KAGRAD: KL-AWARE GRADIENT ALIGNMENT IN MULTITASK LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In gradient-based multitask learning with shared parameters, an update that decreases one task loss can increase another. Existing methods mitigate such conflicts using the Euclidean geometry of task gradients. For probabilistic models, however, gradient geometry alone does not determine local changes in output distributions. Tasks with identical gradients can have different Fisher information matrices and therefore assign different local KL costs to the same update. We propose *KL-Aware Gradient Alignment* (KAGrad), which incorporates this task-specific geometry into the choice of a shared update. Each task gradient and a supplied positive-definite metric define a quadratic model of predicted improvement; with the exact Fisher, its quadratic penalty is proportional to the local KL term. KAGrad maximizes the smallest fraction of each task's independently attainable model improvement, within a Euclidean ball centered at the average-loss gradient. Equivalently, it minimizes the largest normalized squared distance, in each task's supplied metric, from the update preferred by that task's local model. The optimized worst-task fraction is nonnegative if and only if the feasible set contains an update under which every local model used in the optimization predicts non-increase. When the selected direction is applied directly as a gradient step, the constraint gives average-loss descent and an $O(1/T)$ stationarity bound under smoothness, a lower-bounded average loss, and a step-size bound. Because a Fisher metric need not match task-loss curvature, we also derive a sufficient condition under which the local predictions hold for actual losses after a finite step. Against CAGrad, which uses the same average-gradient constraint with a first-order criterion, KAGrad with estimated Fisher or identity metrics attains higher Cityscapes mean intersection over union (74.79 and 75.71 versus 74.16) at a shared constraint radius, but higher relative depth error (39.40 and 37.82 versus 33.70). On MT50, eight-term grouped KAGrad attains higher peak mean task success than eight-term CAGrad-Fast (0.533 versus 0.500) over checkpoints shared across runs. We separately assess the supplied metric. In quadratic tasks, matching it to loss curvature improves worst-task actual loss decrease over an identity metric; rotating the Hessians away diminishes and reverses this advantage. In learned models, diagonal Fisher estimates show no consistent benefit over identity metrics: MT10 peak mean success is 0.700 versus 0.833, and ten paired Multi-Fashion+MNIST runs detect no difference. Cityscapes shows a segmentation–depth trade-off. The benefit of a curvature-matched metric in quadratic tasks does not carry over to diagonal Fisher estimates in these learned models.

1 INTRODUCTION

Multitask learning uses shared parameters to learn from related tasks, but an update to those parameters can decrease one task loss while increasing another. Methods for resolving this conflict typically compare task-loss gradients. MGDA selects a minimum-norm convex combination of the gradients to obtain a common first-order descent direction when one exists (Sener and Koltun, 2018). PCGrad removes the component of a task gradient along another when their inner product is negative (Yu et al., 2020a). CAGrad chooses an update near the average-loss gradient that maximizes the smallest first-order predicted decrease across tasks (Liu et al., 2021a). In each case, a task gradient's inner product with a proposed update describes that task's predicted loss decrease to first order.

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

For probabilistic models, the update also changes each task's output distribution. The Fisher information matrix defines the quadratic term in the second-order local expansion of Kullback–Leibler (KL) divergence (Amari, 1998). This geometry has an optimization interpretation: under a fixed quadratic local KL budget, an individual task's natural-gradient direction gives its greatest first-order predicted loss decrease. Tasks with identical loss gradients can nevertheless have different Fisher matrices. Their gradients then assign the same first-order decrease to an update, while their local KL models assign different costs. Natural-gradient and KL trust-region methods use Fisher geometry for a single objective (Amari, 1998; Schulman et al., 2015); elastic weight consolidation penalizes changes to earlier tasks (Kirkpatrick et al., 2017), while CoMOGA permits a common Fisher metric in constrained multi-objective RL (Kim et al., 2025). We instead use a separate task metric to select one shared update across simultaneous tasks, with a Euclidean average-gradient constraint.

We propose *KL-Aware Gradient Alignment* (KAGrad). For each task, its loss gradient and a supplied positive-definite metric define a quadratic model of predicted improvement: first-order loss decrease minus a quadratic penalty. When the metric is the exact Fisher, this penalty is proportional to the second-order local KL term. The *reference update* for a task maximizes its model improvement when that task is considered alone. A shared update generally cannot attain all task-specific maxima, so KAGrad measures the improvement it predicts for each task as a *retained-improvement fraction* of that task's maximum. KAGrad selects the direction that maximizes the smallest fraction. Equivalently, it minimizes the largest squared distance to a reference update, measured in the corresponding task metric and normalized by the reference update's squared metric norm. This measures how far the shared update departs from what each task would choose alone, relative to the scale of that task's attainable improvement. With the exact Fisher, the distance has a local KL-quadratic interpretation.

An unconstrained maximizer of the worst-task fraction could increase the average task loss. We therefore restrict the direction to lie within Euclidean distance $c\|\bar{g}\|_2$ of the equally weighted average-loss gradient $\bar{g}$, where $0 \leq c < 1$. CAGrad uses the same feasible set (Liu et al., 2021a). When $\bar{g} \neq 0$, every feasible direction has positive inner product with $\bar{g}$ and a norm bounded by a multiple of $\|\bar{g}\|_2$. Consequently, when these directions are applied directly as gradient steps, smoothness and a step-size bound give average-loss descent. If the average loss is also lower bounded, the smallest squared norm of its gradient among the first T iterates is $O(1/T)$.

The worst-task fraction gives a separate guarantee about the local models. Its optimized value is nonnegative if and only if the feasible set contains an update under which every task's model predicts non-increase; KAGrad selects such an update whenever one exists. A positive value means that every model predicts an improvement of at least that fraction of the task's independently attainable maximum. These are model predictions. A Fisher metric describes local KL change, however, and need not match task-loss curvature. We derive a sufficient metric–curvature condition under which the predicted non-increase also holds for actual task losses after a finite step applied directly. For quadratic tasks whose Hessians equal the supplied metrics, the models are exact at unit step size. KAGrad then maximizes the worst-task actual loss decrease relative to each task's independent maximum among updates in the average-gradient feasible set.

For many tasks, we retain selected tasks individually and aggregate the rest. This reduces minimax terms, preserves the average-gradient center, and confines the certificate to represented terms, including the aggregate.

Quadratic tasks isolate metric–curvature alignment. With identical gradients and Hessian-matched metrics, KAGrad improves worst-task actual loss decrease over its identity-metric version. Holding gradients and Hessian spectra fixed while rotating the Hessians diminishes and eventually reverses this advantage.

In learned models, we first compare KAGrad with CAGrad under the same average-gradient constraint. On Cityscapes, both tasks are represented individually; on MT50, the eight-term configurations also use different singleton-selection rules. With all Cityscapes methods at $c = 0.9$, Fisher and Identity KAGrad attain mean intersection over union of 74.79 and 75.71 versus 74.16 for CAGrad. Their relative depth errors are 39.40 and 37.82 versus 33.70, where lower is better. On Meta-World MT50, eight-term grouped KAGrad reaches peak mean task success of 0.533 versus 0.500 for an eight-term CAGrad-Fast baseline over the common evaluated checkpoints. Thus, KAGrad improves the reported Cityscapes segmentation measure and MT50 peak at these settings, while its Cityscapes segmentation gain comes with greater relative depth error.

Within KAGrad, Identity has higher Cityscapes mean intersection over union than Fisher (75.71 versus 74.79). Ten paired Multi-Fashion+MNIST runs detect no difference in final test average loss (0.5803 versus 0.5783). MT10 peak mean success is 0.833 for Identity versus 0.700 for Fisher. Diagonal Fisher estimates show no consistent advantage here.

To interpret the metric comparison, we measure how much sampled loss curvature the diagonal estimates can represent: their mean Frobenius cosine with the loss Hessian is 0.1524 on Multi-Fashion+MNIST, and less than one tenth of sampled Meta-World actor-loss Hessian Frobenius energy lies on the diagonal. In these experiments, Adam applies KAGrad's selected direction; the finite-step and average-loss guarantees concern that direction, not Adam's step.

Our contributions are:

(i) We maximize the smallest retained-improvement fraction, equivalently minimizing the largest normalized metric discrepancy from task reference updates.

(ii) We characterize exactly when every represented local model can predict non-increase and give a sufficient condition transferring that prediction to finite-step task losses.

(iii) In quadratic tasks, matching metrics to loss Hessians improves worst-task actual decrease over identity metrics; rotation diminishes and reverses the advantage. Learned-model tests assess diagonal Fisher estimates and their coverage of sampled loss curvature.

(iv) A grouped representation uses discrepancy screening to reduce minimax terms while preserving the average-gradient center; its certificate covers represented terms.

2 PRELIMINARIES

We first describe loss change locally, then use the Fisher information matrix to describe local KL change. A quadratic KL budget then identifies the natural-gradient direction for an individual task.

Task-loss change. Let $\theta \in \mathbb{R}^p$ be shared parameters and $L_i : \mathbb{R}^p \to \mathbb{R}$, $i \in [K] = \{1, \dots, K\}$, be task losses. Write $g_i = \nabla L_i(\theta)$. The equally weighted average loss and its gradient are

$$\bar{L}(\theta) = \frac{1}{K}\sum_{i=1}^{K} L_i(\theta), \qquad \bar{g} = \nabla \bar{L}(\theta) = \frac{1}{K}\sum_{i=1}^{K} g_i. \tag{1}$$

For a fixed direction d, step size $\alpha > 0$, and update $\theta^+ = \theta - \alpha d$, if L_i is twice differentiable near θ with $H_i = \nabla^2 L_i(\theta)$, then

$$L_i(\theta - \alpha d) - L_i(\theta) = -\alpha g_i^\top d + \frac{\alpha^2}{2} d^\top H_i d + o(\alpha^2) \qquad \text{as } \alpha \to 0. \tag{2}$$

Thus $g_i^\top d > 0$ predicts a first-order decrease in L_i, while $\bar{g}^\top d > 0$ predicts one in $\bar{L}$; the latter does not imply the former for every i. Neither gradient condition describes how the update changes a task's output distribution.

Output-distribution change. To describe that change, consider a conditional output distribution $p_i(y \mid x; \theta)$. Let μ_i be a fixed distribution over inputs and define the score $s_i(x, y; \theta) = \nabla_\theta \log p_i(y \mid x; \theta)$. Its input-averaged Fisher information matrix is

$$F_i^{\mathrm{KL}} = \mathbb{E}_{x\sim\mu_i}\mathbb{E}_{y\sim p_i(\cdot|x;\theta)}\left[s_i(x, y; \theta) s_i(x, y; \theta)^\top\right]. \tag{3}$$

Under standard regularity conditions, the input-averaged Kullback–Leibler (KL) divergence satisfies

$$\mathbb{E}_{x\sim\mu_i} D_{\mathrm{KL}}(p_i(\cdot \mid x; \theta) \,\|\, p_i(\cdot \mid x; \theta + \delta)) = \frac{1}{2}\delta^\top F_i^{\mathrm{KL}} \delta + o(\|\delta\|_2^2). \tag{4}$$

For the update $\theta^+ = \theta - \alpha d$, set $\delta = -\alpha d$. The quadratic KL term is then $\alpha^2 d^\top F_i^{\mathrm{KL}} d/2$.

Natural-gradient direction. For a single task, maximize first-order predicted loss decrease subject to a budget on the quadratic KL term. Write $u = \alpha d$, so $\theta^+ = \theta - u$ and the quadratic KL term in Eq. (4) is $\frac{1}{2}u^\top F_i^{\mathrm{KL}} u$. If $F_i^{\mathrm{KL}} \succ 0$, $g_i \neq 0$, and $\varepsilon > 0$, then

$$\underset{u:\, \frac{1}{2}u^\top F_i^{\mathrm{KL}} u \leq \varepsilon}{\arg\max}\; g_i^\top u = \sqrt{\frac{2\varepsilon}{g_i^\top (F_i^{\mathrm{KL}})^{-1} g_i}}\, (F_i^{\mathrm{KL}})^{-1} g_i. \tag{5}$$

The solution is the natural-gradient direction (Amari, 1998), scaled by the budget ε; it depends on both the loss gradient and the Fisher matrix. The Fisher controls the local quadratic KL term, whereas the loss Hessian H_i controls second-order loss change in Eq. (2). Because these matrices need not coincide, the quadratic KL budget alone does not guarantee a decrease in actual task loss after a finite step.

3 KL-Aware Gradient Alignment

Equation (5) gives the direction of greatest first-order predicted loss decrease for one task under a quadratic KL budget. For shared parameters, a direction must account for all tasks. *KL-Aware Gradient Alignment* (KAGrad) maximizes the smallest fraction of each task's independently attainable model improvement. It constrains the direction near the average-loss gradient to retain an average-loss descent guarantee when the direction is applied directly as a gradient step.

Local improvement and reference updates. For each task, supply a positive-definite *supplied metric* $F_i \succ 0$ and define the *local improvement model*

$$I_i(d) = \alpha \left(g_i^\top d - \frac{1}{2} d^\top F_i d \right), \qquad \alpha > 0. \tag{6}$$

The first term is the first-order loss decrease under $\theta^+ = \theta - \alpha d$. When $F_i = F_i^{\mathrm{KL}}$, the penalty is $1/\alpha$ times the quadratic local KL term in Eq. (4). The learned-model experiments use damped diagonal Fisher estimates as supplied metrics.

The KL expansion determines its quadratic form but does not set its coefficient relative to predicted loss decrease. Changing that coefficient amounts to rescaling F_i and can change the selected direction. With the coefficient in Eq. (6), maximizing I_i alone gives

$$d_i = F_i^{-1} g_i, \qquad I_i^{\max} = I_i(d_i) = \frac{\alpha}{2} g_i^\top F_i^{-1} g_i. \tag{7}$$

We call d_i task i's *reference update*. For the exact Fisher, d_i lies along the natural-gradient direction in Eq. (5); Eq. (6) sets its magnitude in place of the budget ε.

Assume $g_i \neq 0$ for every $i \in [K]$, so $I_i^{\max} > 0$. The *retained-improvement fraction* is

$$r_i(d) = \frac{I_i(d)}{I_i^{\max}}, \qquad r_{\min}(d) = \min_{i \in [K]} r_i(d). \tag{8}$$

The common factor α cancels. The fraction equals one only at d_i, equals zero when the model predicts no change, and is negative when it predicts an increase. These are predictions of Eq. (6), not claims about finite-step losses. In particular,

$$r_{\min}(d) \geq 0 \quad \Longleftrightarrow \quad I_i(d) \geq 0 \quad \text{for every } i \in [K]. \tag{9}$$

Appendix B.4 describes how the implementation excludes numerically negligible gradients from this criterion without changing $\bar{g}$.

Shared-update constraint and criterion. To control progress on the average loss, we use the feasible set introduced by CAGrad (Liu et al., 2021a):

$$\mathcal{A}_c(\bar{g}) = \{ d \in \mathbb{R}^p : \|d - \bar{g}\|_2 \leq c \|\bar{g}\|_2 \}, \qquad 0 \leq c < 1. \tag{10}$$

Every feasible direction satisfies

$$\bar{g}^\top d \geq (1 - c)\|\bar{g}\|_2^2, \qquad \|d\|_2 \leq (1 + c)\|\bar{g}\|_2. \tag{11}$$

When $\bar{g} \neq 0$, the first bound makes d a descent direction for the average loss; when $\bar{g} = 0$, the feasible set contains only $d = 0$. Section 4 uses both bounds to establish average-loss descent for directly applied steps under its smoothness and step-size conditions. These bounds do not depend on the supplied metrics.

KAGrad changes the criterion optimized within the feasible set:

$$\boxed{d^\star = \arg\max_{d \in \mathcal{A}_c(\bar{g})} r_{\min}(d)}, \qquad r^\star_{\min} = r_{\min}(d^\star). \tag{12}$$

For every task, $I_i(d^\star) \geq r^\star_{\min} I_i^{\max}$. Because $d^\star$ maximizes $r_{\min}$ over the feasible set, $r^\star_{\min} \geq 0$ exactly when the feasible set contains a direction under which every task's local model predicts non-increase.

Equivalent normalized discrepancy. Completing the square gives

$$I_i^{\max} - I_i(d) = \frac{\alpha}{2}(d - d_i)^\top F_i (d - d_i). \tag{13}$$

The shortfall from task i's independent maximum is proportional to its squared distance from d_i in the supplied metric. Normalizing this shortfall by $I_i^{\max}$ defines the *normalized metric discrepancy*

$$\rho_i(d) = \frac{(d - d_i)^\top F_i (d - d_i)}{g_i^\top F_i^{-1} g_i}, \tag{14}$$

for which

$$r_i(d) = 1 - \rho_i(d). \tag{15}$$

Writing $\rho_{\max}(d) = \max_{i \in [K]} \rho_i(d)$, KAGrad equivalently solves

$$d^\star = \arg\min_{d \in \mathcal{A}_c(\bar{g})} \rho_{\max}(d), \qquad r^\star_{\min} = 1 - \rho_{\max}(d^\star). \tag{16}$$

Because $F_i \succ 0$ and $g_i \neq 0$, each ρ_i is strongly convex. Their finite maximum therefore has a unique minimizer over the convex feasible set. Consequently, no other feasible direction can weakly increase every retained fraction while strictly increasing one. The model-prediction certificate also takes the form

$$r^\star_{\min} \geq 0 \quad \iff \quad \rho_{\max}(d^\star) \leq 1. \tag{17}$$

For $F_i = F_i^{\mathrm{KL}}$, multiplying the numerator and denominator of Eq. (14) by $\alpha^2/2$ gives a ratio of quadratic local KL terms. Define $\widehat{D}_i(v) = \frac{\alpha^2}{2} v^\top F_i^{\mathrm{KL}} v$, the term in Eq. (4) at $\delta = -\alpha v$. Then

$$\rho_i(d) = \frac{\widehat{D}_i(d - d_i)}{\widehat{D}_i(d_i)}. \tag{18}$$

This normalized metric discrepancy measures departure from d_i; it is not the KL change caused by applying d itself. For an estimated or damped metric, Eq. (14) retains its supplied-metric interpretation. Appendix B gives the reduced dual and numerical algorithm.

The two-task calculation in Appendix B.5 shows that identical gradients can have different reference updates and that KAGrad selects $d^\star = (1 - c)\bar{g}$ for $c < 0.6$, changing the magnitude of the shared direction.

3.1 Grouped representation for many tasks

Representing every task separately gives one term per task in Eq. (16). To use M terms, let S contain $M - 1$ individually represented tasks and let $R = [K] \setminus S$ contain the rest. Define

$$L_R(\theta) = \frac{1}{|R|} \sum_{i \in R} L_i(\theta), \qquad g_R = \nabla L_R(\theta) = \frac{1}{|R|} \sum_{i \in R} g_i, \tag{19}$$

and supply a positive-definite metric F_R for this aggregate. Weighting the aggregate by its number of tasks preserves the average-loss gradient:

$$\frac{1}{K} \sum_{i \in S} g_i + \frac{|R|}{K} g_R = \bar{g}. \tag{20}$$

We optimize the same minimax criterion over the M represented terms while retaining the feasible set centered at $\bar{g}$. The nonnegative-optimum certificate then concerns each singleton in S and the aggregate R, not each task inside R.

For MT50, $M = 8$: the six largest estimated normalized metric discrepancies at $\bar{g}$ select six singletons; a seventh follows a deterministic coverage schedule independent of the screen, and the rest form the aggregate. Grouping reduces minimax terms, but the estimator still evaluates scores for every task. We write $\mathrm{P}q$ for a diagonal metric estimate from q score probes. Appendix C.5 defines the screening score and its score-gradient count.

4 Task-Loss and Average-Loss Guarantees

KAGrad uses local models to choose a direction and an average-gradient constraint to limit that choice. We examine what each part guarantees: when the local models' predictions hold for actual task losses, and when a directly applied update decreases the average loss.

Task-loss change. Since $I_i^{\max} > 0$, $r_i(d) \geq 0$ means that task i's local model predicts non-increase. Optimizing the smallest retained fraction determines whether all represented terms can satisfy this threshold within the feasible set.

Proposition 1 (Certificate of predicted non-increase)**.** *The feasible set contains a direction with nonnegative predicted improvement for every represented term if and only if $r^\star_{\min} \geq 0$. When this holds, $I_i(d^\star) \geq 0$ for every represented term.*

Equation (17) gives the equivalent threshold in terms of normalized metric discrepancy. This result concerns the local models. Their predictions equal actual loss changes in the following quadratic case.

Proposition 2 (Matched quadratic case)**.** *If $L_i(\vartheta) = \frac{1}{2}(\vartheta - \theta_i^\star)^\top F_i(\vartheta - \theta_i^\star)$ and $\alpha = 1$, then $L_i(\theta) - L_i(\theta - d) = I_i(d)$ for every d. Hence KAGrad maximizes, over $\mathcal{A}_c(\bar{g})$, the smallest actual task-loss decrease relative to each task's unconstrained maximum.*

For general losses, the supplied metric need not equal loss curvature. To identify when a model prediction still bounds the actual decrease, let $\Delta_i(d) = L_i(\theta) - L_i(\theta - \alpha d)$. If L_i is twice continuously differentiable along the update segment, Taylor's theorem gives

$$\Delta_i(d) = \alpha g_i^\top d - \frac{\alpha^2}{2} d^\top \overline{H}_i(d;\alpha) d, \tag{21}$$

where

$$\overline{H}_i(d;\alpha) = 2\int_0^1 (1-t)\nabla^2 L_i(\theta - t\alpha d)\, dt. \tag{22}$$

Comparing this expression with Eq. (6) gives a sufficient condition: the metric penalty must be at least as large as the curvature contribution along each feasible update.

Condition 1 (Metric–curvature condition)**.** For every represented term and every $d \in \mathcal{A}_c(\bar{g})$,

$$\alpha d^\top \overline{H}_i(d;\alpha) d \leq d^\top F_i d. \tag{23}$$

Under Condition 1,

$$\Delta_i(d) - I_i(d) = \frac{\alpha}{2}\left[d^\top F_i d - \alpha d^\top \overline{H}_i(d;\alpha) d\right] \geq 0. \tag{24}$$

The local model therefore does not overstate the actual decrease, giving the finite-step result.

Theorem 1 (Finite-step task non-increase)**.** *If Condition 1 holds and $r^\star_{\min} \geq 0$, then $L_i(\theta - \alpha d^\star) \leq L_i(\theta)$ for every represented term.*

Condition 1 is sufficient, not necessary: its failure removes the guarantee but does not imply that a task loss increases. With grouping, the proposition and theorem concern the individually represented tasks and the aggregate, as specified in Section 3.1.

Average-loss descent and stationarity. The average-loss result follows from the feasible set rather than the local models. CAGrad proves a stationarity bound for its update direction (Liu et al., 2021a, Theorem 3.2). Its descent argument uses only membership in the feasible set, so it applies to KAGrad's direction as well. Here smoothness is required only for the average loss.

Theorem 2 (Average-loss descent and stationarity)**.** *Assume $\bar{L}$ is Λ-smooth, with $\Lambda > 0$, and bounded below by $\bar{L}^{\inf}$. If updates are applied directly as $\theta_{t+1} = \theta_t - \alpha d_t$, where $d_t \in \mathcal{A}_c(\bar{g}_t)$ and $0 < \alpha \leq 1/\Lambda$, then*

$$\bar{L}(\theta_{t+1}) \leq \bar{L}(\theta_t) - \frac{\alpha(1-c^2)}{2}\|\bar{g}_t\|_2^2, \tag{25}$$

$$\min_{0\leq t<T} \|\bar{g}_t\|_2^2 \leq \frac{2(\bar{L}(\theta_0) - \bar{L}^{\inf})}{\alpha(1-c^2)T}. \tag{26}$$

Because $\bar{g} = K^{-1}\sum_i g_i$, a zero of $\bar{g}$ is first-order Pareto-stationary: no direction strictly decreases every differentiable task loss to first order. The theorem bounds the smallest mean-gradient norm among the first T iterates; it does not require each task gradient to vanish or establish convergence to a particular point. Complete proofs and an exact quadratic test for Condition 1 appear in Appendix A. Appendix Table 3 summarizes the assumptions and conclusions.

5 Quadratic Experiments

A family is one joint draw of K tasks. These problems test identical gradients, metric–curvature mismatch, and grouping.

Identical gradients. In 500 quadratic families, all tasks share a gradient but have different Hessians H_i, and supplied-metric KAGrad uses $F_i = H_i$. Identity metrics give every task the same reference update, so Identity KAGrad selects $d = \bar{g}$ for every c. Scored with the matched models, its mean worst-task retained fraction is -2.644, with all-task non-increase in 7.2% of families; at $c = 0.5$, supplied-metric KAGrad obtains 0.301 and 100%. Appendix C.2 specifies the generator and the $c = 0.2$ comparison.

Metric–curvature mismatch. We next hold the gradients and supplied metrics fixed while rotating the loss Hessians. Across 300 paired families of $K = 50$ tasks in dimension 4, gradients $g_i \propto b + \sigma z_i$ share a unit vector b, with independent unit vectors z_i and dispersion σ. A task-specific rotation $Q_i(\phi)$, with $Q_i(0) = I$, defines $H_i(\phi) = Q_i(\phi)F_iQ_i(\phi)^\top$, which rotates the Hessian relative to the supplied metric while preserving its spectrum and each task gradient. We use $c = 0.5$, $\alpha = 1$, and compare supplied-metric KAGrad, Identity KAGrad, and an Oracle-Hessian reference using $H_i(\phi)$ in the same feasible set. Appendix C.3 gives the generator, rotation grid, and metric construction.

For a selected direction d, let $\Delta_i(d;\phi)$ be the actual one-step decrease of task i and let $\Delta_i^{\max}(\phi) = \frac{1}{2}g_i^\top H_i(\phi)^{-1}g_i$. We evaluate each method by its worst-task relative actual decrease:

$$r_{\min}^{\text{loss}}(d;\phi) = \min_i \frac{\Delta_i(d;\phi)}{\Delta_i^{\max}(\phi)}. \tag{27}$$

Figure 1 divides the mean of this quantity over families by the Oracle-Hessian mean at each rotation. The reported differences use unnormalized, paired family-level values.

At $\sigma = 0.6$ and $0°$, the supplied metric equals the loss Hessian and KAGrad agrees with the Oracle-Hessian reference. Its paired advantage over Identity KAGrad is 0.0635 (95% paired bootstrap interval $[0.0597, 0.0675]$). At $45°$, the difference is -0.0257 ($[-0.0338, -0.0181]$). In these families, the supplied-metric advantage diminishes and reverses as the eigenspaces separate; Appendix C.3 reports intermediate rotations.

At $45°$, rescaling each ray removes the supplied-metric–Identity difference (-0.00003, interval $[-0.0045, 0.0043]$): the as-applied observed reversal comes from magnitude; no ray-level difference is detected (Appendix C.3).

For the model-to-loss transfer, at $\sigma = 1.0$, $296/300$ families are locally certified regardless of α or rotation; at $\alpha = 1$, $15/296$ have an actual increase at $45°$. Across $\alpha \in \{0.25, 0.5, 0.75, 1\}$, none of

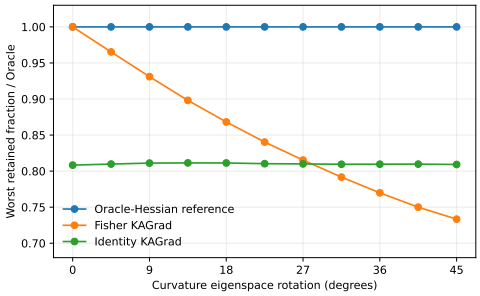

Figure 1: Mean worst-task relative actual decrease divided by the Oracle-Hessian mean at each rotation, for $\sigma = 0.6$. The supplied metric equals the loss Hessian at $0°$; rotation preserves its spectrum. The two KAGrad curves cross between $27°$ and $31.5°$.

the 7,042 cases satisfying both the certificate and Condition 1 violates Theorem 1. The condition is conservative: for $\alpha \leq 0.75$, no violations occur even where it fails; at $\alpha = 1$ it fails at every nonzero rotation (Appendix C.4).

Grouping. Full KAGrad represents each task separately. For grouped KAGrad, we also check $r_i(d) \geq 0$ for every original task, beyond the represented-term certificate. At $K = 50$ and $M = 8$, the all-task check succeeds in 86.0% of families for full KAGrad, 84.0% with singletons selected by exact discrepancy scores, 46.7% with P2 screening ($[36.0\%, 58.0\%]$), and 22.0% with random selection. Appendix C.5 gives the protocol.

6 Learned-Model Experiments

We now evaluate trained models, where metric–curvature alignment is not controlled. Fisher KAGrad supplies damped diagonal task-Fisher estimates; Identity KAGrad supplies identity metrics. The selected direction passes to Adam, so the direct-update guarantees in Section 4 concern the pre-Adam direction.

Cityscapes. We follow CAGrad's two-task setup (Liu et al., 2021a) on Cityscapes (Cordts et al., 2016), jointly training segmentation and depth estimation. All methods use $c = 0.9$; Fisher KAGrad uses P4 predictive-Fisher estimates. We report test metrics averaged over epochs 190–199 of 200 training epochs and three independent seeds; Appendix D.2 gives the protocol.

Table 1: Cityscapes test metrics at $c = 0.9$, averaged over epochs 190–199 within each run and reported as mean $\pm$ standard error across three seeds. Higher is better for segmentation, lower for depth error.

Method	mIoU (%) ↑	Pixel acc. (%) ↑	Abs. err. ↓	Rel. err. ↓
CAGrad	74.16 ± 0.10	93.30 ± 0.06	0.0140 ± 0.0006	33.70 ± 0.25
Identity KAGrad	75.71 ± 0.47	93.69 ± 0.04	0.0136 ± 0.0003	37.82 ± 2.58
Fisher KAGrad	74.79 ± 0.14	93.46 ± 0.03	0.0140 ± 0.0002	39.40 ± 1.97

In Table 1, both KAGrad variants attain higher mean intersection over union and higher relative depth error than CAGrad. Identity KAGrad also exceeds Fisher KAGrad in intersection over union with lower relative depth error.

Multi-Fashion+MNIST. In this two-task classification benchmark, ten paired runs share initialization and minibatch order across CAGrad, Identity KAGrad, and Fisher KAGrad. They train for 50 epochs at $c = 0.2$; Fisher KAGrad uses P8 estimates. Mean final-epoch test average-task loss is 0.5805 for CAGrad, 0.5803 for Identity KAGrad, and 0.5783 for Fisher KAGrad. The paired Identity-minus-Fisher difference is 0.0020 (95% interval $[-0.0184, 0.0225]$, $p = 0.83$); this comparison does not detect a difference between the two metrics. Fisher minus CAGrad is also -0.0022 ($[-0.0192, 0.0148]$). Appendix D.2 reports full statistics and runtime.

Meta-World MT10 and MT50. Meta-World MT10 and MT50 use Soft Actor-Critic (SAC) (Haarnoja et al., 2018; Yu et al., 2020b; Liu et al., 2021a), two million environment steps,

Table 2: Meta-World peak mean task success: the maximum seed-cohort mean over checkpoints common to the methods and seeds in each comparison, as in Eq. (62). Values are mean ± standard error across five seeds.

Benchmark	Method	Peak success ↑
MT10	Identity KAGrad	0.833 ± 0.088
	Fisher KAGrad	0.700 ± 0.058
MT50	CAGrad-Fast-8	0.500 ± 0.012
	Eight-term grouped KAGrad	0.533 ± 0.035

batch size 1280, evaluation every $10{,}000$ steps, and five independent seeds per method. KAGrad selects the actor direction; the critic and entropy-temperature updates follow SAC.

MT10 represents all tasks; Fisher and Identity KAGrad use $c = 0.9$, with P16 policy-Fisher estimates for Fisher KAGrad. MT50 compares eight-term grouped KAGrad with CAGrad-Fast-8, CAGrad's fast variant, at $c = 0.9$. KAGrad picks six singletons by P2 screening and reserves one for coverage, with P4 metrics; CAGrad-Fast-8 takes seven from its task-selection schedule. Each aggregates the remaining tasks. Appendix C.5 gives selection and score-gradient costs.

Table 2 shows that on MT10, Identity KAGrad attains a higher peak than Fisher KAGrad. On MT50, eight-term grouped KAGrad attains a higher peak than CAGrad-Fast-8 at the same constraint radius and representation size. Appendix D.3 reports late-window success and specifies the metric estimator and checkpoint aggregation; Appendix C.5 gives the screening procedure.

Metric diagnostics. On Multi-Fashion+MNIST the mean global Frobenius cosine between the diagonal Fisher estimate and sampled full loss Hessian is 0.1524. For sampled Meta-World actor-loss Hessians, the pooled diagonal-energy fraction ranges from 1.76% to 9.03% when each task is excluded in turn. These measurements show limited global correspondence of the diagonal estimates to sampled loss curvature; they do not test Condition 1. Appendices D.2 and D.4 specify the estimators.

7 Discussion

A Fisher describes local KL change, not necessarily loss curvature. Cityscapes trains L1 depth loss but supplies a Gaussian predictive Fisher associated with squared error; SAC trains an actor objective combining log policy and action value but supplies a policy Fisher. These metrics describe different objectives from the trained losses, so the matched quadratic result does not imply a learned-model gain.

Metric scale changes d_i and the shared solution. We use $\tau_i = 1$; scaling all metrics to $0.1F_i$ rotates the selected ray by 15.9° at measured checkpoints (Appendix D.1). The Euclidean constraint also prevents general reparameterization invariance, even with exact Fishers.

Multi-Fashion+MNIST mean runtime is 1.85 minutes for CAGrad versus 16.52 and 21.40 for Identity and Fisher KAGrad, mainly from repeated dual solves. Grouping reduces minimax terms but not Fisher score-gradient work in this implementation.

8 Conclusion

KAGrad maximizes the smallest retained-improvement fraction within CAGrad's feasible set, equivalently minimizing the largest normalized metric discrepancy. Its nonnegative optimum exactly tests predicted non-increase for represented models. Actual finite-step non-increase needs metric–curvature alignment; direct steps in the feasible set also descend the average loss and bound stationarity.

Matched curvature helps in quadratic tasks, while rotation reverses that advantage. Cityscapes trades segmentation for relative depth error; grouped KAGrad has higher MT50 peak at matched radius and size. The learned comparisons show no consistent Fisher-over-Identity benefit; a Fisher estimate alone does not reproduce the quadratic gain.

Reproducibility Statement

The main text defines the optimization problem, guarantees, quadratic-task results, and benchmark results. The appendices give proofs, the reduced dual and solver verification, quadratic-task generators, metric estimators, grouping diagnostics, and supervised and reinforcement-learning protocols.

AI Use Statement

Generative AI tools were used for research ideation, experimental design and analysis, mathematical derivations and proofs, code generation and debugging, and manuscript drafting and editing. The authors independently reviewed the mathematical claims, experimental code, numerical results, citations, and AI-assisted text and take responsibility for the final submission.

References

Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

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 *ICLR*, 2023.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *ICML*, 2018.

Adrián Javaloy and Isabel Valera. RotoGrad: Gradient homogenization in multitask learning. In *ICLR*, 2022.

Dohyeong Kim, Mineui Hong, Jeongho Park, and Songhwai Oh. Conflict-averse gradient aggregation for constrained multi-objective reinforcement learning. In *ICLR*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *NeurIPS*, 2021.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. FAMO: Fast adaptive multitask optimization. In *NeurIPS*, 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 *ICLR*, 2021.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *ICML*, 2022.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *ICML*, 2015.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *NeurIPS*, 2018.

Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *CVPR*, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *NeurIPS*, 2020.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020.

APPENDIX OVERVIEW

The appendix has four parts. Appendix A proves the task-loss and average-loss results and gives an exact quadratic test for Condition 1. Appendix B derives the numerical problem, works through the two-task example, compares first-order directions, and discusses related methods. Appendix C gives the quadratic-task generators and full results for metric alignment, finite-step losses, and grouping. Appendix D specifies the learned-model protocols and metric diagnostics. Table 3 records the assumptions used by the theoretical statements.

Table 3: Scope of the theoretical statements. Each row lists the assumptions and the corresponding conclusion.

Statement	Requires	Establishes
Eq. (15)	$F_i \succ 0$ and $g_i \neq 0$	$r_i(d) = 1 - \rho_i(d)$ and $r_{\min}(d) = 1 - \rho_{\max}(d)$.
Proposition 1	$F_i \succ 0$, $g_i \neq 0$ for represented terms	$r^\star_{\min} \geq 0$ iff the feasible set contains an update with nonnegative local-model improvement for every represented term.
Proposition 2	Quadratic losses, $H_i = F_i$, and $\alpha = 1$	The local-model improvement equals the actual finite-step loss decrease, and KAGrad maximizes the worst relative loss decrease over the feasible set.
Theorem 1	$r^\star_{\min} \geq 0$ and Condition 1	Non-increase of the loss associated with every represented term after the finite update.
Theorem 2	$\bar{L}$ is Λ-smooth and bounded below; $0 \leq c < 1$; direct updates with $0 < \alpha \leq 1/\Lambda$	Average-loss descent and an $O(1/T)$ bound on $\min_{0 \leq t < T} \|\bar{g}_t\|_2^2$.

A PROOFS AND THE CURVATURE CONDITION

The first proofs connect the normalized criterion to the local-model certificate and the matched quadratic case. The quadratic test then checks Condition 1 over the entire feasible set. The final proof gives the average-loss bound inherited from the feasible-set argument.

A.1 FEASIBLE-SET GEOMETRY

For $d \in \mathcal{A}_c(\bar{g})$, Cauchy–Schwarz gives

$$\bar{g}^\top d = \|\bar{g}\|_2^2 + \bar{g}^\top (d - \bar{g}) \geq \|\bar{g}\|_2^2 - \|\bar{g}\|_2 \|d - \bar{g}\|_2 \geq (1 - c)\|\bar{g}\|_2^2,$$

and the triangle inequality gives

$$\|d\|_2 \leq \|\bar{g}\|_2 + \|d - \bar{g}\|_2 \leq (1 + c)\|\bar{g}\|_2.$$

If $\bar{g} \neq 0$ and $c < 1$, then $0 \notin \mathcal{A}_c(\bar{g})$ because $\|0 - \bar{g}\|_2 = \|\bar{g}\|_2 > c\|\bar{g}\|_2$.

A.2 RETAINED FRACTIONS AND NORMALIZED DISCREPANCY

Using $d_i = F_i^{-1} g_i$,

$$(d - d_i)^\top F_i (d - d_i) = d^\top F_i d - 2 g_i^\top d + g_i^\top F_i^{-1} g_i.$$

Substitution into the definitions of $I_i(d)$ and $I_i^{\max}$ proves Eq. (13). Dividing it by $I_i^{\max} = \frac{\alpha}{2} g_i^\top F_i^{-1} g_i$ gives

$$r_i(d) = 1 - \rho_i(d).$$

For an exact Fisher metric $F_i = F_i^{\mathrm{KL}}$, the numerator of $\rho_i(d)$ is the Fisher quadratic of $d - d_i$ and the denominator is that of d_i. Multiplying both by $\alpha^2/2$ expresses the ratio in the local-KL quadratic units of Eq. (4). For an estimated or damped metric, the same algebra gives a ratio in the supplied quadratic geometry rather than an exact local-KL ratio.

A.3 CERTIFICATE OF PREDICTED NON-INCREASE

For each represented term i, Eq. (15) gives $r_i(d) \geq 0$ if and only if $\rho_i(d) \leq 1$. Hence an update with nonnegative predicted improvement for every represented term exists in $\mathcal{A}_c(\bar{g})$ if and only if

$$\min_{d \in \mathcal{A}_c(\bar{g})} \rho_{\max}(d) \leq 1,$$

which is equivalent to $\rho_{\max}(d^\star) \leq 1$ and therefore to $r^\star_{\min} \geq 0$. If $\rho_{\max}(d^\star) \leq 1$, then $\rho_i(d^\star) \leq 1$ for every represented term, so $I_i(d^\star) \geq 0$ for each of them. If $\rho_{\max}(d^\star) > 1$, every feasible update has negative predicted improvement for at least one represented local model.

A.4 MATCHED QUADRATIC LOSSES

For the quadratic loss in Proposition 2, $g_i = F_i(\theta - \theta_i^\star)$. At $\alpha = 1$,

$$\begin{aligned}\Delta_i(d) &= L_i(\theta) - L_i(\theta - d) \\ &= g_i^\top d - \frac{1}{2} d^\top F_i d \\ &= I_i(d).\end{aligned}$$

The unconstrained quadratic decrease is maximized at $d_i = F_i^{-1} g_i$, giving

$$\Delta_i^{\max} = \frac{1}{2} g_i^\top F_i^{-1} g_i = I_i^{\max}.$$

Thus the normalized actual decrease equals $r_i(d)$ for every d; substituting this identity into Eq. (12) gives the proposition.

A.5 FINITE-STEP TASK NON-INCREASE

By Proposition 1, $r^\star_{\min} \geq 0$ implies

$$I_i(d^\star) \geq 0 \qquad \text{for every represented term } i.$$

Under Condition 1, Eq. (24) gives

$$\Delta_i(d^\star) \geq I_i(d^\star) \geq 0 \qquad \text{for every represented term } i.$$

Since

$$\Delta_i(d^\star) = L_i(\theta) - L_i(\theta - \alpha d^\star),$$

it follows that

$$L_i(\theta - \alpha d^\star) \leq L_i(\theta) \qquad \text{for every represented term } i.$$

A.6 QUADRATIC TEST OF CONDITION 1

Assume $\bar{g} \neq 0$ and $0 < c < 1$, and write $u = \bar{g}/\|\bar{g}\|_2$. Consider a nonzero ray direction x. For $x^\top \bar{g} > 0$, the ray $\{tx : t \geq 0\}$ intersects $\mathcal{A}_c(\bar{g})$ if and only if its minimum Euclidean distance to $\bar{g}$ is at most $c\|\bar{g}\|_2$. This condition is

$$\|\bar{g}\|_2^2 - \frac{(x^\top \bar{g})^2}{\|x\|_2^2} \leq c^2 \|\bar{g}\|_2^2,$$

or equivalently

$$\frac{(x^\top u)^2}{\|x\|_2^2} \geq 1 - c^2 \quad \Longleftrightarrow \quad x^\top C_c x \geq 0, \qquad C_c = uu^\top - (1 - c^2) I.$$

The sign restriction can be dropped when testing Condition 1 because both $x^\top(\alpha H_i - F_i)x$ and $x^\top C_c x$ are even in x. Thus, for a quadratic task, Condition 1 over the feasible set is equivalent to

$$x^\top B_i x \leq 0 \quad \text{whenever} \quad x^\top C_c x \geq 0, \qquad B_i = \alpha H_i - F_i.$$

The quadratic constraint has a strictly feasible direction: $u^\top C_c u = c^2 > 0$. The homogeneous S-lemma therefore gives

$$\text{Condition 1 for task } i \quad \Longleftrightarrow \quad \exists\, \lambda_i \geq 0 \;\text{ such that } B_i + \lambda_i C_c \preceq 0. \tag{28}$$

We test the condition for every task by minimizing $\lambda_{\max}(B_i + \lambda C_c)$ over $\lambda \geq 0$. When $c = 0$ or $\bar{g} = 0$, the feasible set contains only $\bar{g}$ or 0, respectively; Condition 1 is checked at that single direction.

A.7 AVERAGE-LOSS DESCENT AND STATIONARITY

By Λ-smoothness and $\alpha \le 1/\Lambda$,

$$\begin{aligned}\bar{L}(\theta_{t+1}) - \bar{L}(\theta_t) &\le -\alpha \bar{g}_t^\top d_t + \frac{\Lambda \alpha^2}{2}\|d_t\|_2^2 \\ &\le -\frac{\alpha}{2}\left(\|\bar{g}_t\|_2^2 - \|d_t - \bar{g}_t\|_2^2\right) \\ &\le -\frac{\alpha(1-c^2)}{2}\|\bar{g}_t\|_2^2.\end{aligned}$$

The second line uses $2\bar{g}_t^\top d_t - \|d_t\|_2^2 = \|\bar{g}_t\|_2^2 - \|d_t - \bar{g}_t\|_2^2$. The last line uses $d_t \in \mathcal{A}_c(\bar{g}_t)$ and proves Eq. (25). Summing over $t = 0, \dots, T-1$ yields

$$\frac{\alpha(1-c^2)}{2}\sum_{t=0}^{T-1}\|\bar{g}_t\|_2^2 \le \bar{L}(\theta_0) - \bar{L}(\theta_T) \le \bar{L}(\theta_0) - \bar{L}^{\inf}.$$

The minimum is at most the average, giving Eq. (26). The proof uses only the feasible set, which grouping preserves by Eq. (20).

B OPTIMIZATION AND RELATED METHODS

We first give the reduced optimization problem used to select $d^\star$ and the checks on its numerical solution. The following subsections specify metric damping, derive the two-task example, and compare first-order directions.

B.1 ALGORITHM

For a set of represented terms, KAGrad uses their gradients, supplied positive-definite metrics, the weights defining $\bar{g}$, and radius c as follows. An *active represented term* passes the normalization filter specified in Appendix B.4; the average gradient retains every term.

1. Form the weighted average gradient $\bar{g}$ that defines $\mathcal{A}_c(\bar{g})$.
2. For each active represented term, compute $q_i = g_i^\top F_i^{-1} g_i$, the reference update $d_i = F_i^{-1} g_i$, and the normalized quadratic $\rho_i(d)$ in Eq. (14).
3. Solve the convex minimax problem in Eq. (16) through the reduced dual below.
4. Return the primal direction $d^\star$ and recompute its feasibility and normalized worst-task objective directly from $\{g_i, F_i\}$.

The normalization filter for numerically negligible gradients and the damping used for learned diagonal metrics are given in Appendix B.4.

B.2 REDUCED DUAL

For each active represented term define

$$q_i = g_i^\top F_i^{-1} g_i, \qquad \widetilde{F}_i = F_i/q_i, \qquad \widetilde{g}_i = g_i/q_i. \tag{29}$$

Then

$$\rho_i(d) = d^\top \widetilde{F}_i d - 2\widetilde{g}_i^\top d + 1. \tag{30}$$

Let $\mathcal{I}$ be the active represented terms and $B = c\|\bar{g}\|_2$. The epigraph form is

$$\min_{d,t}\ t \quad \text{s.t.} \quad \rho_i(d) \le t\ (i \in \mathcal{I}), \qquad \|d - \bar{g}\|_2^2 \le B^2. \tag{31}$$

Using multipliers $\lambda_i \ge 0$ with $\sum_{i\in\mathcal{I}} \lambda_i = 1$ and $\mu \ge 0$, define

$$A(\lambda, \mu) = \sum_{i\in\mathcal{I}} \lambda_i \widetilde{F}_i + \mu I, \tag{32}$$

$$b(\lambda, \mu) = \sum_{i\in\mathcal{I}} \lambda_i \widetilde{g}_i + \mu \bar{g}. \tag{33}$$

Minimization of the Lagrangian over d gives

$$d(\lambda, \mu) = A(\lambda, \mu)^{-1} b(\lambda, \mu), \tag{34}$$

and the dual objective

$$\max_{\lambda \in \Delta,\, \mu \geq 0} \left\{ 1 + \mu(1 - c^2)\|\bar{g}\|_2^2 - b(\lambda, \mu)^\top A(\lambda, \mu)^{-1} b(\lambda, \mu) \right\}, \tag{35}$$

where Δ is the probability simplex over $\mathcal{I}$. For $0 < c < 1$ and $\bar{g} \neq 0$, Slater's condition holds for Eq. (31). The boundary case $c = 0$ has the single feasible update $d = \bar{g}$, and if $\bar{g} = 0$ then the feasible set contains only the zero update for $c < 1$.

With diagonal metrics, the inverse in Eq. (34) is elementwise, so the numerical optimization dimension is $|\mathcal{I}| + 1$ rather than the neural-network parameter dimension. The reported implementation solves the reduced problem with SLSQP and independently checks primal feasibility, the returned objective, and the primal–dual gap. The supervised implementation uses a maximum of 200 iterations, 10^{-10} function tolerance, and 10^{-5} feasibility and primal–dual-gap tolerances. Numerical validity is distinct from the local-model certificate: a valid solution may have $\rho^\star_{\max} > 1$.

B.3 Numerical verification

To check the solver's reported objective against Eq. (14), note that the implementation stores a half-scaled epigraph value,

$$t_{\text{solver}} = \frac{1}{2} \max_{i \in \mathcal{I}} \rho_i(d), \tag{36}$$

so the paper-level quantity is $\rho_{\max}(d) = 2t_{\text{solver}}$. We independently recompute $\rho_{\max}(d)$ from the returned direction and task data. Across the reported verification checkpoints, the maximum absolute residual relative to $2t_{\text{solver}}$ is 9.13×10^{-10}.

B.4 Metric damping and active terms

The normalized discrepancy in Eq. (14) requires $q_i = g_i^\top F_i^{-1} g_i > 0$ for each term in the solve. Let m be the median of the finite q_i values above 10^{-12}. When such values exist, retain terms with finite $q_i > \max\{10^{-12}, 10^{-8} m\}$; otherwise the threshold is 10^{-12}. The full average gradient $\bar{g}$ still centers the feasible set. For learned-model experiments, diagonal estimates are damped as

$$F = \widehat{F} + \left(\delta_{\text{abs}} + \delta_{\text{rel}} \operatorname{mean}(\widehat{F}) \right) I \tag{37}$$

The probe counts below are fixed by configuration. Reinforcement-learning experiments use $\delta_{\text{abs}} = 10^{-6}$ and $\delta_{\text{rel}} = 0.01$. MT10 uses P16 Fisher estimates; MT50 uses P2 estimates for screening and P4 estimates for the represented terms. Cityscapes uses P4 estimates and Multi-Fashion+MNIST uses P8 estimates.

B.5 Two-task example

The example in Section 3 has identical gradients and different task metrics. Here we compute the update and the smallest radius that admits nonnegative model improvement for both tasks.

With $g_1 = g_2 = (1, 1)^\top$, $F_1 = \operatorname{diag}(4, 1)$, and $F_2 = \operatorname{diag}(1, 4)$, we have $d_1 = (1/4, 1)^\top$ and $d_2 = (1, 1/4)^\top$. At $\alpha = 1$,

$$I_1(x, y) = x + y - \tfrac{1}{2}(4x^2 + y^2), \qquad I_2(x, y) = x + y - \tfrac{1}{2}(x^2 + 4y^2),$$

and

$$I_1^{\max} = I_2^{\max} = \frac{5}{8}.$$

At $\bar{g} = (1, 1)$, $I_1(\bar{g}) = I_2(\bar{g}) = -\tfrac{1}{2}$, hence $r_1(\bar{g}) = r_2(\bar{g}) = -0.8$.

The problem is symmetric under exchange of the two coordinates. Along the diagonal $x = y = z$,

$$I_1(z, z) = I_2(z, z) = 2z - \frac{5}{2} z^2,$$

which is maximized at $z = 0.4$. Therefore the unconstrained max–min update is $d = (0.4, 0.4)^\top$ and $r_1(d) = r_2(d) = 0.64$.

Along the diagonal, nonnegative local-model improvement requires $2z - \frac{5}{2}z^2 \geq 0$, or $0 \leq z \leq 0.8$. By symmetry and convexity of the intersection of the two local-model safe sets, the feasible point nearest $\bar{g} = (1, 1)$ is $(0.8, 0.8)$. Its distance from $\bar{g}$ is $0.2\|\bar{g}\|_2$, so $c = 0.2$ is the smallest feasible-set radius that contains an update with nonnegative predicted improvement for both tasks. At that point $r_1 = r_2 = 0$ and $\rho_{\max} = 1$. The unconstrained optimum $(0.4, 0.4)$ lies at distance $0.6\|\bar{g}\|_2$ from $\bar{g}$ and first becomes feasible at $c = 0.6$. Thus $d^\star = (1 - c)\bar{g}$ for $0 \leq c < 0.6$, and $d^\star = (0.4, 0.4)^\top$ for $c \geq 0.6$.

B.6 Comparison with first-order methods

The methods below use task gradients without KAGrad's task-specific quadratic models. Their update criteria clarify which properties arise from the feasible set and which arise from the retained-improvement criterion.

If d_{MGDA} is the minimum-norm element of the convex hull of the task gradients, its projection property gives

$$g_i^\top d_{\mathrm{MGDA}} \geq \|d_{\mathrm{MGDA}}\|_2^2 \qquad \text{for every } i.$$

Thus, a nonzero MGDA solution is a common first-order descent update, while a zero solution gives the usual Pareto-stationarity condition. This statement is first-order and does not use the KAGrad feasible set.

CAGrad uses the same feasible set $\mathcal{A}_c(\bar{g})$ as KAGrad. Its criterion maximizes the smallest first-order projection $g_i^\top d$, whereas KAGrad maximizes the smallest retained-improvement fraction $r_i(d)$ defined by the task-specific local improvement models. The shared feasible set explains why the direct-update convergence argument in Theorem 2 does not depend on which of these criteria selects the feasible update. The two criteria are otherwise different; neither is obtained from the other merely by choosing a particular F_i.

B.6.1 Two-task Nash-MTL direction

For comparison with KAGrad's metric-dependent references, we derive the two-task Nash-MTL direction from its first-order coefficients. Let $G = [g_1, g_2]$ and let the positive Nash-MTL coefficients $a = (a_1, a_2)$ satisfy

$$G^\top G a = 1/a, \tag{38}$$

where the division on the right is elementwise. Write $\gamma = g_1^\top g_2/(\|g_1\|_2\|g_2\|_2)$, $x = \|g_1\|_2 a_1$, and $y = \|g_2\|_2 a_2$. The two stationarity equations become

$$x^2 + \gamma xy = 1, \qquad y^2 + \gamma xy = 1. \tag{39}$$

Subtracting them gives $x = y$ because $x, y > 0$. For $\gamma > -1$, $x = y = (1 + \gamma)^{-1/2}$, and hence

$$Ga = \frac{1}{\sqrt{1 + \gamma}} \left(\frac{g_1}{\|g_1\|_2} + \frac{g_2}{\|g_2\|_2} \right). \tag{40}$$

The two-task pre-optimizer direction is proportional to the sum of the unit-normalized task gradients.

B.7 Related methods

Gradient-based multitask methods. MGDA finds a minimum-norm convex combination of task gradients (Sener and Koltun, 2018), while PCGrad removes pairwise conflicting components (Yu et al., 2020a). CAGrad maximizes the smallest first-order task projection within a Euclidean ball around the average-loss gradient (Liu et al., 2021a); Nash-MTL formulates gradient combination as a bargaining problem (Navon et al., 2022). KAGrad retains CAGrad's feasible set but scores each candidate using a local improvement model for each task. It maximizes the smallest ratio of predicted improvement to the model's maximum at the reference update.

Other gradient-based methods. GradNorm adjusts task weights to balance gradient magnitudes (Chen et al., 2018), and IMTL-G chooses weights so the aggregate gradient has equal projections onto task gradients (Liu et al., 2021b). RotoGrad rescales task gradients and rotates the shared feature representation separately for each task to reduce directional conflict (Javaloy and Valera, 2022). Aligned-MTL uses the condition number of the task-gradient system to guide component alignment (Senushkin et al., 2023). MoCo corrects bias in stochastic multi-objective gradient estimates (Fernando et al., 2023), while FAMO adapts task weights from changes in task losses without computing every task gradient (Liu et al., 2023). KAGrad uses a supplied metric for each task to determine the improvement assigned to a candidate shared direction.

Fisher geometry. For a single probabilistic objective, the natural gradient uses the inverse Fisher information matrix (Amari, 1998); trust region policy optimization constrains KL change in a policy distribution (Schulman et al., 2015). In continual learning, elastic weight consolidation uses diagonal Fisher-weighted quadratic penalties to retain parameters learned on earlier tasks (Kirkpatrick et al., 2017). CoMOGA permits a common Fisher metric to define the local region for constrained multi-objective policy updates, with linearized constraints on objective improvement (Kim et al., 2025). KAGrad places a separate supplied metric in each task's local improvement model for simultaneous training. With exact Fishers, its normalized metric discrepancy is the ratio of quadratic local-KL terms for $d - d_i$ and d_i; the constraint around the average-loss gradient remains Euclidean rather than a KL trust region.

C Quadratic Experiments

We first specify the common generator. The following subsections report the identical-gradient case, the Hessian-rotation study, the finite-step condition check, and the effect of grouping.

C.1 Task generation

For the rotation study, draw independent Gaussian matrices X_i and set

$$\widetilde{A}_i = \frac{X_i + X_i^\top}{2}. \tag{41}$$

Center them across tasks,

$$\bar{A}_i = \widetilde{A}_i - \frac{1}{K}\sum_{j=1}^{K} \widetilde{A}_j, \tag{42}$$

and apply one common rescaling,

$$A_i = \frac{\bar{A}_i}{\max_j \|\bar{A}_j\|_2}, \tag{43}$$

so that $\max_i \|A_i\|_2 = 1$. The supplied metric is

$$F_i = I + \tanh(\beta) A_i, \tag{44}$$

which is positive definite because $\|A_i\|_2 \le 1$ and $|\tanh(\beta)| < 1$. A unit base vector b and independent unit vectors z_i give

$$g_i = \frac{b + \sigma z_i}{\|b + \sigma z_i\|_2}. \tag{45}$$

All random quantities are drawn once per family and reused across the entire rotation grid, so every comparison across rotation is paired.

For the spectrum-preserving mismatch, draw a Gaussian matrix Y_i and form the skew-symmetric generator $S_i = Y_i - Y_i^\top$. Set $\Omega_i = S_i / \|S_i\|_2$ and define

$$H_i(\phi) = Q_i(\phi) F_i Q_i(\phi)^\top, \qquad Q_i(\phi) = \exp\left(\frac{\pi\phi}{180}\Omega_i\right). \tag{46}$$

Orthogonality of $Q_i(\phi)$ gives $\mathrm{spec}(H_i(\phi)) = \mathrm{spec}(F_i)$ exactly. Numerical checks give maximum eigenvalue error below 4.3×10^{-15}, relative condition-number error below 1.1×10^{-14}, and gradient reconstruction error below 1.3×10^{-15}. At $\beta = 0$, $F_i = I$, so rotation leaves the Hessian unchanged and gives worst-task fraction 1 for both KAGrad variants, a zero-anisotropy check.

The identical-gradient experiment is described separately in Appendix C.2.

C.2 Identical gradients and different metrics

The experiment uses 500 families, each with $K = 4$ tasks in dimension 2. Within a family, fix a common unit gradient g (its orientation is immaterial under uniform rotations) and draw independent rotation angles uniformly on $[0, 2\pi)$. Let Q_i be the corresponding rotation matrices and set

$$F_i = H_i = Q_i \operatorname{diag}(e^{\beta}, e^{-\beta}) Q_i^{\top}, \qquad \beta = 1.25. \tag{47}$$

The task metrics share eigenvalues and unit determinant; they differ only in orientation. Setting $\theta_i^{\star} = \theta - F_i^{-1} g$ gives quadratic Gaussian location tasks with covariance F_i^{-1} and identical gradients at θ. We score the directions selected by both methods using these same Hessian-matched task models. Identity KAGrad uses identity metrics to select its direction and returns $d = \bar{g} = g$ for every c. Under the common evaluation, its mean worst-task retained fraction is -2.644, with nonnegative fractions for all tasks in 7.2% of families. Supplied-metric KAGrad gives -0.963 and 20.0% at $c = 0.2$, and 0.301 and 100% at $c = 0.5$. Identity KAGrad's evaluation values do not change with c. Since $H_i = F_i$, the matched-model score equals the relative actual decrease at $\alpha = 1$.

The following sweep varies anisotropy through β and reports Identity KAGrad's matched-model score across 500 families. **Table: Identical-gradient anisotropy sweep.**

β	Mean worst-task fraction	All-task nonnegative (%)
0	1.000	100.0
0.25	0.925	100.0
0.50	0.643	100.0
0.75	0.049	46.6
1.00	−0.992	15.2
1.25	−2.644	7.2

At $\beta = 0$, every task has $F_i = I$ and the fraction equals one.

C.3 Rotation of loss curvature

C.3.1 Rotation design

The primary grid uses 300 paired families, $K = 50$, dimension 4, $c = 0.5$, $\alpha = 1$, $\beta = 1.5$, and

$$\phi \in \{0, 4.5, 9, \ldots, 45\} \text{ degrees.} \tag{48}$$

We report $\sigma \in \{0.6, 0.8, 1.0\}$. Supplied-metric KAGrad, Identity KAGrad, and the Oracle-Hessian reference are all solved over the same feasible set $\mathcal{A}_c(\bar{g})$. The Oracle-Hessian method uses $H_i(\phi)$ as its metric and, at $\alpha = 1$, therefore maximizes $r_{\min}^{\text{loss}}(d; \phi)$ exactly over that feasible set.

A linear interpolation between isospectral endpoints does not preserve the curvature spectrum. If

$$J_i = Q_i(45) F_i Q_i(45)^{\top} \tag{49}$$

and

$$H_i^{\text{lin}} = (1 - m) F_i + m J_i, \tag{50}$$

then this task generator gives median $\kappa(H_i^{\text{lin}})/\kappa(F_i)$ of 0.928, 0.904, and 0.927 at $m = 0.25, 0.5, 0.75$, with 5th percentiles 0.780, 0.717, and 0.778. Equation (46) avoids this confound because orthogonal conjugation preserves the eigenvalues exactly. We report the accumulated rotation ϕ directly.

C.3.2 Worst-task loss decrease

The table reports the full $\sigma = 0.6$ rotation sweep. The Oracle, Metric, and Identity columns are means of Eq. (27) across the 300 families. The ratio columns divide each method's mean by the Oracle mean at that rotation; the difference column averages the paired family-level Metric-minus-Identity differences. **Table: Worst-task relative actual decrease by rotation.**

ϕ	Oracle	Metric	Identity	Metric/Oracle	Identity/Oracle	Metric−Identity
0°	.3315	.3315	.2679	1.000	.808	$+.0635$
4.5°	.3319	.3204	.2688	.965	.810	$+.0516$
9°	.3322	.3093	.2695	.931	.811	$+.0398$
13.5°	.3326	.2987	.2699	.898	.811	$+.0289$
18°	.3331	.2892	.2702	.868	.811	$+.0190$
22.5°	.3337	.2804	.2704	.840	.810	$+.0100$
27°	.3344	.2727	.2709	.815	.810	$+.0018$
31.5°	.3353	.2654	.2714	.792	.810	$-.0060$
36°	.3362	.2588	.2722	.770	.810	$-.0134$
40.5°	.3373	.2530	.2731	.750	.810	$-.0201$
45°	.3383	.2481	.2737	.733	.809	$-.0257$

Paired 95% bootstrap intervals for metric-minus-Identity at 22.5°, 27°, 31.5°, 36°, and 45° are $[0.0039, 0.0160]$, $[-0.0047, 0.0083]$, $[-0.0131, 0.0009]$, $[-0.0212, -0.0061]$, and $[-0.0338, -0.0181]$. The Oracle-minus-metric gap is strictly positive with interval excluding zero at every nonzero rotation; at 45° it is 0.0902 with interval $[0.0831, 0.0977]$.

At $\sigma = 0.8$, Metric/Oracle falls from 1.000 to 0.716 while Identity/Oracle remains near 0.771; the metric–Identity pointwise interval first excludes zero on the negative side at 40.5°. At $\sigma = 1.0$, Metric/Oracle falls from 1.000 to 0.624 while Identity/Oracle remains near 0.644; the endpoint paired difference is -0.0014 with interval $[-0.0037, 0.0010]$, which includes zero. The same generator and rotation grid are used at $\sigma = 0.6, 0.8, 1.0$; only task-gradient dispersion changes.

C.3.3 Direction and magnitude

For every method we maximize $r_{\min}^{\text{loss}}(td; \phi)$ over scalars $t \geq 0$ that keep the scaled update in $\mathcal{A}_c(\bar{g})$. Let R_M^{ray} denote the resulting maximum for method M and R_M^{applied} the value of its applied update. If R_O denotes the Oracle-Hessian value, then

$$R_O - R_M^{\text{applied}} = (R_O - R_M^{\text{ray}}) + (R_M^{\text{ray}} - R_M^{\text{applied}}). \tag{51}$$

The first term is the gap remaining after optimizing magnitude along the selected ray; the second is the gain from that rescaling. The Oracle-Hessian update has best scalar exactly 1 at every rotation. The supplied-metric KAGrad update also has best scalar exactly 1 at zero rotation.

For supplied-metric KAGrad, the two mean contributions are: **Table: Ray and scale contributions to the Oracle gap.**

ϕ	Ray cost	Scale cost	Ray share	Best scalar
4.5°	.0050	.0065	.434	.991
13.5°	.0157	.0182	.462	.976
22.5°	.0257	.0276	.482	.963
31.5°	.0347	.0352	.497	.953
36°	.0389	.0385	.502	.948
45°	.0470	.0432	.521	.945

The supplied-metric KAGrad update's applied norm is fixed at $0.9547\|\bar{g}\|_2$ because the quantities (g_i, F_i) that determine its update do not change across the rotation sweep. Its scalar-optimal norm falls to $0.8978\|\bar{g}\|_2$ at 45°. Figure 2 shows the corresponding gap decomposition.

As an additional first-order comparison, we evaluate CAGrad on the same quadratic families. Identity KAGrad and CAGrad are nearly collinear: their mean angular separation is 0.414°, whereas supplied-metric KAGrad is 9.87° from Identity KAGrad and 9.79° from CAGrad. At 45°, supplied-metric KAGrad is 9.04° from the Oracle-Hessian update. After scalar calibration, mean values at 45° are 0.2917 for CAGrad, 0.2913 for supplied-metric KAGrad, and 0.2913 for Identity KAGrad. Paired mean differences are 0.00040 for CAGrad-minus-supplied-metric (interval $[-0.00398, 0.00476]$), 0.00037 for CAGrad-minus-Identity (interval $[-0.00003, 0.00078]$), and -0.00003 for metric-minus-Identity (interval $[-0.00452, 0.00429]$). The narrow CAGrad–Identity comparison reflects their nearly identical rays: 72.3% of families differ by less than 0.001. Supplied-metric comparisons are heterogeneous across families: supplied-metric KAGrad beats Identity KAGrad in 54.3% of families and loses in 45.7%, with paired-difference standard deviation 0.0387.

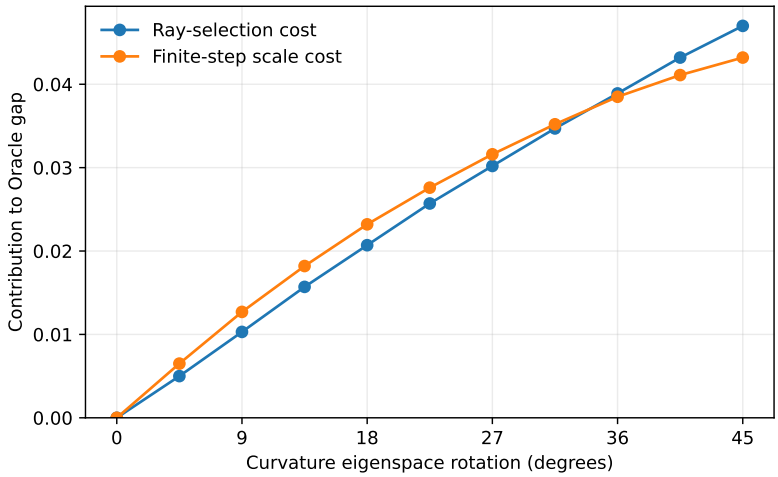

Figure 2: Decomposition of the supplied-metric–Oracle gap in the $\sigma = 0.6$ family. Both the selected update ray and the applied finite-step magnitude contribute as the supplied metric and loss-curvature eigenspaces separate.

CAGrad uses the full ball radius, $\|d - \bar{g}\|_2/(c\|\bar{g}\|_2) = 1$, and its mean norm is $1.492\|\bar{g}\|_2$, while its scalar-optimal norm is about $0.94\|\bar{g}\|_2$. The difference between this applied magnitude and its scalar optimum explains its as-applied quadratic loss decrease in this construction. CAGrad's linear first-order objective contains no curvature-dependent magnitude term, whereas KAGrad's quadratic local model selects both a ray and a finite update magnitude within the feasible set.

C.4 Finite-step condition and violations

The model-prediction test $r^\star_{\min} \geq 0$ does not itself guarantee finite-step task-loss non-increase. We check the additional Condition 1 for the $\sigma = 1.0$ rotation families using Eq. (28). Among all 300 families, the fraction satisfying that condition at each rotation and step size is: **Table: Fraction satisfying Condition 1.**

Rotation	$\alpha = .25$	$\alpha = .50$	$\alpha = .75$	$\alpha = 1.00$
0°	1.000	1.000	1.000	1.000
4.5°	1.000	1.000	1.000	0.000
9°	1.000	1.000	.887	0.000
13.5°	1.000	1.000	.527	0.000
18°	1.000	.967	.170	0.000
22.5°	1.000	.877	.040	0.000
27°	1.000	.773	0.000	0.000
31.5°	.997	.633	0.000	0.000
36°	.987	.487	0.000	0.000
40.5°	.963	.323	0.000	0.000
45°	.930	.217	0.000	0.000

At $\alpha = 1$, supplied-metric KAGrad is locally certified in the same $296/300$ families at every rotation. The finite-step loss-violation counts in this fixed cohort are $0, 0, 0, 2, 5, 7, 9, 10, 12, 12, 15$ from $0°$ to $45°$. At the endpoint this is 5.07% with Wilson 95% interval $[3.09\%, 8.19\%]$. Endpoint violation rates are 1.0% and 1.67% at $\sigma = 0.6$ and 0.8, respectively. Figure 3 shows the violation rate across the rotation sweep.

The condition rates above and the fixed $296/300$ certified cohort are family-level marginals. Their exact intersection contains 7,042 $(\text{family}, \phi, \alpha)$ cases, none of which has a finite-step loss violation, as required by Theorem 1. At $\alpha = .25, .5, .75$, violations are zero both when Condition 1 holds and when it fails. At $\alpha = 1$, the condition holds only at zero rotation ($0/296$ violations) and fails in all nonzero-rotation certified cases, with $72/2960 = 2.43\%$ violations overall.

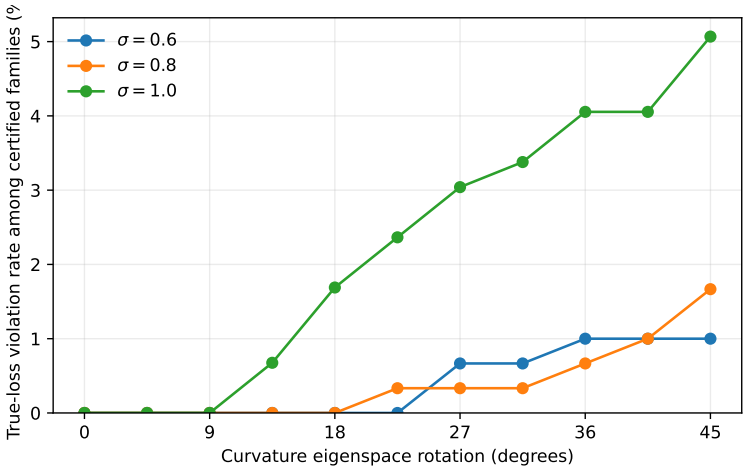

Figure 3: Finite-step loss-violation rate among families with $r^{\star}_{\min} \geq 0$ under the curvature-rotation experiment. Within each family the local-model certificate is unchanged across rotation, while the finite-step loss curvature changes.

C.5 Grouping and screening

Grouping preserves the average-gradient center but changes which terms enter the minimax objective. We specify the aggregate metric, check original-task improvement, and then give a separate Meta-World screening-cost diagnostic.

We rank tasks using the normalized metric discrepancy of $\bar{g}$ computed with estimated metrics:

$$\widehat{\rho}_i(\bar{g}) = \frac{(\bar{g} - \widehat{F}_i^{-1} g_i)^\top \widehat{F}_i (\bar{g} - \widehat{F}_i^{-1} g_i)}{g_i^\top \widehat{F}_i^{-1} g_i}. \tag{52}$$

With P probes per task, evaluating the singleton terms and aggregate uses

$$(M-1)P + |R|P = KP \tag{53}$$

score-gradient evaluations. Thus grouping reduces minimax terms but not this estimator's score-gradient work.

C.5.1 Aggregate metric and feasible set

For singleton set S and aggregate set R, the grouped problem uses $|S| = M - 1$ singleton terms and the aggregate loss $L_R = |R|^{-1} \sum_{i \in R} L_i$, with gradient $g_R = |R|^{-1} \sum_{i \in R} g_i$. Equation (20) shows that group-size weighting recovers the original average task gradient exactly, so the feasible set $\mathcal{A}_c(\bar{g})$ is unchanged. In the quadratic-task experiments we use $F_R = |R|^{-1} \sum_{i \in R} F_i$. In stochastic implementations, an aggregate probe assigns independent Rademacher signs to the tasks in R, sums their signed score gradients with factor $|R|^{-1/2}$, and forms the outer-product estimate. Its expectation is the mean task Fisher because the cross terms vanish. The local-model certificate for this term therefore concerns the aggregate loss, not each task in R individually.

Figure 4 summarizes the grouped representation.

C.5.2 Original-task check

The screening comparison uses $K = 50$, $\beta = 1.5$, $\sigma = 0.60$, $c = 0.5$, 50 families, and three screening draws per family, with $M = 8$. Tasks are ranked by Eq. (52) using P1, P2, or P4 estimates with 1% relative isotropic damping. Here "top-7" selects the seven largest estimated scores; "6+1" selects six and reserves one for deterministic coverage. Mean top-7 recovered counts the overlap with the seven largest exact-metric scores; median worst retained is the median $r_{\min}$ over all original tasks. For this comparison, selected terms are re-estimated at P16 for the KAGrad solve; the MT50 benchmark configuration is specified separately in Appendix D.3.

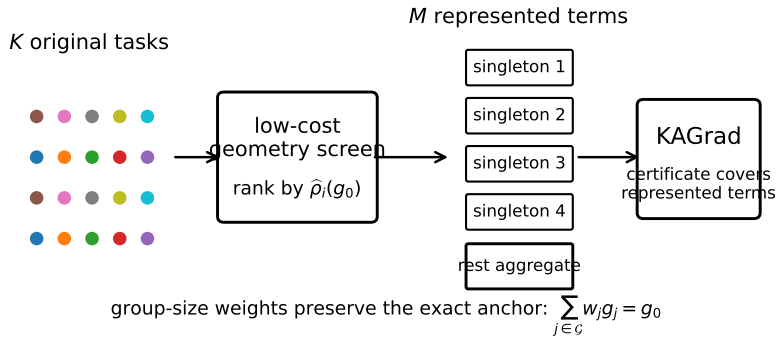

Figure 4: Grouped task representation. A subset of tasks is represented individually and the remainder form one aggregate. Group-size weighting preserves the exact average task gradient used to center the KAGrad feasible set.

Table: Screening recovery and original-task checks.

Screen	Mean top-7 recovered	S_{safe}	95% CI	Median worst retained
P1, top-7	1.84	36.0%	[27.3, 44.7]%	−0.157
P1, 6+1	1.67	35.3%	[25.3, 45.3]%	−0.160
P2, top-7	2.20	46.0%	[35.3, 56.7]%	−0.063
P2, 6+1	2.00	46.7%	[36.0, 58.0]%	−0.085
P4, top-7	2.70	51.3%	[40.7, 62.0]%	+0.021
P4, 6+1	2.51	53.3%	[42.0, 64.0]%	+0.025

For a grouped solution $d^{\star}_{\text{grp}}$, define the full-task local-model-safe event by

$$r_i(d^{\star}_{\text{grp}}) \geq 0 \qquad \text{for every original task } i \in [K]. \tag{54}$$

Here S_{safe} is the fraction of evaluated directions satisfying Eq. (54). Random and approximate screens use three draws per family; intervals resample families. Random selection reaches 22.0%, the full method 86.0%, and exact-score selection with one coverage slot 84.0%. Using a 35% selection threshold, P2 is the lowest tested probe count whose interval clears the threshold; it is therefore used for MT50. P4 is included as a higher-probe comparison. At $K = 100$, exact-score selection with $M = 8$ reaches 54.7%, compared with 62.0% for full KAGrad.

Figure 5 compares these screening choices.

C.5.3 Screening accuracy and cost

In a separate higher-probe screening diagnostic, tasks are ranked at three actor checkpoints using a gradient-only score and P1, P2, and P16 versions of Eq. (52). Mean top-six recovery relative to the P16 ranking is $0.667/6$, $2.000/6$, and $2.667/6$ for the gradient-only, P1, and P2 rankings, respectively. Mean Spearman correlations against the P16 ranking are 0.4313 for P1 and 0.5970 for P2.

Cost accounting. Measured phase times are 0.101 s for all task gradients, 0.105 s for the P1 screen, a further 0.095 s for P2, 1.283 s to refine represented terms to P16, and 1.483 s for an all-task P16 computation. These higher-probe measurements characterize the screening mechanism rather than the MT50 benchmark probe count.

The staged metric computation totals 1.483 s, matching the all-task P16 computation. This is consistent with Eq. (53): the aggregate estimator still processes every rest-task sample, so grouping does not reduce Fisher-estimation work in this implementation. It can still reduce the number of terms in the minimax solve; the measured Multi-Fashion+MNIST runtimes appear in Appendix D.2.

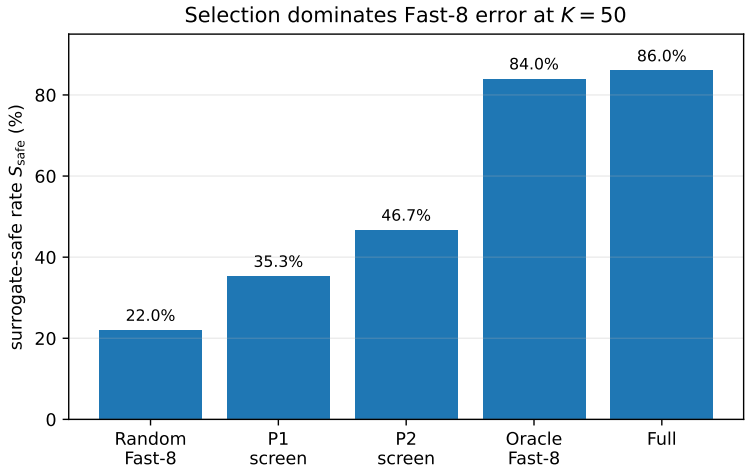

Figure 5: Post-hoc full-task local-model-safe rate at $K = 50$, $M = 8$ for random selection, P1/P2/P4 approximate screening, exact-score screening (all grouped methods use 6+1 selection), and full KAGrad. The P2 screen is used for MT50.

D Learned-Model Protocols and Diagnostics

In the supervised and Meta-World experiments, KAGrad computes a direction before an Adam step. The following protocols identify the metrics supplied to that computation and the checkpoints used for comparisons. The diagnostics separately measure estimation error, metric scale, and the sampled loss-Hessian structure.

D.1 Fisher estimation and metric scale

The learned-model experiments supply damped diagonal Fisher estimates. The first diagnostics measure how probe count changes the estimate and the selected direction; the final diagnostic varies metric scale.

The reinforcement-learning implementation uses a damped diagonal policy Fisher

$$F_i = \widehat{F}_i + \left(10^{-6} + 0.01\,\mathrm{mean}(\widehat{F}_i)\right) I. \tag{55}$$

MT10 uses P16 estimates. On MT50, screening uses P2 estimates and the represented-term metrics are then refined to P4. These choices keep score-gradient computation comparable across benchmarks because Eq. (53) gives total Fisher-estimation work proportional to KP.

For the MT50 probe count, a P4–P64 diagnostic gives normalized diagonal-Fisher discrepancy 0.603, where the discrepancy is the Euclidean norm of the difference between the two diagonal estimates divided by the norm of the P64 estimate. This quantity measures estimator variability and does not measure whether the Fisher is an appropriate model of loss curvature.

Separately, we evaluate solution sensitivity in a higher-probe setting. Let d_{16} and d_{256} denote KAGrad solutions obtained from P16 and P256 metrics, and let $\rho^{(256)}_{\max}$ denote the normalized KAGrad objective evaluated using the common P256 reference metrics. We report

$$\rho^{(256)}_{\max}(d_{16}) - \rho^{(256)}_{\max}(d_{256}). \tag{56}$$

Across six evaluation checkpoints, this reference-objective gap has mean 5.14×10^{-4} and maximum 9.14×10^{-4}, with mean angular difference 2.97° between the two solutions. This experiment characterizes solution sensitivity in the higher-probe setting and is distinct from the P4 estimator-fidelity diagnostic used to characterize the MT50 configuration.

D.1.1 METRIC-SCALE SENSITIVITY

For a positive common multiplier τ, the scaled metric τF_i gives the reference update

$$d_i^{(\tau)} = \frac{1}{\tau} F_i^{-1} g_i. \tag{57}$$

At saved checkpoints we recompute the KAGrad solution for $\tau \in \{0.1, 0.5, 1, 2, 10\}$. Relative to $\tau = 1$, the values $\tau = 0.1, 0.5, 2$, and 10 change the selected direction by $15.9°, 4.5°, 2.9°$, and $5.7°$, respectively. Thus the selected ray changes with metric scale. These scales are diagnostic references rather than values selected by benchmark performance.

D.2 SUPERVISED TASKS

We specify the Cityscapes protocol, then report paired Multi-Fashion+MNIST results and a separate Fisher–Hessian diagnostic.

D.2.1 CITYSCAPES

The Cityscapes experiment follows the two-task MTAN/SegNet setup used by the CAGrad benchmark, with KAGrad applied to the parameters returned by the model's shared modules.

Metric construction. The segmentation metric is a predictive categorical Fisher: labels are sampled from the model probabilities at valid pixels and score gradients are accumulated over the shared parameters. For depth, the metric is the predictive Fisher of a unit-variance Gaussian observation model over valid depth entries. Both metrics are estimated diagonally and damped by $10^{-6} + 0.01\ \mathrm{mean}(\widehat{F}_i)$ with a diagonal floor of 10^{-8}. Fisher KAGrad uses P4 estimates and $c = 0.9$; Identity KAGrad uses the same solver and radius with $F_i = I$.

Training protocol. The selected shared direction replaces the shared gradient before the benchmark Adam step. Models are trained for 200 epochs with batch size 8 and learning rate 10^{-4}, reduced to 5×10^{-5} after epoch 100. The reported test metrics are averaged over epochs 190–199 and then across seeds. Architecture, data processing, augmentation, and task losses otherwise follow the benchmark implementation. Theoretical direct-update statements concern the pre-optimizer KAGrad direction rather than Adam's transformed parameter step.

D.2.2 MULTI-FASHION+MNIST

The Multi-Fashion+MNIST experiment uses ten paired seeds, 50 epochs, batch size 256, $c = 0.2$, and P8 Fisher estimates. Training uses Adam with learning rate 10^{-3} and weight decay 10^{-2}, matching the benchmark. Methods share initialization and minibatch order within each pair, and the final epoch is reported. KAGrad replaces the shared gradient before the Adam step, so the direct-update guarantees again refer to the pre-optimizer direction.

Table: Final test average task loss (mean ± standard error across ten runs) and mean runtime.

Method	Final test average task loss	Mean runtime
CAGrad	0.5805 ± 0.0126	1.85 min
Identity KAGrad	0.5803 ± 0.0163	16.52 min
Fisher KAGrad	0.5783 ± 0.0165	21.40 min

Identity KAGrad uses no Fisher probes, yet is substantially more expensive than CAGrad. Most of the additional runtime in this implementation therefore arises from repeated evaluations of the simplex dual and its refinement rather than from Fisher estimation alone. We use these experiments to characterize the optimization formulation and the effect of the supplied metric rather than to claim a computational advantage over CAGrad.

For final test average task loss, the paired Identity-minus-Fisher difference is 0.002016, interval $[-0.018428, 0.022459]$, $t_9 = 0.2231$, $p = 0.8285$, and $d_z = 0.0705$, with Fisher KAGrad lower in 5 of 10 pairs. Against CAGrad, the Fisher-minus-CAGrad difference is -0.002210, interval

$[-0.019239, 0.014819]$, with $p = 0.7757$. No multiplicity correction is applied to these comparisons. Fisher KAGrad satisfies $r^\star_{\min} \geq 0$ on 14.7% of updates overall and 2.27% in the final epoch. Thus most updates have no feasible direction whose local models predict non-increase for both tasks. We do not compare certificate rates across metrics because each metric defines a different local model.

D.2.3 FISHER–HESSIAN COMPARISON

Three separately trained models are used only for diagnosis. For each model, five deterministic minibatches, both tasks, and eight common Rademacher Hessian-vector probes are evaluated against a P256 Fisher estimate at the same checkpoints and over the same shared parameters. We report the global Frobenius cosine

$$C_{FH} = \frac{\langle H, F \rangle_F}{\|H\|_F \|F\|_F}, \tag{58}$$

which is invariant to a positive scalar rescaling of either matrix. The mean cosine is 0.1524.

The Fisher-estimation discrepancy is

$$\frac{\|\widehat{F}_{\mathrm{P8}} - \widehat{F}_{\mathrm{P256}}\|_F}{\|\widehat{F}_{\mathrm{P256}}\|_F}, \tag{59}$$

and has mean 0.5241. Because the supplied Fisher approximation is diagonal whereas the sampled Hessian is full, the Frobenius cosine characterizes global correspondence between those two objects but does not isolate Fisher-specific metric–curvature mismatch. The quadratic rotation isolates eigenspace mismatch at fixed eigenvalues; this diagnostic combines Fisher–curvature disagreement with the off-diagonal information lost by a diagonal estimate.

D.3 META-WORLD PROTOCOL AND PEAK SUCCESS

This subsection specifies where KAGrad enters the SAC update, the fixed MT10 and MT50 configurations, and the checkpoints used to compare methods.

D.3.1 REFERENCE PROTOCOL

We use Soft Actor-Critic (SAC) (Haarnoja et al., 2018) on Meta-World MT10 and MT50 (Yu et al., 2020b) with two million environment steps, batch size 1280, and evaluation every 10,000 steps. Each method is evaluated with five independent seeds.

D.3.2 OPTIMIZATION AND METRIC CONSTRUCTION

KAGrad is applied only to the SAC actor. For task i, the actor-loss gradient g_i is computed on the task batch and the diagonal empirical policy Fisher is estimated from score gradients $\nabla_\theta \log \pi_\theta(a \mid s, i)$ with the sampled action held fixed. A Rademacher/Hutchinson estimator is used for singleton and aggregate terms. The critic and entropy-temperature parameters follow the standard SAC update paths and are not modified by KAGrad.

KAGrad produces a pre-optimizer actor direction $d^\star$. That vector is written as the actor gradient and the standard actor Adam optimizer applies the actual parameter step. Consequently, $r^\star_{\min}$ and the feasible-set statements in Sections 3–4 refer to the pre-Adam direction. The direct-update theorems do not claim guarantees for the post-Adam parameter displacement.

D.3.3 FIXED CONFIGURATIONS

MT10. MT10 represents all ten tasks and uses $c = 0.9$. We compare Identity KAGrad and Fisher KAGrad under the same training and evaluation protocol. Identity KAGrad uses $F_i = I$. Fisher KAGrad uses P16 diagonal policy-Fisher estimates with relative damping 0.01 and absolute damping 10^{-6}.

MT50. MT50 uses $M = 8$ represented terms: six singleton tasks selected by the screening score in Eq. (52), one coverage singleton, and one aggregate over the remaining tasks. Eight-term grouped KAGrad and CAGrad-Fast-8 both use $c = 0.9$ and eight represented terms. CAGrad-Fast-8 uses

seven singleton tasks selected by the benchmark's pseudorandom task-selection schedule and one aggregate over the remaining 43 tasks, so all 50 tasks enter its represented objective.

Singleton selection and probe budget. The six screened singleton slots use the largest P2 screening scores; the seventh follows a deterministic coverage schedule independent of the screening scores. The P2 screen is the low-probe stage of the MT50 P4 budget, after which the represented-term metrics are refined to P4 with the same damping. The screening rule, probe counts, damping coefficients, feasible-set radii, and representation rules are fixed across the reported runs. From Eq. (53), score-gradient work is KP per update irrespective of the grouped representation: MT10 uses $10 \times 16 = 160$ score-gradient evaluations and MT50 uses $50 \times 4 = 200$.

D.3.4 Synchronized success endpoints

For benchmark $B \in \{\mathrm{MT10}, \mathrm{MT50}\}$, let $\mathcal{M}_B$ denote the methods reported for benchmark B, and let $\mathcal{T}_{M,s}$ denote the set of evaluation checkpoints available for method M and seed s. We define one checkpoint set for the entire comparison,

$$\mathcal{T}_B = \bigcap_{M \in \mathcal{M}_B} \bigcap_{s=1}^{n} \mathcal{T}_{M,s}, \qquad n = 5. \tag{60}$$

All methods use this same checkpoint intersection. If $S_{M,s}(t)$ is the mean task success reported for method M, seed s, at checkpoint t, define the synchronized cohort mean

$$\bar{S}_M(t) = \frac{1}{n} \sum_{s=1}^{n} S_{M,s}(t). \tag{61}$$

Peak success is

$$S_{\mathrm{peak}}(M) = \max_{t \in \mathcal{T}_B} \bar{S}_M(t), \tag{62}$$

which is the maximum of the seed-cohort mean, not the mean of seed-specific maxima. The peak standard error is computed across $\{S_{M,s}(t^\star_M)\}_{s=1}^n$, where $t^\star_M$ attains the maximum in Eq. (62). We report descriptive statistics without significance tests.

For the final five checkpoints $\mathcal{T}_B^{(5)} \subseteq \mathcal{T}_B$, the late-window endpoint is

$$S_{\mathrm{late}}(M) = \frac{1}{5} \sum_{t \in \mathcal{T}_B^{(5)}} \bar{S}_M(t). \tag{63}$$

Its standard error is computed across seed-level means over those five checkpoints.

Table 4: Meta-World late-window mean task success over the final five common checkpoints. Values are mean ± standard error across five seeds.

Benchmark	Method	Late-window success ↑
MT10	Identity KAGrad	0.673 ± 0.103
	Fisher KAGrad	0.627 ± 0.075
MT50	CAGrad-Fast-8	0.401 ± 0.017
	Eight-term grouped KAGrad	0.441 ± 0.005

Both benchmarks retain the method ordering seen at peak success.

D.4 Meta-World actor-loss Hessian structure

This diagnostic characterizes how much of the sampled SAC actor-loss Hessian is contained on its diagonal. It is not a test of Condition 1. For task i, let H_i denote the Hessian of the sampled actor loss at a fixed trained checkpoint and state batch. Define its total Frobenius energy and diagonal energy as

$$E_i = \|H_i\|_F^2, \qquad D_i = \sum_{j=1}^{p} (H_i)_{jj}^2. \tag{64}$$

The fraction of curvature energy on the diagonal is

$$\eta_{\text{diag}} = \frac{\sum_i D_i}{\sum_i E_i}. \tag{65}$$

We estimate these quantities using $P = 8$ independent Rademacher vectors z_{ip} per task and Hessian-vector products $H_i z_{ip}$. The total-energy estimator is

$$\widehat{E}_i = \frac{1}{P} \sum_{p=1}^{P} \|H_i z_{ip}\|_2^2, \tag{66}$$

and the stochastic diagonal estimate is

$$\widehat{h}_i = \frac{1}{P} \sum_{p=1}^{P} z_{ip} \odot H_i z_{ip}. \tag{67}$$

The diagonal estimate satisfies $\mathbb{E}\|\widehat{h}_i\|_2^2 = D_i + (E_i - D_i)/P$, so its squared norm is upward biased for D_i. We use the unbiased correction

$$\widehat{D}_i = \frac{\|\widehat{h}_i\|_2^2 - \widehat{E}_i/P}{1 - 1/P}. \tag{68}$$

The pooled estimate substitutes $\widehat{D}_i$ and $\widehat{E}_i$ into Eq. (65). Across three independently trained Meta-World checkpoints, we recompute the pooled ratio after omitting each task in turn. The resulting diagonal-energy fractions lie between 1.76% and 9.03%. Thus at least 90.97% of the estimated Hessian Frobenius energy remains off-diagonal over this sensitivity range. This measures the fraction of these sampled Hessians that a diagonal metric could represent; it does not identify whether the Fisher is the appropriate non-diagonal metric and does not evaluate the path-averaged curvature inequality in Condition 1.