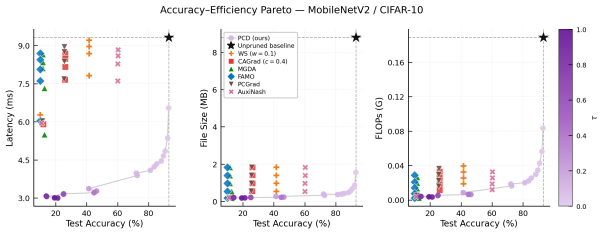
(b) ResNet-34 / CIFAR-10



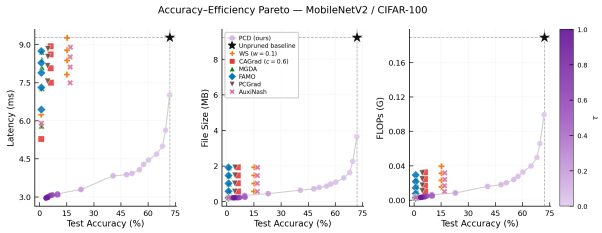
(c) Inception / CIFAR-10



(d) Inception / CIFAR-100

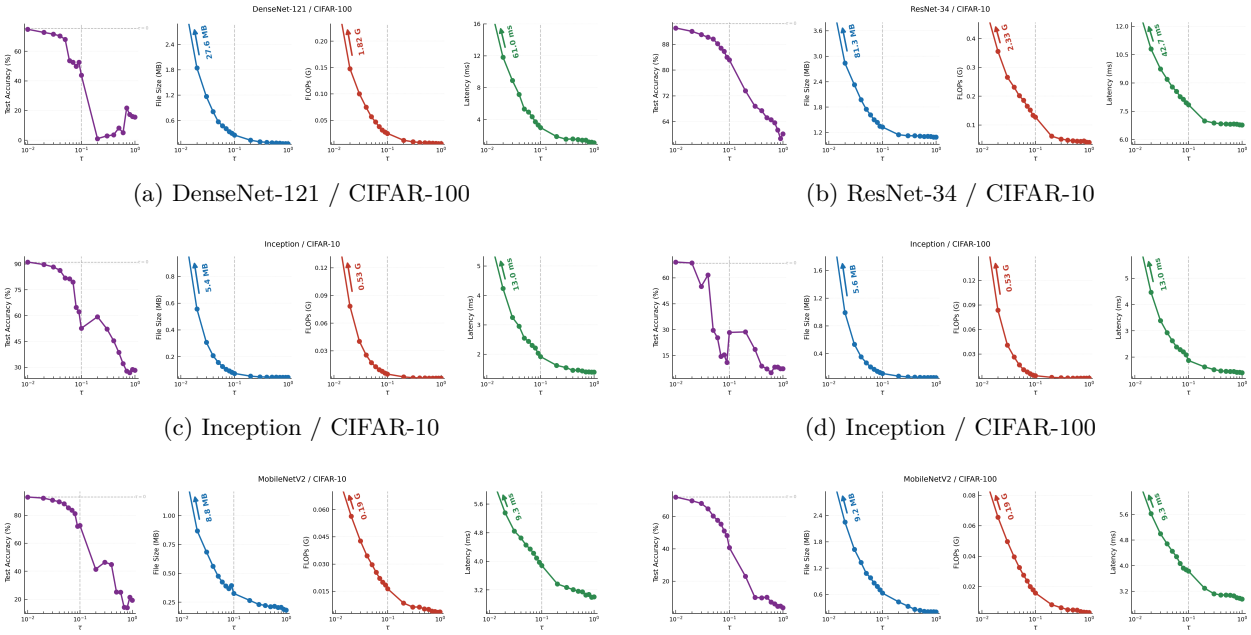


(e) MobileNetV2 / CIFAR-10



(f) MobileNetV2 / CIFAR-100

Figure 13: Accuracy–efficiency Pareto frontiers for PCD across all six configurations, plotted against inference latency (ms), file size (MB), and FLOPs (G). Unpruned baselines marked with $\star$.



(a) DenseNet-121 / CIFAR-100 (b) ResNet-34 / CIFAR-10

(c) Inception / CIFAR-10 (d) Inception / CIFAR-100

(e) MobileNetV2 / CIFAR-10 (f) MobileNetV2 / CIFAR-100

Figure 14: Accuracy, file size (MB), FLOPs (G), and latency (ms) as a function of the priority tolerance τ for all six configurations not included in the main body of the paper. Consistently, the majority of compression benefit is realized for $\tau \in (0, 0.1]$. Arrows pointing up represent efficiency metric values at $\tau = 0$ (unpruned baseline).

Table 6: Normalization ablation results (DenseNet-121, CIFAR-10, τ=0.01, λ=10^{-3}, seed 42). Direction deviation is $\|d^* - g_1\|/\|g_1\|$; CE efficiency is $\cos(d^*, \tilde{g}_1)$, both epoch-averaged.

Normalization	**Acc. (%)**	**Group Spar. (%)**	**Dir. Dev.**	**CE Eff.**
None	83.25	98.14	0.082	0.996
Standard (unit norm)	93.89	79.94	0.010	1.000
EMA β=0.9	94.24	81.29	0.044	0.993
EMA β=0.99	94.05	81.34	0.048	0.992
EMA β=0.999	94.08	81.72	0.051	0.991