GEOMETRY-AWARE MULTI-OBJECTIVE OPTIMIZATION: NORM-CONSTRAINED ORACLES WITH ENTROPIC WEIGHT UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple-gradient descent is typically formulated in Euclidean geometry, whereas matrix-aware optimizers use directions induced by other norms. We analyze a single-loop multi-objective method combining a norm-constrained linear minimization oracle with entropic updates of task weights. For any finite-dimensional norm pair, we establish an $O(T^{-1/2})$ ergodic stationarity rate in the deterministic setting, deriving the required trajectory gradient bound from smoothness and lower-boundedness in the unregularized case. With stochastic gradients and a bounded-gradient assumption, one persistent blended-gradient momentum buffer yields an $O(T^{-1/4})$ rate whose leading term has no explicit dependence on the number of tasks. We also establish static regret for the inner weighting objective on arbitrary bounded gradient sequences and, with entropy regularization and a weight floor, iteration-wise inner-gap control after burn-in while retaining ergodic stationarity, conditional on a trajectory gradient bound. The optimal blended gradient can be nonunique in spectral and sign geometries, motivating a selection-independent inner gap. Our analysis extends the spectral construction studied by concurrent MOON to general norms. Experiments compare the method with existing multi-task optimizers and examine its oracle, weighting, and momentum choices.

1 INTRODUCTION

Many learning problems optimize several objectives at once: a shared backbone serving segmentation and depth, a language model balancing helpfulness against safety. Writing $f_1, \dots, f_K : \mathbb{R}^d \to \mathbb{R}$ for the losses, one seeks a point that cannot be improved on one objective without degrading another: a Pareto-optimal point. Pareto stationarity is a first-order necessary condition for such optimality. A standard algorithm targeting this condition is multiple gradient descent (Désidéri, 2012; Fliege & Svaiter, 2000; Sener & Koltun, 2018): choose the convex combination of task gradients with smallest norm,

$$w^\star(x) \in \arg\min_{w \in \Delta_K} \|\nabla F(x) w\|_2, \qquad \nabla F(x) w := \sum_{i=1}^{K} w_i \nabla f_i(x), \tag{1}$$

and step along $-\nabla F(x) w^\star(x)$, the *conflict-avoidant* direction that minimizes $\max_i \langle \nabla f_i(x), d\rangle + \frac{1}{2}\|d\|_2^2$. A large literature refines the weighting (Yu et al., 2020; Chen et al., 2020; Liu et al., 2021; Navon et al., 2022; Liu et al., 2023; Ban & Ji, 2024) and proves rates for stochastic variants (Liu & Vicente, 2021; Zhou et al., 2022; Fernando et al., 2023; Chen et al., 2024; Xiao et al., 2023; Zhang et al., 2025).

Two choices in (1) are Euclidean by inheritance. The *parameter update* is steepest descent in ℓ_2, whereas a network's parameters are matrices with their own scales; optimizers that take an LMO step in an operator norm, as in Muon (Jordan et al., 2024), Scion (Pethick et al., 2025) and the duality view of Bernstein & Newhouse (2024; 2025); Large et al. (2024), have improved large-scale training performance (Liu et al., 2025). These methods are single-objective; concurrent MOON (Zhou et al., 2026) brings the spectral construction to MOO. The *weight update* lives on the simplex, where entropic mirror descent gives an $O(M\sqrt{\log K/T})$ regret bound under an ℓ_∞ subgradient bound M,

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.

compared with the $O(M\sqrt{K/T})$ bound obtained by converting that same bound to ℓ_2 for Euclidean projected descent (Nemirovsky & Yudin, 1983; Beck & Teboulle, 2003; Bubeck, 2015).

The closest methods already use elements of this construction. MSMD (Yang et al., 2024) uses entropic mirror descent on the weights with a Euclidean direction variable and a double loop. MOON (Zhou et al., 2026) combines a spectral-norm oracle with logit-based task weighting and establishes the same deterministic and stochastic rate exponents. Our contribution is a general-norm analysis, a derived deterministic trajectory gradient bound, and refinements for blended momentum and inner-problem tracking.

- **General-norm deterministic convergence.** Theorem 4.4 gives $\frac{1}{T}\sum_t \|\nabla F(x_t)w_t\|_* \leq (\sqrt{2\Delta L} + M\sqrt{\log K/2})/\sqrt{T}$. For the unregularized, unclipped method, Lemma 4.3 derives $M = O(\sqrt{L\Delta \log K})$ from smoothness and lower-boundedness, yielding $O(L\Delta \log^2 K/\epsilon^2)$ iterations without a separate gradient-bound assumption. Each iteration uses K task-gradient evaluations, one oracle call, and one entropic weight update.
- **Stochastic convergence with one persistent buffer.** Under a trajectory gradient bound, Theorem 4.9 gives $O(\epsilon^{-4})$ iterations with momentum on the blended gradient. A weighted-average bound on noisy weight increments keeps explicit K dependence out of the leading $T^{-1/4}$ term; lower-order terms retain linear K dependence under bounded variance, replaced by $\log K$ under a stronger tail assumption (Corollary D.4).
- **Selection-independent inner optimality and iteration-wise tracking.** The optimal blended gradient can be nonunique in the Muon and Signum geometries (Proposition 3.1), motivating the inner gap. Theorem 4.5 gives static regret on arbitrary bounded gradient sequences. Average gap and regret are also bounded by the blended-gradient residual, so alone they do not distinguish adaptation from a converged scalarization; with regularization, clipping and a trajectory gradient bound, Theorem 5.1 instead controls the inner gap at *every* iteration after a $\tilde{O}(M^4 \log^2 K/\epsilon^4)$ burn-in while retaining ergodic stationarity, in any norm but with a non-competitive exponent.

2 Related work

Geometry-aware optimizers. Non-Euclidean steepest descent is classical (Nesterov, 2012; Carlson et al., 2015); the current line begins with Bernstein & Newhouse (2024) and Muon (Jordan et al., 2024). Pethick et al. (2025) give the LMO formulation we build on, with $O(n^{-1/4})$ stochastic rates and layer-wise norm and radius choices; Bernstein & Newhouse (2025); Large et al. (2024) develop the duality view; Liu et al. (2025); Riabinin et al. (2025) address scaling, and Yang et al. (2021; 2023) hyperparameter transfer. All are single-objective.

Stochastic MGDA. Liu & Vicente (2021) give the first stochastic multi-gradient method; Zhou et al. (2022) show naive stochastic gradient manipulation need not converge; Fernando et al. (2023) fix the bias with a tracking variable; Chen et al. (2024) show a three-way trade-off between optimization, generalization and conflict avoidance; Xiao et al. (2023) add a direction-oriented regularizer. Zhang et al. (2025) analyze MGDA under generalized smoothness with no gradient bound, obtaining $O(\epsilon^{-2})$ and $O(\epsilon^{-4})$ for average CA distance and $O(\epsilon^{-11})$, $O(\epsilon^{-17})$ for iteration-wise CA distance; these are our reference point. All are Euclidean with projected weights.

The two closest methods. *MSMD* (Yang et al., 2024) solves the MGDA saddle problem by stochastic mirror descent with the negative entropy on the weights, so the $\log K$ benefit of entropic weights is theirs; its direction variable is Euclidean, it is double-loop, and it assumes bounded gradients. *MOON* (Zhou et al., 2026) uses spectral–nuclear geometry with momentum and logit-based task weighting. Its analysis assumes bounded gradients and losses and carries polynomial dependence on the number of objectives. Our spectral specialization shares this geometric construction; our stochastic analysis evaluates the oracle at the previous buffer to decorrelate the direction from the current minibatch, and our tracking variant adds a weight floor. Practical implementations also differ in momentum, logit updates, and auxiliary optimizers, as detailed in Appendix I; the empirical comparison does not isolate a single change. Our analytical additions are the arbitrary norm pair, the derived deterministic M, comparator-dependent regret, iteration-wise inner gap, nonuniqueness analysis, and the one-buffer K argument. Conditional-gradient MOO (Assunção et al., 2021) ap-

Primal $\|\cdot\|$	**Dual** $\|\cdot\|_*$	$\mathrm{lmo}(g)$	**Known as**
ℓ_2	ℓ_2	$-R\,g/\|g\|_2$	normalized SGD
ℓ_∞	ℓ_1	$-R\,\mathrm{sign}(g)$	Signum (Bernstein et al., 2018)
spectral	nuclear	$-RUV^\top,\ G=U\Sigma V^\top$	Muon (Jordan et al., 2024)

Table 1: Norm-constrained oracles; Pethick et al. (2025) apply them layer-wise with per-layer radii ρ_ℓ.

plies an oracle to the constraint set, a different object. See Chen et al. (2026) for a survey and Nemirovsky & Yudin (1983); Beck & Teboulle (2003); Kivinen & Warmuth (1997); Bubeck (2015) for the entropic mirror-descent bounds we use off the shelf.

3 SETUP

Norms and the oracle. Fix a norm $\|\cdot\|$ on $\mathbb{R}^d$, the *primal* norm, measuring parameter displacements, with dual $\|g\|_* := \sup_{\|v\|\le 1}\langle g, v\rangle$ measuring gradients; Hölder's inequality $\langle g, v\rangle \le \|g\|_*\|v\|$ holds throughout. Let $B := \{v : \|v\| \le R\}$. The *linear minimization oracle* is any selection

$$\mathrm{lmo}(g) \in \arg\min_{v\in B}\langle g, v\rangle, \qquad \text{so that} \qquad \|\mathrm{lmo}(g)\| \le R, \quad \langle g, \mathrm{lmo}(g)\rangle = -R\|g\|_*. \tag{2}$$

The identity holds for every element of the argmin and is the only property of lmo our analysis uses; we set $\mathrm{lmo}(0) = 0$, so that an exactly zero blend produces a zero parameter step. With entropy regularization, the weights can still change; the full state need not be a fixed point. Table 1 lists the instances of interest. For a network with parameter blocks $x = (x_\ell)_\ell$ the norm used in practice is $\|x\| := \max_\ell \|x_\ell\|_{(\ell)}/\rho_\ell$, with dual $\|g\|_* = \sum_\ell \rho_\ell \|g_\ell\|_{(\ell),*}$, blockwise oracle $\mathrm{lmo}(g) = (\rho_\ell\ \mathrm{lmo}_\ell(g_\ell))_\ell$ of per-block radius $R\rho_\ell$, Assumption 1 imposed in this pair, and $\zeta^2 \le \sum_\ell \rho_\ell^2\zeta_\ell^2$. The minus sign is inside the oracle: the update is $x_{t+1} = x_t + \gamma D_t$ with $D_t = \mathrm{lmo}(G_t)$, which for the Euclidean ball is normalized gradient descent, $x - \gamma R g/\|g\|_2$ (Pethick et al., 2025).

MOO in a norm pair. Write $F(x) = (f_1(x), \dots, f_K(x))^\top$ and, for $w \in \Delta_K := \{w \ge 0,\ \sum_i w_i = 1\}$, $\nabla F(x)w := \sum_i w_i \nabla f_i(x)$ and $F(x)w := \sum_i w_i f_i(x)$.

Definition 3.1 (Stationarity, CA distance, inner gap)**.** x is *ϵ-accurate Pareto stationary* if $\min_{w\in\Delta_K}\|\nabla F(x)w\|_* \le \epsilon$; $w^\star(x)$ denotes a minimizer and $d^\star(x) := \nabla F(x)w^\star(x)$. For an iterate–weight pair (x_t, w_t), the *CA distance* is $\mathrm{CA}_t := \|\nabla F(x_t)w_t - d^\star(x_t)\|_*$ and the *inner gap* is $\mathrm{gap}_t := \|\nabla F(x_t)w_t\|_* - \min_{w\in\Delta_K}\|\nabla F(x_t)w\|_* \ge 0$. By the reverse triangle inequality $\mathrm{gap}_t \le \mathrm{CA}_t$ for every selection $w^\star(x_t)$; $\mathrm{gap}_t = 0$ iff w_t solves the inner problem; and, unlike CA_t, the gap involves no selection.

Two facts about these objects differ from the Euclidean intuition (proofs in Appendix B).

Proposition 3.1 (Uniqueness of the optimal blended gradient)**.** *If the dual unit ball is strictly convex, $d^\star(x)$ is the same for every minimizer $w^\star(x)$, for every instance. If it is not, there is an instance with $K = 2$ in which every $w \in \Delta_K$ is optimal and $d^\star$ ranges over a whole segment: take $\nabla f_1 = a$, $\nabla f_2 = b$ for distinct unit vectors a, b with $\|(a+b)/2\|_* = 1$ (for ℓ_1: $a = (1,0)$, $b = (0,1)$).*

Proposition 3.2 (Stationary points of fixed scalarizations are Pareto stationary)**.** *If $\bar{x}$ is a first-order stationary point of $x \mapsto F(x)u$ for a fixed $u \in \Delta_K$, i.e. $\nabla F(\bar{x})u = 0$, then $\min_{w\in\Delta_K}\|\nabla F(\bar{x})w\|_* = 0$, attained at $w = u$.*

Strict convexity holds for ℓ_p, $1 < p < \infty$, and fails for ℓ_1 and the nuclear norm, so in exactly the Muon and Signum geometries the CA distance can depend on the selected minimizer; our guarantees hold for any measurable selection or, better, for the selection-free gap. Throughout, $K \ge 2$. To distinguish Pareto stationarity from the residual of the weights actually used, write $\psi_t(w) := \|\nabla F(x_t)w\|_*$ and $S_T := T^{-1}\sum_t \psi_t(w_t)$. Directly from nonnegativity,

$$\frac{1}{T}\sum_t \mathrm{gap}_t \le S_T, \qquad \frac{1}{T}\Big[\sum_t \psi_t(w_t) - \min_{u\in\Delta_K}\sum_t \psi_t(u)\Big] \le S_T. \tag{3}$$

Algorithm 1 Entropic LMO-MGDA (with optional regularization $\tau \geq 0$ and clipping $\kappa \geq 0$)

Require: $x_0 \in \mathbb{R}^d$; step sizes $\gamma, \eta > 0$; radius R; regularization $\tau \geq 0$; clipping $\kappa \in [0, 1/K]$

1: $w_0 \leftarrow (1/K, \ldots, 1/K)$
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: compute $\nabla f_i(x_t)$ for $i \in [K]$ ▷ K backward passes
4: $G_t \leftarrow \sum_{i=1}^K w_{t,i} \nabla f_i(x_t)$; $D_t \leftarrow \mathrm{lmo}(G_t)$ ▷ one oracle call, layer-wise
5: $g_{t,i} \leftarrow \langle D_t, \nabla f_i(x_t) \rangle$ for $i \in [K]$
6: $w_{t+1} \leftarrow \arg\min_{w \in \Delta_K^\kappa} \Big\{ -\eta \langle g_t, w \rangle + \eta R \tau \, \mathrm{KL}(w \| u_K) + \mathrm{KL}(w \| w_t) \Big\}$ ▷ softmax then clip, (4)
7: $x_{t+1} \leftarrow x_t + \gamma D_t$
8: **end for**

Also $\mathrm{CA}_t \leq 2\psi_t(w_t)$ in every norm; in Hilbert dual geometry, $\mathrm{CA}_t^2 \leq \psi_t(w_t)^2 - \psi_t(w^\star(x_t))^2$ (Zhang et al., 2025). Thus convergence of the actual blended-gradient residual implies average-gap, average-CA, and static-regret upper bounds, including for a converged fixed scalarization. Smallness of $\min_w \psi_t(w)$ alone does not control a separately chosen w_t.

The gap is useful as a selection-independent inner optimality measure, but does not certify simultaneous per-task descent for every oracle selection. Appendix B gives the saddle condition and a sign-oracle counterexample. The additional results are a comparator-dependent regret bound valid independently of outer progress (Theorem 4.5) and an every-iterate tracking guarantee after burn-in (Theorem 5.1).

Assumptions.

Assumption 1 (Smoothness in the norm pair)**.** *Each f_i is differentiable and $\|\nabla f_i(x) - \nabla f_i(y)\|_* \leq L\|x - y\|$.*

Assumption 2 (Lower bound)**.** *$f_i^\star := \inf_x f_i(x) > -\infty$; $\Delta := \max_i \big(f_i(x_0) - f_i^\star\big) > 0$ and $L > 0$ (if $\Delta = 0$ or $L = 0$ every gradient vanishes at x_0 and there is nothing to prove; excluding these keeps $M > 0$ and the step sizes defined).*

Assumption 3 (Stochastic oracle)**.** *With $\mathcal{F}_t$ the history up to the computation of x_t, w_t, the sample ξ_t is independent of $\mathcal{F}_t$ and $\mathbb{E}[\nabla f_i(x_t; \xi_t) \mid \mathcal{F}_t] = \nabla f_i(x_t)$, $\mathbb{E}[\|\nabla f_i(x_t; \xi_t) - \nabla f_i(x_t)\|_2^2 \mid \mathcal{F}_t] \leq \sigma^2$ for every i (a shared minibatch across tasks is allowed).*

Assumption 4 (Gradient bound; stochastic setting only)**.** $\max_i \|\nabla f_i(x_t)\|_* \leq M$ *a.s. along the trajectory of Algorithm 2.*

Assumption 3 is in ℓ_2 because the variance recursion needs an inner product; we convert through $\zeta := \max_{x \neq 0} \|x\|_* / \|x\|_2$ (1 for ℓ_2, $\sqrt{d}$ for ℓ_1, $\sqrt{\min(m, n)}$ for the nuclear norm on $m \times n$ matrices), as in Pethick et al. (2025). In the deterministic analysis M is derived, not assumed (Lemma 4.3).

4 Entropic LMO-MGDA

Line 6 is entropic mirror descent on the scores g_t, with an optional pull of strength τ toward the uniform weighting u_K and an optional floor κ on every weight, $\Delta_K^\kappa := \{w \in \Delta_K : w_i \geq \kappa\}$. The sign is correct and counterintuitive: $-g_t/R$ is a subgradient of $w \mapsto \|\nabla F(x_t) w\|_*$ at w_t (Lemma A.3), so descending on it ascends in g_t: a task well aligned with D_t has $g_{t,i} \ll 0$ and loses weight, while a task being harmed gains it. The closed form is a softmax followed by clipping (Lemma F.1): with $\rho := \eta R \tau$,

$$w_{t+1,i} = \max\left\{\kappa, \; \frac{w_{t,i}^{1/(1+\rho)} \, e^{\eta g_{t,i}/(1+\rho)}}{Z_t}\right\}, \qquad Z_t \text{ such that } \textstyle\sum_i w_{t+1,i} = 1, \tag{4}$$

where Z_t and the set of clipped coordinates are found by a monotone recurrence in at most K rounds (Appendix F.1). At $\tau = \kappa = 0$ this is exponentiated gradient, $w_{t+1,i} \propto w_{t,i} e^{\eta g_{t,i}}$; the entropy regularizer shrinks the logits toward a uniform weighting, related to MOON's logit decay (Zhou et al., 2026). The method costs K backward passes, one oracle call and one softmax per iteration (for $\kappa > 0$, Appendix F.1 gives a clipping recurrence with $O(K^2)$ worst-case cost, or $O(K \log K)$

using sorting), stores the K task gradients within an iteration (the scores need D_t), and needs no inner optimization loop. Sections 4.1–4.2 analyse $\tau = \kappa = 0$; Section 5 shows that $\tau, \kappa > 0$ make the weights track the inner optimum at every iterate after a burn-in, without entering the stationarity inequality.

4.1 Deterministic analysis

Three lemmas, proved in Appendix C: the descent lemma in the norm pair turned into a one-step inequality by the oracle identity, the regularized mirror-descent bound, and a self-consistent bound on the gradients along the trajectory.

Lemma 4.1 (One-step inequality). *Under Assumption 1, for every $w \in \Delta_K$ and every t,*

$$F(x_{t+1})w \;\le\; F(x_t)w \;-\; \gamma R\,\|G_t\|_* \;+\; \gamma\langle g_t, w - w_t\rangle \;+\; \tfrac{1}{2}L\gamma^2R^2. \tag{5}$$

The descent $-\gamma R\|G_t\|_*$ is *linear* in the dual norm, not quadratic as for gradient descent; this is why our stationarity criterion is first-moment, as in Pethick et al. (2025).

Lemma 4.2 (Regularized entropic mirror descent). *Line 6 of Algorithm 1 satisfies, for every $u \in \Delta_K^\kappa$, with $\rho = \eta R\tau$ and $r(w) := \mathrm{KL}(w\|u_K)$,*

$$\eta\langle g_t, u - w_t\rangle \;\le\; \mathrm{KL}(u\|w_t) - (1+\rho)\,\mathrm{KL}(u\|w_{t+1}) + \rho\big(r(u) - r(w_{t+1})\big) + \tfrac{1}{2}\eta^2\|g_t\|_\infty^2. \tag{6}$$

For $u = u_K$ the regularization terms are ≤ 0; for general u they are $\le \rho \ln K$; at $\tau = \kappa = 0$ this is the standard bound for every $u \in \Delta_K$.

Lemma 4.3 (Trajectory gradient bound). *Under Assumptions 1–2, run Algorithm 1 with $\tau = \kappa = 0$, $\gamma R = \sqrt{2\Delta/(LT)}$ and $\eta R = \sqrt{2\log K}/(M\sqrt{T})$. If $M \ge 2\sqrt{L\Delta\log K} + 2\sqrt{L\Delta(\log K + 1)}$ then $\|\nabla f_i(x_t)\|_* \le M$ for all $t \le T$ and $i \in [K]$. In particular $M = O(\sqrt{L\Delta\log K})$, the larger root of an explicit quadratic, suffices, with no circularity.*

Theorem 4.4 (Deterministic rate). *Under Assumptions 1–2, run Algorithm 1 with any $\tau \ge 0$, $\kappa \in [0, 1/K]$, and suppose $\max_i \|\nabla f_i(x_t)\|_* \le M$ for $t < T$. Then for every $\gamma, \eta > 0$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)w_t\|_* \;\le\; \frac{\Delta}{\gamma RT} + \frac{\eta RM^2}{2} + \frac{L\gamma R}{2}, \tag{7}$$

in which neither τ, κ nor K appears. With the step sizes of Lemma 4.3,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla F(x_t)w_t\|_* \;\le\; \frac{\sqrt{2\Delta L} + M\sqrt{\log K/2}}{\sqrt{T}}, \tag{8}$$

so with $\hat{t} \in \arg\min_{t<T}\|\nabla F(x_t)w_t\|_*$ the output $x_{\hat{t}}$ is ϵ-accurate Pareto stationary, $\min_w \|\nabla F(x_{\hat{t}})w\|_* \le \|\nabla F(x_{\hat{t}})w_{\hat{t}}\|_* \le \epsilon$, after $T = O\big((\Delta L + M^2\log K)/\epsilon^2\big)$ iterations; for $\tau = \kappa = 0$, M is supplied by Lemma 4.3, and substituting $M^2 = O(L\Delta\log K)$ gives $T = O(L\Delta\log^2 K/\epsilon^2)$. No property of the primal norm beyond Assumption 1 is used.

The proof compares against u_K, for which $\mathrm{KL}(u_K\|w_0) = 0$ and the regularization only helps: given the same trajectory gradient bound M, τ, κ do not enter the inequality (they may change the trajectory and hence the smallest valid M), and $\log K$ enters (8) only through η, as the price of adaptivity rather than of stationarity.

Since $\mathrm{CA}_t \le 2\|\nabla F(x_t)w_t\|_*$ by the triangle inequality, the average CA distance obeys twice (8) for every selection (Corollary C.1); equally for a fixed weighting, so this is a consistency statement, not evidence of tracking. The same stationarity bound controls the average inner gap by (3). The next result gives a comparator-dependent guarantee without requiring outer progress.

Theorem 4.5 (Inner regret on bounded gradient sequences). *Let $A_t : \mathbb{R}^K \to \mathbb{R}^d$, $t < T$, be any sequence of linear maps with $\max_i \|A_t e_i\|_* \le M$, where $M > 0$. Start from $w_0 = u_K$ and set $D_t = \mathrm{lmo}(A_t w_t)$, $g_t = A_t^\top D_t$, and $w_{t+1,i} \propto w_{t,i}e^{\eta g_{t,i}}$ with $\eta R = \sqrt{2\log K}/(M\sqrt{T})$. Then, for every $u \in \Delta_K$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\|A_t w_t\|_* \;\le\; \frac{1}{T}\sum_{t=0}^{T-1}\|A_t u\|_* + M\sqrt{\frac{2\log K}{T}}. \tag{9}$$

In particular, this holds for $A_t = \nabla F(x_t)$ under the conditions of (8) *with $\tau = \kappa = 0$.*

The subgradient inequality gives $\|A_t w_t\|_* - \|A_t u\|_* \le \langle g_t, u - w_t\rangle / R$, and Lemma 4.2 sums the right-hand side. The proof is pathwise, allows adaptively chosen sequences, and needs neither smooth losses nor outer descent. The comparator is evaluated on the algorithm's own trajectory; on a prescribed sequence with constant columns $a, -a$, a fixed e_1 has regret $\|a\|_*$ against the uniform weighting, which says nothing about scalarization on its own trajectory.

Corollary 4.6 (Static regret to gap decomposition)**.** *Under the conditions of Theorem 4.5 with $A_t = \nabla F(x_t)$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathrm{gap}_t \;\le\; M\sqrt{\frac{2\log K}{T}} + V_T, \qquad V_T := \min_{u\in\Delta_K} \frac{1}{T}\sum_{t=0}^{T-1} \Big[\|\nabla F(x_t)u\|_* - \min_{w\in\Delta_K} \|\nabla F(x_t)w\|_*\Big], \tag{10}$$

where the comparator-variation term $V_T \ge 0$ is the shortfall of the best fixed *weighting against the per-iterate optimum, a property of the trajectory rather than of the update. This decomposition is informative when V_T is controlled. Along the trajectory covered by Theorem 4.4,* (3) *already gives average-gap convergence at the stationarity rate, without requiring $V_T \to 0$; one may take the smaller of the two bounds. In the Euclidean case $\mathrm{CA}_t^2 \le 2M\,\mathrm{gap}_t$, and under the uniform convexity inequality* (S2) *of Appendix G, $\mathrm{CA}_t^q \le (q/\mu)M^{q-1}\mathrm{gap}_t$ (Remark C.1).*

4.2 Stochastic analysis with momentum on the blended gradient

The oracle is nonlinear, so single-minibatch noise biases the direction at any fixed batch size. Averaging before the oracle call controls this error (Pethick et al., 2025); our analysis uses momentum with a horizon-dependent averaging parameter. Stochastic MGDA methods average each task gradient separately (Fernando et al., 2023; Xiao et al., 2023), storing K buffers; the alternative, which any implementation handing the MGDA direction to a momentum optimizer already uses, blends first:

$$\widehat{G}_t = \sum_{i=1}^{K} w_{t,i}\nabla f_i(x_t;\xi_{t,i}), \qquad \mathcal{M}_t = (1-\alpha)\mathcal{M}_{t-1} + \alpha\widehat{G}_t, \qquad D_t = \mathrm{lmo}(\mathcal{M}_{t-1}). \tag{11}$$

This is Algorithm 2 (Appendix D). The tracking error is one vector $\lambda_t := \mathcal{M}_t - \nabla F(x_t)w_t$ whose noise part is K-free by Jensen, $\mathbb{E}\|\sum_i w_{t,i}n_{t,i}\|_2^2 \le \sum_i w_{t,i}\mathbb{E}\|n_{t,i}\|_2^2 \le \sigma^2$ (the per-task bound in Appendix E carries $\sqrt{K}$). Evaluating the oracle at $\mathcal{M}_{t-1}$ makes D_t measurable before ξ_t, so the score-noise cross term has zero conditional mean. The buffer chases a target moving with w_t, which moves by *noisy* scores; the pathwise bound $\|w_{t+1} - w_t\|_1 \le 2\eta\|g_t\|_\infty$ would bring in a $\sqrt{K}$, and the next lemma avoids it.

Lemma 4.7 (Weight increment under noisy scores)**.** *Let w_{t+1} be obtained from w_t by line 6 of Algorithm 2. Pathwise, $\|w_{t+1} - w_t\|_1 \le 2\eta\|g_t\|_\infty$. If Assumptions 3–4 hold and $\eta RM \le \frac{1}{2}$,*

$$\mathbb{E}\big[\|w_{t+1} - w_t\|_1 \mid \mathcal{F}_t\big] \;\le\; \omega := 16\,\eta R\,(M + \zeta\sigma) + 8K\eta^2R^2\zeta^2\sigma^2.$$

K appears only at second order in η: a task can move the weights by a large amount only if its noisy score is large, which has probability $O(\eta^2)$.

Lemma 4.8 (Tracking error)**.** *Under Assumptions 1–4 and $\eta RM \le \frac{1}{2}$, splitting $\lambda_t = \nu_t + \delta_t$ into noise and drift, $\mathbb{E}\|\nu_t\|_2^2 \le \sigma^2(\alpha + (1-\alpha)^{2t})$ and $\mathbb{E}\|\delta_t\|_* \le (L\gamma R + M\omega)/\alpha$ (the x-part admits a deterministic bound since $\|x_t - x_{t-1}\| \le \gamma R$ pathwise), so $\mathbb{E}\|\lambda_t\|_* \le \zeta\sigma(\sqrt{\alpha} + (1-\alpha)^t) + (L\gamma R + M\omega)/\alpha$.*

Theorem 4.9 (Stochastic rate)**.** *Under Assumptions 1–4, run Algorithm 2 with $T \ge (2/M)^{4/3}$, $\alpha = T^{-1/2}$, $\gamma R = T^{-3/4}$ and $\eta R = T^{-3/4}/M^2$. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla F(x_t)w_t\|_* \;\le\; \frac{\Delta + 2L + 32 + 2\zeta\sigma(1 + 16/M)}{T^{1/4}} + \frac{C_2}{T^{1/2}} + \frac{C_3}{T^{3/4}} + \frac{C_4}{T}, \tag{12}$$

with $C_2 = 2\zeta\sigma$, $C_3 = 1 + \frac{5}{2}L + 32(1 + \zeta\sigma/M) + \zeta^2K\sigma^2/M^2$ and $C_4 = 2M + 32K\zeta^2\sigma^2/M^3$. Writing C_1 for the numerator of the first term, the right-hand side is at most ϵ once $T \ge \max\{(4C_1/\epsilon)^4, (4C_2/\epsilon)^2, (4C_3/\epsilon)^{4/3}, 4C_4/\epsilon\}$; the average CA distance obeys the same bound up

to a factor 2 *for every measurable selection, and for* $\hat{t}$ *uniform on* $\{0, \dots, T-1\}$ *drawn after the run,* $\mathbb{E} \min_w \|\nabla F(x_{\hat{t}})w\|_* \le \mathbb{E}\|\nabla F(x_{\hat{t}})w_{\hat{t}}\|_*$ *is bounded by* (12). C_1 *is free of* K*;* K *enters* C_3 *through the* ℓ_∞ *error term of mirror descent on noisy scores and* C_4 *through Lemma 4.7. The* K*-free requirement* $T \ge (4C_1/\epsilon)^4$ *determines the count whenever* $C_3 \le 16C_1^3/\epsilon^2$ *and* $C_4 \le 64C_1^4/\epsilon^3$*, i.e. for* $K \lesssim \epsilon^{-2}M^2/(\zeta^2\sigma^2)$ *as* $\epsilon \to 0$*.*

Here "K-free" means no explicit dependence with $L, M, \Delta, \zeta, \sigma$ fixed (these can change with task scaling and norm choices). By (3), (12) also bounds the expected average inner gap. Two corollaries in Appendix D give further comparator and noise-tail bounds. Corollary D.3 bounds the stochastic inner regret against any deterministic u by (12) with $\Delta \to M^2 \ln K$, hence the average expected gap up to $\bar{V}_T := \min_u \mathbb{E}\frac{1}{T}\sum_t[\|\nabla F(x_t)u\|_* - \min_w \|\nabla F(x_t)w\|_*] \ge \mathbb{E}V_T$. Corollary D.4 shows that under a sub-Gaussian tail bound the linear-K terms of C_3, C_4 become $\ln K$, so every constant is polylogarithmic in K.

The condition $T \ge (2/M)^{4/3}$ is $\eta RM \le \frac{1}{2}$. The gradient bound is assumed here because Lemma 4.3's induction controls f_i along a deterministic path, whereas the stochastic one-step inequality has terms controlled only in expectation (Remark D.1). The weight step is $\Theta(T^{-3/4}/M^2)$ rather than $\Theta(T^{-1/2}/M)$ because the latter leaves a residual $\Theta(M)$ in the drift; this costs a factor $M\sqrt{\log K}$ in the regret constant only (Appendix E compares with per-task tracking).

5 Iteration-wise inner gap

A per-iterate guarantee, which Zhang et al. (2025) obtain for the CA distance in the Euclidean case at $O(\epsilon^{-11})$ and $O(\epsilon^{-17})$, is stronger than an ergodic one; to track a *moving* target we make the inner problem strongly convex. Here $\tau > 0$ makes $\tilde{\psi}_t(w) := \|\nabla F(x_t)w\|_* + \tau\,\mathrm{KL}(w\|u_K)$ strongly convex with a unique minimizer p_t over Δ_K^κ, and $\kappa > 0$ keeps the divergence to that target finite; neither enters (7). The analysis uses only three norm-free facts: $-g_t/R$ is a subgradient of $w \mapsto \|\nabla F(x_t)w\|_*$, that function is M-Lipschitz in $\|\cdot\|_1$, and it moves by at most $L\|x - x'\|$ when x moves.

Theorem 5.1 (Iteration-wise inner gap, with progress)**.** *Let Assumptions 1–2 hold and suppose* $\max_i \|\nabla f_i(x_t)\|_* \le M$ *along the trajectory of Algorithm 1. Fix a tracking accuracy* $\epsilon \in (0, 1]$ *with* $\epsilon \le 2M$ *and a bias accuracy* $\beta \in (0, \epsilon]$*, and set*

$$\tau = \frac{\beta}{2\ln K}, \qquad \kappa = \frac{\beta}{4MK}, \qquad \eta R = \frac{\epsilon^2\beta}{768\,M^4 \ln K}, \qquad \gamma R = \frac{\epsilon^6\beta^6}{27 \cdot 2^{28}\,(2+\Lambda)\,L\,M^{11}K\ln^5 K},$$

with $\Lambda := \ln(4MK/\beta)$*. Then, for any primal norm:*

(i) ***Tracking.*** *For every* $t \ge T_0 := \lceil \frac{6144\,M^4\ln^2 K}{\epsilon^2\beta^2} \ln \frac{24M^2 \ln K}{\epsilon^2} \rceil$*:* $\|w_t - p_t\|_1 \le \epsilon/(2M)$ *and* $\mathrm{gap}(p_t) := \|\nabla F(x_t)p_t\|_* - \min_w \|\nabla F(x_t)w\|_* \le \beta$*, hence* $\mathrm{gap}_t \le \epsilon/2 + \beta$*. With* $\beta = \epsilon/2$*:* $\mathrm{gap}_t \le \epsilon$ *for all* $t \ge T_0 = \tilde{O}(M^4 \ln^2 K/\epsilon^4)$*.*

(ii) ***Progress.*** *For every* T*,* (7) *holds, and its right-hand side is at most* $\epsilon/2$ *for* $T \ge T_1 := \lceil 4\Delta/(\gamma R\,\epsilon)\rceil$*; with* $\beta = \epsilon/2$*,* $T_1 = \tilde{O}(\Delta L M^{11}K/\epsilon^{13})$*.*

(iii) ***Tail output.*** *For* $T \ge \max\{2T_0, T_1\}$ *and* $\hat{t} \in \arg\min_{T_0 \le t < T} \|\nabla F(x_t)w_t\|_*$*, the single iterate* $(x_{\hat{t}}, w_{\hat{t}})$ *satisfies both* $\|\nabla F(x_{\hat{t}})w_{\hat{t}}\|_* \le \epsilon$*, hence* $\min_w \|\nabla F(x_{\hat{t}})w\|_* \le \epsilon$*, and* $\mathrm{gap}_{\hat{t}} \le \epsilon/2 + \beta$*.*

If $\epsilon > 2M$ *all conclusions are trivial.*

Corollary 5.2 (Iteration-wise CA distance under uniform convexity)**.** *If the dual norm satisfies* (S2) *with power type* q *and constant* μ *(all* ℓ_p*, Schatten-*p*,* $1 < p < \infty$*; Proposition G.1), then* $\mathrm{CA}_t \le M\|w_t - p_t\|_1 + \big((q/\mu)M^{q-1}\mathrm{gap}(p_t)\big)^{1/q}$ *for every selection, so Theorem 5.1 with* $\beta = \mu\epsilon^q/(q2^qM^{q-1})$ *gives* $\mathrm{CA}_t \le \epsilon$ *for* $t \ge T_0$*; in the Euclidean case* $T_0 = \tilde{O}(M^6 \ln^2 K/\epsilon^6)$*,* $T_1 = \tilde{O}(\Delta L M^{17}K/\epsilon^{19})$*.*

The ϵ^{-13} bound concerns inner gap, whereas the ϵ^{-11} deterministic bound of Zhang et al. (2025) concerns Euclidean CA distance under different smoothness assumptions. For the latter metric our

Method	mIoU↑	Pix Acc↑	Abs Err↓	Rel Err↓	$\Delta m\% \downarrow$
STL	74.11	93.44	0.0137	38.22	0.00
LS	74.80	93.37	0.0188	42.41	11.83
SI	67.77	89.97	0.0151	34.65	3.29
RLW	74.24	93.30	0.0169	47.09	11.63
DWA	74.42	93.41	0.0159	41.94	6.35
UW	71.56	92.15	0.0191	38.75	11.41
MGDA	73.12	92.80	0.0157	36.03	2.72
PCGrad	75.89	93.72	0.0150	45.83	6.67
GradDrop	75.85	93.70	0.0151	43.60	5.42
CAGrad	75.83	93.50	0.0145	35.89	-0.66
MoCo	74.37	93.17	0.0221	81.06	43.33
MoDo	74.71	93.51	0.0152	44.31	6.50
Nash-MTL	74.76	93.29	0.0144	33.47	-2.01
FAMO	74.19	93.28	0.0154	**30.62**	-1.85
MGDA (warm)	75.64	93.60	0.0147	47.00	7.01
MOON	<u>78.53</u>	<u>94.15</u>	<u>0.0141</u>	<u>31.39</u>	**-5.42**
Entropic LMO-MGDA (ours)	**79.03**	**94.48**	**0.0136**	36.29	<u>-3.38</u>

Table 2: Results on CityScapes dataset. The best and second-best values are highlighted in **bold** and <u>underlined</u>, respectively. The task-specific metrics and the performance drop $\Delta m\%$ are reported.

conversion gives $\tilde{O}(\epsilon^{-19})$, so we claim no competitive exponent. The content is an iteration-wise inner-gap guarantee with simultaneous ergodic progress and one tail-selected iterate carrying both, in any norm, conditional on the gradient bound M (a hypothesis here, unlike in Theorem 4.4; Remark F.2). The tracked object is the regularized floored target p_t; the gap to the original inner problem comes through β.

Why we track the unsquared objective, unlike the Euclidean analyses, and what it costs, is explained in Remark F.1; proofs are in Appendix F.

6 Experiments

We evaluate on Cityscapes ($K = 2$: segmentation and depth) and NYU-v2 ($K = 3$: segmentation, depth, surface normals) with the SegNet/MTAN protocol of Navon et al. (2022); Liu et al. (2023), against single-task learning (STL), linear scalarization (LS), SI, RLW, DWA and UW weighting, MGDA (cold and warm-started), PCGrad, GradDrop, CAGrad, MoCo, MoDo, Nash-MTL, FAMO and MOON. Our method is Algorithm 2 with the spectral oracle on matrix-valued parameters. Tables 2 and 3 report per-task metrics and $\Delta m\%$, the mean relative degradation against the single-task (STL) row of each table (lower is better). The reported architectures, hyperparameters, and tuning settings are in Appendix I. The dense-prediction results use one seed, averaging the final ten epochs; they do not quantify variation across independent training runs. Our step sizes and schedule were tuned on the split on which results are reported, whereas the baselines use their reference hyperparameters (Appendix I.1).

On Cityscapes, our method achieves the highest segmentation scores and the lowest absolute depth error among the reported methods. MOON achieves a lower aggregate degradation, -5.42% versus -3.38%, driven by its lower relative depth error. On NYU-v2, our method achieves the lowest reported aggregate degradation, -6.80% versus -3.72% for MOON. These are descriptive comparisons from single runs. With $K = 2, 3$, they do not establish the predicted scaling advantage with the number of objectives, and the implementation differences do not isolate the contribution of geometry or weighting.

6.1 MultiMNIST: comparison with MOON and ablations

We follow MOON's MultiMNIST setup ($K = 2$: left and right digit, small ViT). Because its dataset instance is not released, we regenerate the data and rerun each method (Appendix I.2); only within-table comparisons are meaningful. The reported mean accuracy is 90.54% for our configuration and 87.47% for MOON. Hyperparameters for our method, including the weight step, are selected on a fixed validation split drawn from MNIST-train and frozen before evaluation on MNIST-test. Every value of the η and κ grids (Tables 5, 6) gives at least 89.5%, above all baselines. The spectral

	Segmentation		Depth		Surface normal					
Method	mIoU↑	Pix Acc↑	Abs↓	Rel↓	Mean↓	Med.↓	$<11.25\uparrow$	$<22.5\uparrow$	$<30\uparrow$	$\Delta m\%\downarrow$
STL	38.77	63.98	0.6350	0.2654	25.10	19.06	30.12	57.66	69.53	0.00
LS	38.13	64.83	0.5523	0.2334	27.79	23.56	23.03	48.23	61.64	6.76
SI	38.59	64.57	0.5439	0.2339	27.57	23.43	22.81	48.45	62.18	6.32
RLW	38.45	64.61	0.5814	0.2454	27.83	23.46	23.20	48.51	61.81	7.53
DWA	39.57	65.19	0.5691	0.2399	27.40	22.87	24.26	49.61	62.81	5.37
UW	38.48	63.88	0.5555	0.2294	27.47	23.38	23.18	48.60	62.29	6.23
MGDA	33.79	62.12	0.5906	0.2398	<u>24.47</u>	**18.65**	**31.28**	**58.50**	**70.28**	-1.33
PCGrad	38.79	65.08	0.5501	0.2212	27.18	22.74	24.10	49.93	63.36	4.23
GradDrop	40.37	66.04	0.5536	0.2353	27.97	23.76	22.78	47.97	61.39	6.39
CAGrad	<u>41.61</u>	66.95	0.5444	0.2251	25.75	20.79	27.44	53.99	66.85	-1.18
MoCo	35.47	61.76	0.5808	0.2328	31.21	27.97	15.31	39.38	54.01	18.38
MoDo	39.41	65.82	0.5353	0.2225	25.54	20.08	28.57	55.34	67.78	-1.96
Nash-MTL	38.18	65.39	0.5575	0.2325	25.12	19.85	29.13	55.97	68.48	-1.48
FAMO	32.57	61.67	0.5411	0.2220	25.16	19.25	30.18	57.16	69.10	-1.00
MGDA (warm)	38.41	63.76	0.5813	0.2453	28.13	24.03	21.43	47.24	61.13	9.15
MOON	39.88	<u>67.30</u>	**0.4951**	<u>0.2152</u>	25.50	20.02	29.44	55.34	67.71	<u>-3.72</u>
Entropic LMO-MGDA (ours)	**42.71**	**68.07**	<u>0.5075</u>	**0.2129**	**23.76**	<u>19.18</u>	<u>30.27</u>	<u>57.34</u>	<u>69.60</u>	**-6.80**

Table 3: Results on NYU-v2 dataset. The best and second-best values are highlighted in **bold** and <u>underlined</u>, respectively. The task-specific metrics and the performance drop $\Delta m\%$ are reported.

Method	Left↑	Right↑	Avg.↑
MGDA	83.63	85.13	84.38
MGDA + Muon	84.48	86.43	85.45
FAMO	85.24	87.29	86.27
FAMO + Muon	87.28	88.02	87.65
Muon (LS)	<u>87.45</u>	88.04	<u>87.75</u>
MOON	86.87	<u>88.06</u>	87.47
Entropic LMO-MGDA (Ours)	**89.86**	**91.23**	**90.54**

Component	Variant	Avg.↑	Gap↓
Oracle	ℓ_2	49.59	0.55
	sign	89.01	6.38
	spectral	90.54	0.90
Weights	entropic	90.54	0.90
	projected	90.51	0.97
Momentum	blended	90.54	0.90
	per-task	90.16	1.41
	none	81.68	1.65
Step η	10^{-5} to 10^{-3}	89.5–90.5	0.29–1.97

Table 4: MultiMNIST with a ViT backbone. Left: test accuracy (%), mean over three seeds on the same regenerated data. Best and second-best values are shown in bold and underlined. Right: fixed-hyperparameter substitutions with the final encoder-only inner gap, in each oracle's dual norm (not comparable across oracle rows). Details in Appendix I.2.

oracle is best under the fixed step sizes used in the ablation, removing momentum reduces accuracy by 8.86 points, and entropic and projected weights give similar means at $K = 2$; these are fixed-hyperparameter substitutions, not separately tuned comparisons.

7 Conclusion

We analyzed a single-loop entropic-LMO method for multi-objective optimization in arbitrary norm pairs: $O(\epsilon^{-2})$ deterministic stationarity with a derived gradient bound, $O(\epsilon^{-4})$ stochastic stationarity with one persistent buffer and no explicit K in the leading term, static regret of the weights on bounded gradient sequences, and, with regularization and clipping, iteration-wise inner-gap control after burn-in.

The inner gap avoids the minimizer-selection dependence of CA distance in spectral and sign geometries. It does not by itself certify simultaneous descent, and its average, like average static regret, is already controlled by the actual blended-gradient residual. The distinct tracking conclusion is the every-iterate guarantee after burn-in, not average-gap convergence alone.

Several limitations remain. Appendix H tolerates fixed relative oracle error for deterministic stationarity and average gap, but not yet for the stochastic or iteration-wise guarantees. The gradient bound is derived only in the deterministic unregularized case, and the tracking exponents are not competitive with Euclidean CA-distance results. Experiments have $K \leq 3$ and $\tau = \kappa = 0$, so they do not test many-objective scaling or regularized tracking.

AI USE STATEMENT

We used a large language model as an assistant for restructuring and editing the manuscript, checking and tightening proofs, drafting parts of the experiment code, and suggesting related work. The authors verified all such output. The research ideas, the method, and the main results are the authors' own, and the authors take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

All theoretical results are stated together with their assumptions (Assumptions 1–4), and complete proofs are in Appendices A–H: the structural results in Appendix B, the deterministic analysis in Appendix C, the stochastic analysis in Appendix D, the iteration-wise analysis in Appendix F, the uniform-convexity conversions in Appendix G and the inexact-oracle analysis in Appendix H; explicit constants are given throughout. The method is fully specified by Algorithms 1 and 2, the closed-form weight update (4) and the clipping recurrence of Algorithm 3. Appendix I describes the architectures, data construction, reported hyperparameters and tuning grids, seed counts, validation splits, reporting protocol, ablations, and empirical encoder-only gap. The $\Delta m\%$ values are computed from the per-task metrics with the reference values given there. An anonymous implementation of all methods and experiments is available at `https://anonymous.4open.science/r/LMO-MOO-D648/`.

REFERENCES

P. B. Assunção, O. P. Ferreira, and L. F. Prudente. Conditional gradient method for multiobjective optimization. *Computational Optimization and Applications*, 78:741–768, 2021.

Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones Mathematicae*, 115:463–482, 1994.

Hao Ban and Kaiyi Ji. Fair resource allocation in multi-task learning. In *International Conference on Machine Learning (ICML)*, 2024.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv:2409.20325*, 2024.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *International Conference on Machine Learning (ICML)*, 2025.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning (ICML)*, 2018.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic spectral descent for restricted Boltzmann machines. In *Artificial Intelligence and Statistics (AISTATS)*, 2015.

Lisha Chen, Heshan Fernando, Yiming Ying, and Tianyi Chen. Three-way trade-off in multi-objective learning: Optimization, generalization and conflict-avoidance. *Journal of Machine Learning Research*, 25:1–53, 2024.

Weiyu Chen, Baijiong Lin, Xiaoyuan Zhang, Xi Lin, Han Zhao, Qingfu Zhang, and James Kwok. Gradient-based multi-objective deep learning: Algorithms, theories, applications, and beyond. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=eCUcXXH3PS`.

Zhao Chen, Jiquan Ngiam, Yanping Huang, Thang Luong, Henrik Kretzschmar, Yuning Chai, and Dragomir Anguelov. Just pick a sign: Optimizing deep multitask models with gradient sign dropout. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5-6):313–318, 2012.

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 *International Conference on Learning Representations (ICLR)*, 2023.

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

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024.

Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. FAMO: Fast adaptive multitask optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, et al. Muon is scalable for LLM training. *arXiv:2502.16982 [cs.LG]*, 2025.

Suyun Liu and Luis Nunes Vicente. The stochastic multi-gradient algorithm for multi-objective optimization and its application to supervised machine learning. *Annals of Operations Research*, 2021.

Charles A. McCarthy. c_p. *Israel Journal of Mathematics*, 5:249–271, 1967.

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 (ICML)*, 2022.

Arkadi Nemirovsky and David Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley, 1983.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *International Conference on Machine Learning (ICML)*, 2025.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making Muon and Scion great again (bridging theory and practice of lmo-based optimizers for llms). *arXiv:2505.13416 [cs.LG]*, 2025.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Peiyao Xiao, Hao Ban, and Kaiyi Ji. Direction-oriented multi-objective learning: Simple and provable stochastic algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Greg Yang, Edward J. Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Greg Yang, James B. Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv:2310.17813 [cs.LG]*, 2023.

Linxi Yang, Liping Tang, Jiahao Lv, Yuehong He, and Xinmin Yang. Mirror descent method for stochastic multi-objective optimization. *arXiv:2410.06632*, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Qi Zhang, Peiyao Xiao, Shaofeng Zou, and Kaiyi Ji. MGDA converges under generalized smoothness, provably. In *International Conference on Learning Representations (ICLR)*, 2025.

Shiji Zhou, Wenpeng Zhang, Jiyan Jiang, Wenliang Zhong, Jinjie Gu, and Wenwu Zhu. On the convergence of stochastic multi-objective gradient manipulation and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shiji Zhou, Kunlin Lyu, Lei Zhang, Ruodong Wang, and Yifan Sun. MOON: Multi-objective orthonormalized updates for multitask learning. *arXiv:2608.11749*, 2026.

A PRELIMINARIES FOR THE PROOFS

Throughout, $\|\cdot\|$ is the primal norm on $\mathbb{R}^d$, $\|\cdot\|_*$ its dual, $B = \{v : \|v\| \le R\}$, and lmo is as in (2). We write $A := \nabla F(x)$ for the linear map $\mathbb{R}^K \to \mathbb{R}^d$, $Aw = \sum_i w_i \nabla f_i(x)$, and $A^\top$ for its adjoint, so $(A^\top d)_i = \langle \nabla f_i(x), d\rangle$. We record three elementary facts.

Lemma A.1 (Descent lemma in a norm pair)**.** *Under Assumption 1, for all $x, y \in \mathbb{R}^d$ and each i,*

$$f_i(y) \;\le\; f_i(x) + \langle \nabla f_i(x), y - x\rangle + \tfrac{L}{2}\|y - x\|^2.$$

Proof. By the fundamental theorem of calculus along the segment,

$$f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle = \int_0^1 \langle \nabla f_i(x + t(y - x)) - \nabla f_i(x), y - x\rangle \, dt.$$

Applying Hölder's inequality $\langle g, v\rangle \le \|g\|_*\|v\|$ and then Assumption 1 with $\|(x+t(y-x)) - x\| = t\|y - x\|$,

$$\le \int_0^1 \|\nabla f_i(x + t(y - x)) - \nabla f_i(x)\|_* \|y - x\| \, dt \le \int_0^1 Lt\|y - x\|^2 dt = \tfrac{L}{2}\|y - x\|^2. \qquad \square$$

Lemma A.2 (Gradient bound from the function value)**.** *Under Assumptions 1–2, $\|\nabla f_i(x)\|_*^2 \le 2L\big(f_i(x) - f_i^\star\big)$ for all x and i.*

Proof. If $\nabla f_i(x) = 0$ there is nothing to prove, so assume $\nabla f_i(x) \neq 0$. Since $\mathbb{R}^d$ is finite-dimensional the primal unit ball is compact, so the supremum defining $\|\nabla f_i(x)\|_*$ is attained: there is v with $\|v\| = 1$ and $\langle \nabla f_i(x), v\rangle = -\|\nabla f_i(x)\|_*$ (negate a maximizer). For $s > 0$, Lemma A.1 applied at $y = x + sv$ gives

$$f_i^\star \;\le\; f_i(x + sv) \;\le\; f_i(x) - s\,\|\nabla f_i(x)\|_* + \tfrac{L}{2}s^2.$$

Minimizing the right-hand side over $s > 0$ at $s = \|\nabla f_i(x)\|_*/L$ yields $f_i^\star \le f_i(x) - \|\nabla f_i(x)\|_*^2/(2L)$, which rearranges to the claim. $\square$

Lemma A.3 (Subgradient of the weight objective)**.** *Fix x and let $\psi(w) := \|Aw\|_*$ on $\mathbb{R}^K$. For every $d \in \mathrm{lmo}(Aw)$ we have $-A^\top d/R \in \partial\psi(w)$; componentwise, $-g/R \in \partial\psi(w)$ where $g_i = \langle d, \nabla f_i(x)\rangle$.*

Proof. Because $B = -B$ is symmetric, the dual norm is the support function $\|h\|_* = \frac{1}{R}\max_{v\in B}\langle h, v\rangle$, which is convex with $\partial\|\cdot\|_*(h) = \frac{1}{R}\arg\max_{v\in B}\langle h, v\rangle = -\frac{1}{R}\arg\min_{v\in B}\langle h, v\rangle = -\frac{1}{R}\,\mathrm{lmo}(h)$. The chain rule for the composition of a convex function with a linear map gives $A^\top\partial\|\cdot\|_*(Aw) \subseteq \partial\psi(w)$, whence the claim. $\square$

Lemma A.3 justifies the sign in line 6 of Algorithm 1: since $-g_t/R$ is a subgradient of ψ at w_t, a mirror-descent step *minimizing* ψ moves along $-\eta(-g_t/R)$, i.e. along $+\eta g_t$ after absorbing R into η. Note that the lemma holds for *every* element of $\mathrm{lmo}(Aw)$, so no selection rule is needed for our analysis.

B STRUCTURAL PROPERTIES OF THE CA DIRECTION

Proof of Proposition 3.1. The set $S = \{Aw : w \in \Delta_K\} = \mathrm{conv}\{\nabla f_1(x), \dots, \nabla f_K(x)\}$ is convex and compact. Let $m := \min_{h\in S}\|h\|_*$ and suppose $g_1, g_2 \in S$ both attain m with $g_1 \neq g_2$. If $m = 0$ then $g_1 = g_2 = 0$, a contradiction; so $m > 0$. Then g_1/m and g_2/m are distinct points of the dual unit sphere, and strict convexity of the dual unit ball gives $\|\frac{g_1+g_2}{2m}\|_* < 1$, i.e. $\|\frac{g_1+g_2}{2}\|_* < m$. But $\frac{g_1+g_2}{2} \in S$ by convexity, contradicting minimality of m. Hence the minimizing value $Aw^\star$ is unique.

For the converse, suppose the dual unit ball is not strictly convex: there are distinct a, b with $\|a\|_* = \|b\|_* = 1$ and $\|m\|_* = 1$ for $m := (a + b)/2$. Let φ be a norming functional at m, i.e. $\varphi(m) = 1$ and $|\varphi(y)| \le \|y\|_*$ for all y (it exists by compactness of the primal unit ball, since $\|m\|_* =$

$\max_{\|v\|\le 1}\langle m, v\rangle$). Then $\varphi(a) \le 1$, $\varphi(b) \le 1$ and $\frac{1}{2}(\varphi(a)+\varphi(b)) = \varphi(m) = 1$ force $\varphi(a) = \varphi(b) = 1$, so for every $\theta \in [0,1]$,

$$1 = \varphi\big(\theta a + (1-\theta)b\big) \le \|\theta a + (1-\theta)b\|_* \le \theta\|a\|_* + (1-\theta)\|b\|_* = 1,$$

and the whole segment $[a, b]$ lies on the unit sphere. Take $K = 2$ with $\nabla f_1(x) = a$, $\nabla f_2(x) = b$: then $\|Aw\|_* = 1$ for every $w \in \Delta_K$, every w is optimal, and $Aw^\star$ ranges over $[a, b]$. For $\|\cdot\|_* = \ell_1$ one may take $a = (1, 0)$, $b = (0, 1)$; for the nuclear norm, $a = E_{11}$, $b = E_{22}$. Hence strict convexity of the dual ball is necessary and sufficient for the optimal blended gradient to be unique in every instance. □

Proof of Proposition 3.2. The function $x \mapsto F(x)u = \sum_i u_i f_i(x)$ is differentiable with gradient $\nabla F(x)u$, so first-order stationarity is $\nabla F(\bar{x})u = 0$ (a minimizer is a special case), and then $\min_{w\in\Delta_K} \|\nabla F(\bar{x})w\|_* \le \|\nabla F(\bar{x})u\|_* = 0$, attained at $w = u$. □

We now record the selection criterion mentioned in Section 3. Let $\Phi_x(w, d) = \langle \nabla F(x)w, d\rangle$ on $\Delta_K \times B$. Since $\min_{d\in B} \Phi_x(w, d) = -R\|\nabla F(x)w\|_*$ by definition of the dual norm, Sion's minimax theorem (Sion, 1958) gives

$$\max_{w\in\Delta_K} \min_{d\in B} \Phi_x(w, d) = \min_{d\in B} \max_{w\in\Delta_K} \Phi_x(w, d) = -R \min_{w\in\Delta_K} \|\nabla F(x)w\|_*, \tag{13}$$

so the CA problem is $\max_w \min_d \Phi_x$, with w the maximizing and d the minimizing player.

Proposition B.1 (When the oracle output is conflict-avoidant)**.** *Fix x, let $w \in \Delta_K$, $G = \nabla F(x)w$, and let $D \in \mathrm{lmo}(G)$ with $g_i = \langle \nabla f_i(x), D\rangle$. Then (w, D) is a saddle point of Φ_x on $\Delta_K \times B$ if and only if*

$$\max_{i\in[K]} g_i \;=\; -R\,\|G\|_*. \tag{14}$$

When (14) *holds, every objective improves by at least $R\|G\|_*$ along D, which by* (13) *is the optimal guaranteed worst-case improvement over B. If $\|\cdot\|_*$ is differentiable at G then* $\mathrm{lmo}(G)$ *is a singleton and* (14) *holds automatically for any minimizer w of ψ; if not,* (14) *can fail for some elements of* $\mathrm{lmo}(G)$.

Proof. $D \in \mathrm{lmo}(G)$ is exactly the statement that D minimizes $\Phi_x(w, \cdot)$ over B. The remaining saddle condition is that w maximizes $\Phi_x(\cdot, D) = \langle g, \cdot\rangle$ over Δ_K, i.e. that w is supported on $\arg\max_i g_i$, i.e. that $\langle g, w\rangle = \max_i g_i$. Since $\langle g, w\rangle = \langle \nabla F(x)w, D\rangle = \langle G, D\rangle = -R\|G\|_*$ by (2), this is (14). Under (14) we get $-g_i \ge R\|G\|_*$ for every i, which is the stated improvement. If $\|\cdot\|_*$ is differentiable at G then $\partial\|\cdot\|_*(G)$ is a singleton, so by Lemma A.3 $\partial\psi(w) = \{-g/R\}$ and first-order optimality of w on Δ_K coincides with (14).

For the failure, take the instance in the proof of Proposition 3.1 with $w = e_1$, so $G = (1, 0)$ and $\mathrm{lmo}(G) = \{(-R, d_2) : |d_2| \le R\}$, every element of which satisfies $\langle G, v\rangle = -R = -R\|G\|_1$. Here $g = (-R, d_2)$ and (14) reads $\max(-R, d_2) = -R$, forcing $d_2 = -R$. The selection $d_2 = +R$ gives $g_2 = R > 0$: the second objective is actively degraded, while the choice $D = (-R, -R)$ improves both by R. □

The usual sign selection with $\mathrm{sign}(0) = 0$ can also harm a task at an inner minimizer. Take $R = 1$, primal ℓ_∞, gradients $a = (2, -1)$, $b = (-2, 3)$, and $w = (1/2, 1/2)$. Then $Aw = (0, 1)$ and $\|A(q, 1-q)\|_1 = |4q-2|+|3-4q| \ge 1$ for all $q \in [0, 1]$, with equality at $q = 1/2$. Thus the inner gap is zero, but $D = -\mathrm{sign}(Aw) = (0, -1)$ gives $\langle a, D\rangle = 1 > 0$. These gradients occur at the origin for the lower-bounded smooth quadratics $f_1(x) = \frac{1}{2}\|x\|_2^2 + \langle a, x\rangle$ and $f_2(x) = \frac{1}{2}\|x\|_2^2 + \langle b, x\rangle$. Inner optimality therefore does not alone certify per-task descent, even for a conventional oracle selection.

We stress that Theorem 4.4 and Theorem 4.9 do not require (14). The reason is that the mirror-descent bound (6) holds for an *arbitrary* vector g_t, and by Lemma A.3 every element of $\mathrm{lmo}(G_t)$ produces a valid subgradient; the only property of D_t used elsewhere is the identity $\langle G_t, D_t\rangle = -R\|G_t\|_*$, which likewise holds for every element. Selection therefore affects the per-iteration guarantee of Proposition B.1, not the average-case rates.

C PROOFS FOR THE DETERMINISTIC SETTING

Proof of Lemma 4.1. Fix $w \in \Delta_K$. Multiplying Lemma A.1 (with $y = x_{t+1}$, $x = x_t$) by $w_i \geq 0$ and summing over i, using $\sum_i w_i = 1$ for the quadratic term,

$$F(x_{t+1})w \;\leq\; F(x_t)w + \langle \nabla F(x_t)w, x_{t+1} - x_t\rangle + \tfrac{L}{2}\|x_{t+1} - x_t\|^2.$$

By line 7 of Algorithm 1, $x_{t+1} - x_t = \gamma D_t$ with $\|D_t\| \leq R$, so the last term is at most $\frac{1}{2}L\gamma^2R^2$. For the inner product, decompose $\nabla F(x_t)w = G_t + \nabla F(x_t)(w - w_t)$ and use (2):

$$\langle \nabla F(x_t)w, D_t\rangle = \langle G_t, D_t\rangle + \sum_i (w_i - w_{t,i})\langle \nabla f_i(x_t), D_t\rangle = -R\|G_t\|_* + \langle g_t, w - w_t\rangle. \quad \square$$

Proof of Lemma 4.2. Let $\Phi(w) = \sum_i w_i \ln w_i$ with $\nabla\Phi(w) = (1 + \ln w_i)_i$, whose Bregman divergence is $D_\Phi(u, v) = \mathrm{KL}(u\|v)$ for $u \in \Delta_K$ and v in the relative interior (with $0 \ln 0 = 0$). Write $C := \Delta_K^\kappa$ and

$$h(w) := -\eta\langle g_t, w\rangle + \rho\, r(w) + \mathrm{KL}(w\|w_t), \qquad \rho = \eta R\tau,$$

so that $w_{t+1} = \arg\min_C h$. Both $r = \mathrm{KL}(\cdot\|u_K)$ and $\mathrm{KL}(\cdot\|w_t)$ have Bregman divergence $\mathrm{KL}(u\|v)$ with respect to themselves, and the linear term contributes nothing, so h is $(1+\rho)$-strongly convex relative to Φ: for all $u \in \Delta_K$ and v in the relative interior,

$$h(u) = h(v) + \langle \nabla h(v), u - v\rangle + (1 + \rho)\,\mathrm{KL}(u\|v). \tag{15}$$

The function h is differentiable at w_{t+1}: if $\kappa > 0$, every point of C has all coordinates at least $\kappa > 0$, so h is differentiable on all of C, including at points with some $w_i = \kappa$ (where the clipped solution may well lie); if $\kappa = 0$, the minimizer has all coordinates positive because $\partial_i \mathrm{KL}(w\|w_t) = \ln(w_i/w_{t,i}) + 1 \to -\infty$ as $w_i \downarrow 0$ while w_t has positive coordinates. Hence first-order optimality over the convex set C gives $\langle \nabla h(w_{t+1}), u - w_{t+1}\rangle \geq 0$ for all $u \in C$. Hence, by (15) at $v = w_{t+1}$,

$$h(u) \geq h(w_{t+1}) + (1 + \rho)\,\mathrm{KL}(u\|w_{t+1}) \qquad (u \in C).$$

Expanding h on both sides and rearranging,

$$\eta\langle g_t, u - w_{t+1}\rangle \leq \mathrm{KL}(u\|w_t) - (1+\rho)\,\mathrm{KL}(u\|w_{t+1}) - \mathrm{KL}(w_{t+1}\|w_t) + \rho\big(r(u) - r(w_{t+1})\big), \tag{16}$$

with equality when $\kappa = 0$, since then $C = \Delta_K$ and $\nabla h(w_{t+1})$ is a multiple of $\mathbf{1}$, orthogonal to $u - w_{t+1}$. Next, split $\langle g_t, u - w_t\rangle = \langle g_t, u - w_{t+1}\rangle + \langle g_t, w_{t+1} - w_t\rangle$ and bound the second term by Hölder, Young's inequality $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ and Pinsker's inequality $\|p - q\|_1^2 \leq 2\,\mathrm{KL}(p\|q)$:

$$\eta\langle g_t, w_{t+1} - w_t\rangle \leq \eta\|g_t\|_\infty\|w_{t+1} - w_t\|_1 \leq \tfrac{1}{2}\eta^2\|g_t\|_\infty^2 + \tfrac{1}{2}\|w_{t+1} - w_t\|_1^2 \leq \tfrac{1}{2}\eta^2\|g_t\|_\infty^2 + \mathrm{KL}(w_{t+1}\|w_t). \tag{17}$$

Adding (16) and (17), the terms $\pm\,\mathrm{KL}(w_{t+1}\|w_t)$ cancel and we obtain (6). The claims about $u = u_K$ and general u follow from $r(u_K) = 0$, $r \geq 0$ and $r \leq \ln K$ on Δ_K. At $\tau = \kappa = 0$ we have $\rho = 0$ and $C = \Delta_K$, and (16) is the three-point identity $\eta\langle g_t, u - w_{t+1}\rangle = \mathrm{KL}(u\|w_t) - \mathrm{KL}(u\|w_{t+1}) - \mathrm{KL}(w_{t+1}\|w_t)$ of exponentiated gradient; the iterates then stay in the relative interior since w_0 is uniform and the update multiplies by positive factors, so (6) holds for every $u \in \Delta_K$, including the vertices $u = e_i$ used in Lemma 4.3. $\square$

Proof of Lemma 4.3. Set $\bar\gamma := \gamma R = \sqrt{2\Delta/(LT)}$ and $\bar\eta := \eta R = \sqrt{2\ln K}/(M\sqrt{T})$, and define

$$\mathcal{C} \;:=\; \frac{\bar\gamma \ln K}{\bar\eta} \;+\; \frac{T\bar\gamma\bar\eta M^2}{2} \;+\; \frac{TL\bar\gamma^2}{2}.$$

Substituting the step sizes,

$$\frac{\bar\gamma \ln K}{\bar\eta} = M\sqrt{\tfrac{\Delta \ln K}{L}}, \qquad \frac{T\bar\gamma\bar\eta M^2}{2} = M\sqrt{\tfrac{\Delta \ln K}{L}}, \qquad \frac{TL\bar\gamma^2}{2} = \Delta, \tag{18}$$

so $\mathcal{C} = 2M\sqrt{\Delta \ln K/L} + \Delta$.

We prove by induction on t that $\|\nabla f_i(x_t)\|_* \leq M$ for all i.

Base case. By Assumption 2 and Lemma A.2, $\|\nabla f_i(x_0)\|_* \leq \sqrt{2L\Delta}$. Since $K \geq 2$ we have $4\ln K \geq 4\ln 2 > 2$, so $2\sqrt{L\Delta \ln K} \geq \sqrt{2L\Delta}$ and the hypothesis on M gives $\|\nabla f_i(x_0)\|_* \leq M$.

Inductive step. Suppose $\|\nabla f_i(x_s)\|_* \le M$ for all i and all $s \le t$. Then $\|g_s\|_\infty \le \max_i \|D_s\| \|\nabla f_i(x_s)\|_* \le RM$ for $s \le t$ by Hölder. Apply Lemma 4.1 with $w = e_i$ and add γ/η times (6) with $u = e_i$; summing over $s = 0, \dots, t$ and using $\mathrm{KL}(e_i\|w_0) = \ln K$, $\mathrm{KL}(e_i\|w_{t+1}) \ge 0$ and $t+1 \le T$,

$$f_i(x_{t+1}) - f_i(x_0) \;\le\; -\bar\gamma \sum_{s=0}^{t} \|G_s\|_* + \frac{\bar\gamma \ln K}{\bar\eta} + \frac{T\bar\gamma\bar\eta M^2}{2} + \frac{TL\bar\gamma^2}{2} \;\le\; \mathcal{C},$$

where we dropped the non-positive first sum. Hence $f_i(x_{t+1}) - f_i^\star \le \Delta + \mathcal{C} = 2\Delta + 2M\sqrt{\Delta \ln K / L}$, and Lemma A.2 gives

$$\|\nabla f_i(x_{t+1})\|_*^2 \;\le\; 2L\Big(2\Delta + 2M\sqrt{\tfrac{\Delta \ln K}{L}}\Big) \;=\; 4L\Delta + 4M\sqrt{L\Delta \ln K}.$$

It therefore suffices that $M^2 \ge 4L\Delta + 4M\sqrt{L\Delta \ln K}$, i.e. that M be at least the positive root of $M^2 - 4\sqrt{L\Delta \ln K}\, M - 4L\Delta = 0$, namely

$$M \;\ge\; 2\sqrt{L\Delta \ln K} + \sqrt{4L\Delta \ln K + 4L\Delta} \;=\; 2\sqrt{L\Delta \ln K} + 2\sqrt{L\Delta(\ln K + 1)},$$

which is the hypothesis. This closes the induction and shows $M = O(\sqrt{L\Delta \ln K})$ suffices. Note that M is defined by an explicit formula in L, Δ, K only, so the step sizes are well defined and there is no circularity. □

Proof of Theorem 4.4. Take the uniform comparator $u = u_K$, which lies in Δ_K^κ for every $\kappa \le 1/K$. Adding Lemma 4.1 with $w = u_K$ to γ/η times (6) with $u = u_K$ gives, for each t, using $r(u_K) = 0$, $r(w_{t+1}) \ge 0$ and $(1+\rho)\,\mathrm{KL} \ge \mathrm{KL}$,

$$F(x_{t+1})u_K \le F(x_t)u_K - \gamma R\|G_t\|_* + \frac{\gamma}{\eta}\Big(\mathrm{KL}(u_K\|w_t) - \mathrm{KL}(u_K\|w_{t+1})\Big) + \frac{\gamma\eta}{2}\|g_t\|_\infty^2 + \frac{L\gamma^2R^2}{2}.$$

By hypothesis (or by Lemma 4.3 when $\tau = \kappa = 0$), $\|g_t\|_\infty \le RM$ for all $t < T$. Summing over $t = 0, \dots, T-1$, telescoping the divergences, and using $\mathrm{KL}(u_K\|w_0) = 0$ (as $w_0 = u_K$) and $\mathrm{KL}(u_K\|w_T) \ge 0$,

$$F(x_T)u_K - F(x_0)u_K \le -\gamma R \sum_{t=0}^{T-1} \|G_t\|_* + \frac{T\gamma\eta R^2M^2}{2} + \frac{TL\gamma^2R^2}{2}.$$

Moreover $F(x_0)u_K - F(x_T)u_K \le \frac{1}{K}\sum_i (f_i(x_0) - f_i^\star) \le \Delta$ by Assumption 2. Rearranging and dividing by γRT gives (7); neither τ, κ nor K has entered. With $\bar\gamma = \gamma R = \sqrt{2\Delta/(LT)}$ the first and third terms of (7) are each $\sqrt{\Delta L/(2T)}$, summing to $\sqrt{2\Delta L/T}$; with $\bar\eta = \eta R = \sqrt{2\ln K}/(M\sqrt{T})$ the second is $M\sqrt{\ln K/(2T)}$. Adding gives (8). The right-hand side of (8) is at most ϵ once $T \ge (\sqrt{2\Delta L} + M\sqrt{\ln K/2})^2/\epsilon^2$, and $(\sqrt{2\Delta L} + M\sqrt{\ln K/2})^2 \le 4\Delta L + M^2 \ln K$, so $T = \lceil (4\Delta L + M^2 \ln K)/\epsilon^2 \rceil$ iterations suffice. Every step used only Assumptions 1–2 and the identity (2); no property of $\|\cdot\|$ beyond being a norm was invoked. □

Corollary C.1 (Average CA distance)**.** *Under the conditions of* (8)*,* $\frac{1}{T}\sum_t \mathrm{CA}_t \le 2\big(\sqrt{2\Delta L} + M\sqrt{\log K/2}\big)/\sqrt{T}$ *for every measurable selection* $w^\star(x_t)$*.*

Proof of Corollary C.1. Let $w^\star(x_t)$ be any minimizer of $w \mapsto \|\nabla F(x_t)w\|_*$ over Δ_K. By the triangle inequality and then optimality,

$$\mathrm{CA}_t = \|\nabla F(x_t)w_t - \nabla F(x_t)w^\star(x_t)\|_* \le \|\nabla F(x_t)w_t\|_* + \|\nabla F(x_t)w^\star(x_t)\|_* \le 2\|\nabla F(x_t)w_t\|_*.$$

Averaging over t and applying Theorem 4.4 gives the claim. Only the triangle inequality and the optimality of $w^\star$ were used, so the bound is valid simultaneously for every measurable selection; in particular it does not presuppose the uniqueness of $\nabla F(x_t)w^\star(x_t)$ discussed in Proposition 3.1. □

Remark C.1. In a Hilbert dual norm the constant 2 can be removed: the first-order optimality condition for $w^\star$ gives $\langle \nabla F w^\star, \nabla F w_t - \nabla F w^\star\rangle \ge 0$, whence $\|\nabla F w_t - \nabla F w^\star\|_*^2 \le \|\nabla F w_t\|_*^2 - \|\nabla F w^\star\|_*^2 \le \|\nabla F w_t\|_*^2$. This is the argument used in the Euclidean literature. It

relies on the parallelogram identity and is unavailable for a general dual norm, which is why we state Corollary C.1 with the triangle inequality instead. The same identity gives the conversion used in Corollary 4.6: writing $a = \|\nabla F w_t\|_*$ and $b = \|\nabla F w^\star\|_*$,

$$\mathrm{CA}_t^2 \le a^2 - b^2 = (a+b)(a-b) \le 2M\,\mathrm{gap}_t,$$

since $a, b \le M$ along the trajectory (Lemma 4.3) and $a - b = \mathrm{gap}_t$. More generally, if the dual norm satisfies (S2) with power type q, then with $w^\star$ minimizing $\varphi_q(\nabla F w)$ over Δ_K (the same minimizers as $\|\nabla F w\|_*$), first-order optimality and (S2) give $\frac{\mu}{q}\mathrm{CA}_t^q \le \varphi_q(\nabla F w_t) - \varphi_q(\nabla F w^\star) = \frac{1}{q}(a^q - b^q) \le a^{q-1}(a-b) \le M^{q-1}\mathrm{gap}_t$, i.e. $\mathrm{CA}_t^q \le (q/\mu)M^{q-1}\mathrm{gap}_t$. The direct inequalities $\mathrm{gap}_t \le \|\nabla F(x_t)w_t\|_*$ and $\mathrm{CA}_t \le 2\|\nabla F(x_t)w_t\|_*$ imply average-gap and average-CA convergence whenever the actual blended-gradient residual converges in average. This includes a fixed scalarization on its own convergent trajectory. Corollary 4.6 gives an additional comparator-dependent decomposition; it is not needed to infer average-gap convergence from Theorem 4.4.

Proof of Theorem 4.5. Fix $u \in \Delta_K$ and write $\psi_t(w) := \|A_t w\|_*$. The support-function argument of Lemma A.3 applies to any linear map A_t, with $d = D_t \in \mathrm{lmo}(G_t)$ and $G_t = A_t w_t$. Thus $-g_t/R$ is a subgradient of ψ_t at w_t. The subgradient inequality $\psi_t(u) \ge \psi_t(w_t) + \langle -g_t/R, u - w_t\rangle$ rearranges to

$$\psi_t(w_t) - \psi_t(u) \le \frac{1}{R}\langle g_t, u - w_t\rangle \qquad \text{for every } t. \tag{19}$$

Summing (6) at $\tau = \kappa = 0$ over $t = 0, \dots, T-1$, telescoping and dropping $-\mathrm{KL}(u\|w_T) \le 0$,

$$\eta \sum_{t=0}^{T-1} \langle g_t, u - w_t\rangle \le \mathrm{KL}(u\|w_0) + \frac{\eta^2}{2}\sum_{t=0}^{T-1}\|g_t\|_\infty^2 \le \ln K + \frac{\eta^2 T R^2 M^2}{2},$$

using $\mathrm{KL}(u\|w_0) \le \ln K$ for uniform w_0 and $\|g_t\|_\infty \le RM$ from the column bound $\max_i \|A_t e_i\|_* \le M$ and Hölder's inequality. Combining with (19) and dividing by RT,

$$\frac{1}{T}\sum_{t=0}^{T-1}\big[\psi_t(w_t) - \psi_t(u)\big] \le \frac{\ln K}{\eta RT} + \frac{\eta R M^2}{2} = M\sqrt{\frac{2\ln K}{T}},$$

the last equality by the choice $\eta R = \sqrt{2\ln K}/(M\sqrt{T})$, as in the proof of Theorem 4.4. This is (9). Since it holds for every $u \in \Delta_K$, it holds in particular for a minimizer of $u \mapsto \frac{1}{T}\sum_t \psi_t(u)$ (which exists: each ψ_t is continuous and Δ_K is compact), the best fixed weighting in hindsight. The argument used Lemma A.3, Lemma 4.2 and the bound $\|g_t\|_\infty \le RM$ only. Hence it holds for arbitrary bounded matrix sequences, including sequences depending on preceding iterates; no outer progress is needed. □

Proof of Corollary 4.6. Write $\psi_t(w) = \|\nabla F(x_t)w\|_*$. For any $u \in \Delta_K$ and any t, by definition of the gap,

$$\mathrm{gap}_t = \big[\psi_t(w_t) - \psi_t(u)\big] + \Big[\psi_t(u) - \min_{w\in\Delta_K}\psi_t(w)\Big].$$

Average over t and apply (9) to the first bracket:

$$\frac{1}{T}\sum_t \mathrm{gap}_t \le M\sqrt{\frac{2\ln K}{T}} + \frac{1}{T}\sum_t\Big[\psi_t(u) - \min_w \psi_t(w)\Big] \qquad \text{for every } u \in \Delta_K.$$

The second sum is a continuous function of u on the compact set Δ_K and is non-negative termwise; minimizing over u gives V_T and hence (10). If some $\bar u$ minimizes ψ_t for every t then each bracket vanishes at $u = \bar u$ and $V_T = 0$. The Euclidean conversion $\mathrm{CA}_t^2 \le 2M\,\mathrm{gap}_t$ is the remark preceding this proof; averaging and using the concavity of $\sqrt{\cdot}$ (Jensen), $\frac{1}{T}\sum_t \mathrm{CA}_t \le \frac{1}{T}\sum_t\sqrt{2M\,\mathrm{gap}_t} \le \sqrt{2M\cdot\frac{1}{T}\sum_t \mathrm{gap}_t}$. □

D Proofs for the stochastic setting

Only *one* vector of dimension d is carried between iterations. The convention $\alpha_0 = 1$ removes the initialization bias of the moving average. At $t = 0$ the oracle is evaluated at $\mathcal{M}_{-1} = 0$; with the convention $\mathrm{lmo}(0) = 0$ of Section 3 the first step is $D_0 = 0$, and the argument below in fact allows any selection in B.

Algorithm 2 Entropic LMO-MGDA with momentum on the blended gradient ($\tau = \kappa = 0$)

Require: x_0; step sizes $\gamma, \eta > 0$; momentum $\alpha \in (0, 1]$; radius R; horizon T

1: $w_0 \leftarrow (1/K, \dots, 1/K)$; $\mathcal{M}_{-1} \leftarrow 0$
2: **for** $t = 0, 1, \dots, T-1$ **do**
3: $D_t \leftarrow \mathrm{lmo}(\mathcal{M}_{t-1})$ ▷ oracle at the *previous* buffer: D_t is $\mathcal{F}_t$-measurable
4: sample $\xi_{t,i}$ and compute $\nabla f_i(x_t; \xi_{t,i})$ for $i \in [K]$
5: $g_{t,i} \leftarrow \langle D_t, \nabla f_i(x_t; \xi_{t,i}) \rangle$ for $i \in [K]$
6: $w_{t+1,i} \leftarrow w_{t,i} e^{\eta g_{t,i}} / \sum_j w_{t,j} e^{\eta g_{t,j}}$ ▷ line 6 of Algorithm 1 at $\tau = \kappa = 0$
7: $\widehat{G}_t \leftarrow \sum_i w_{t,i} \nabla f_i(x_t; \xi_{t,i})$; $\mathcal{M}_t \leftarrow (1 - \alpha_t)\mathcal{M}_{t-1} + \alpha_t \widehat{G}_t$, $\alpha_0 = 1$, $\alpha_t = \alpha$ for $t \geq 1$
8: $x_{t+1} \leftarrow x_t + \gamma D_t$
9: **end for**

Filtration and notation. Let $\mathcal{F}_t$ be the σ-algebra generated by $\xi_{0,\cdot}, \dots, \xi_{t-1,\cdot}$ (and x_0), so that x_t, w_t, $\mathcal{M}_{t-1}$ and hence D_t are $\mathcal{F}_t$-measurable while the $\xi_{t,i}$ are not. Write $A_t := \nabla F(x_t)$ and

$$\lambda_t := \mathcal{M}_t - A_t w_t, \qquad n_t := \widehat{G}_t - A_t w_t = \sum_i w_{t,i} n_{t,i}, \qquad n_{t,i} := \nabla f_i(x_t; \xi_{t,i}) - \nabla f_i(x_t),$$

together with the conventions $\lambda_{-1} := \mathcal{M}_{-1} - A_0 w_0 = -A_0 w_0$ and $\Theta_0 := 0$, where for $t \geq 1$

$$\Theta_t := A_{t-1} w_{t-1} - A_t w_t.$$

By Assumption 3, $\mathbb{E}[n_{t,i} \mid \mathcal{F}_t] = 0$ and $\mathbb{E}[\|n_{t,i}\|_2^2 \mid \mathcal{F}_t] \leq \sigma^2$; consequently $\mathbb{E}[n_t \mid \mathcal{F}_t] = 0$. Since $w_t \in \Delta_K$ is $\mathcal{F}_t$-measurable, the triangle inequality and Jensen's inequality for the convex map $s \mapsto s^2$ with weights w_t give

$$\mathbb{E}\big[\|n_t\|_2^2 \mid \mathcal{F}_t\big] \leq \mathbb{E}\Big[\Big(\sum_i w_{t,i} \|n_{t,i}\|_2\Big)^2 \Big| \mathcal{F}_t\Big] \leq \sum_i w_{t,i}\, \mathbb{E}\big[\|n_{t,i}\|_2^2 \mid \mathcal{F}_t\big] \leq \sigma^2. \tag{20}$$

This bound is free of K and, notably, uses no independence across tasks: a shared minibatch for all K objectives is allowed. Throughout this appendix we work under Assumption 4, so that $\max_i \|\nabla f_i(x_t)\|_* \leq M$ for every t, and we write $\zeta = \max_{x \neq 0} \|x\|_* / \|x\|_2$ as in Section 3.

D.1 The weight increment

The tracking analysis needs to control how far the weights move in one step. The pathwise bound (a) below is what the deterministic analysis would use; it involves $\|g_t\|_\infty$, a maximum over tasks, which in the stochastic setting contains $\max_i \|n_{t,i}\|_*$ and would introduce a $\sqrt{K}$ through a maximal inequality. Part (b) shows that *in expectation* the increment is governed by the w_t*-weighted average* of the noise instead, with the maximum over tasks confined to a term of second order in η.

Lemma D.1 (Weight increment; restatement of Lemma 4.7)**.** *Let $w \in \Delta_K$, $g \in \mathbb{R}^K$, and $w_i' = w_i e^{\eta g_i} / \sum_j w_j e^{\eta g_j}$.*

(a) Pathwise, $\|w' - w\|_1 \leq 2\eta \|g\|_\infty$.

(b) Suppose w, x and D with $\|D\| \leq R$ are $\mathcal{F}$-measurable, $g_i = \langle D, \nabla f_i(x) \rangle + \langle D, n_i \rangle$ with $\max_i \|\nabla f_i(x)\|_* \leq M$, $\mathbb{E}[n_i \mid \mathcal{F}] = 0$ and $\mathbb{E}[\|n_i\|_2^2 \mid \mathcal{F}] \leq \sigma^2$ for each i (no independence across i required). If $\eta R M \leq \frac{1}{2}$, then

$$\mathbb{E}\big[\|w' - w\|_1 \mid \mathcal{F}\big] \;\leq\; \omega \;:=\; 16\,\eta R\,(M + \zeta\sigma) \;+\; 8K\eta^2 R^2 \zeta^2 \sigma^2. \tag{21}$$

Proof. (a) Take $u = w$ in (16), which at $\tau = \kappa = 0$ is the three-point identity of exponentiated gradient (with $w_t = w$, $w_{t+1} = w'$, $g_t = g$). Since $\mathrm{KL}(w\|w) = 0$ it reads $\eta\langle g, w - w' \rangle = -\mathrm{KL}(w\|w') - \mathrm{KL}(w'\|w)$, i.e.

$$\mathrm{KL}(w'\|w) + \mathrm{KL}(w\|w') = \eta\langle g, w' - w \rangle \leq \eta \|g\|_\infty \|w' - w\|_1.$$

Both divergences are non-negative, so $\mathrm{KL}(w'\|w) \leq \eta\|g\|_\infty \|w' - w\|_1$. Pinsker's inequality $\frac{1}{2}\|w' - w\|_1^2 \leq \mathrm{KL}(w'\|w)$ gives $\frac{1}{2}\|w' - w\|_1^2 \leq \eta\|g\|_\infty \|w' - w\|_1$; if $w' \neq w$ divide by $\|w' - w\|_1$, and otherwise the claim is trivial.

(b) *Step 1: a weighted-average bound.* Put $b_i := \eta g_i$, $a_i := e^{b_i}$, $Z := \sum_j w_j a_j$ and $\bar{b} := \sum_j w_j b_j$. Then $w_i' - w_i = w_i(a_i - Z)/Z$, so $\|w' - w\|_1 = Z^{-1} \sum_i w_i |a_i - Z|$. For any constant c, since $\sum_i w_i(a_i - c) = Z - c$,

$$\sum_i w_i |a_i - Z| \le \sum_i w_i |a_i - c| + |c - Z| = \sum_i w_i |a_i - c| + \Big| \sum_i w_i (c - a_i) \Big| \le 2 \sum_i w_i |a_i - c|.$$

Take $c = e^{\bar{b}}$. By Jensen's inequality for exp, $Z = \sum_j w_j e^{b_j} \ge e^{\bar{b}}$, hence $e^{\bar{b}}/Z \le 1$ and

$$\|w' - w\|_1 \;\le\; \frac{2e^{\bar{b}}}{Z} \sum_i w_i \big| e^{b_i - \bar{b}} - 1 \big| \;\le\; 2 \sum_i w_i \big| e^{y_i} - 1 \big|, \qquad y_i := b_i - \bar{b}. \tag{22}$$

Step 2: splitting the scores. Write $b_i = d_i + \nu_i$ with $d_i := \eta \langle D, \nabla f_i(x) \rangle$ and $\nu_i := \eta \langle D, n_i \rangle$. By Hölder, $|d_i| \le \eta R M$ and $|\nu_i| \le \eta R \|n_i\|_* \le \eta R \zeta \|n_i\|_2$. Hence

$$|y_i| \le |d_i - \bar{d}| + |\nu_i - \bar{\nu}| \le 2\eta R M + |\nu_i| + \sum_j w_j |\nu_j|,$$

and consequently $\sum_i w_i |y_i| \le 2\eta R M + 2 \sum_i w_i |\nu_i|$.

Step 3: the good event. Let $\mathcal{A} := \{\max_j |\nu_j| \le \frac{1}{2}\}$. On $\mathcal{A}$, using $2\eta R M \le 1$, every $|y_i| \le 1 + \frac{1}{2} + \frac{1}{2} = 2$. Since $y \mapsto (e^y - 1)/y$ is increasing on $(0, 2]$ with value $(e^2 - 1)/2 < 4$ at $y = 2$, we have $|e^y - 1| \le e^{|y|} - 1 \le 4|y|$ for $|y| \le 2$, and (22) gives, on $\mathcal{A}$,

$$\|w' - w\|_1 \le 8 \sum_i w_i |y_i| \le 16 \eta R M + 16 \sum_i w_i |\nu_i|.$$

Off $\mathcal{A}$ we use the trivial bound $\|w' - w\|_1 \le 2$, together with the union bound and Chebyshev's inequality:

$$\Pr(\mathcal{A}^c \mid \mathcal{F}) \le \sum_j \Pr\big(|\nu_j| > \tfrac{1}{2} \,\big|\, \mathcal{F}\big) \le 4 \sum_j \mathbb{E}[\nu_j^2 \mid \mathcal{F}] \le 4\eta^2 R^2 \zeta^2 \sum_j \mathbb{E}[\|n_j\|_2^2 \mid \mathcal{F}] \le 4K\eta^2 R^2 \zeta^2 \sigma^2.$$

Step 4. Combining, and using $\mathbb{E}[|\nu_i| \mid \mathcal{F}] \le \eta R \zeta \, \mathbb{E}[\|n_i\|_2 \mid \mathcal{F}] \le \eta R \zeta \sigma$ (Jensen),

$$\mathbb{E}\big[\|w' - w\|_1 \mid \mathcal{F}\big] \le 16\eta R M + 16 \eta R \zeta \sigma \sum_i w_i + 2 \Pr(\mathcal{A}^c \mid \mathcal{F}) \le 16 \eta R (M + \zeta\sigma) + 8K\eta^2 R^2 \zeta^2 \sigma^2.$$

□

The constant 16 is not optimized. What matters is the structure of (21): the first-order term is a w-weighted average of the per-task noise, not a maximum, and the number of objectives enters only through the second-order term $8K\eta^2 R^2 \zeta^2 \sigma^2$, which is what makes the leading term of Theorem 4.9 free of K apart from $\log K$.

D.2 Tracking error

Proof of Lemma 4.8. From line 7, for $t \ge 1$,

$$\begin{aligned} \lambda_t &= (1 - \alpha)\mathcal{M}_{t-1} + \alpha \widehat{G}_t - A_t w_t \\ &= (1 - \alpha)\big(\mathcal{M}_{t-1} - A_{t-1} w_{t-1}\big) + (1 - \alpha)\big(A_{t-1} w_{t-1} - A_t w_t\big) + \alpha n_t \\ &= (1 - \alpha)\lambda_{t-1} + (1 - \alpha)\Theta_t + \alpha n_t, \end{aligned}$$

and $\lambda_0 = \widehat{G}_0 - A_0 w_0 = n_0$ because $\alpha_0 = 1$. Unrolling,

$$\lambda_t = \underbrace{\alpha \sum_{s=1}^{t} (1 - \alpha)^{t-s} n_s + (1 - \alpha)^t n_0}_{=:\, \nu_t} \;+\; \underbrace{\sum_{s=1}^{t} (1 - \alpha)^{t-s+1} \Theta_s}_{=:\, \delta_t}. \tag{23}$$

(i) Drift. Split Θ_s according to whether x or w moved:

$$\Theta_s = \big[A_{s-1} - A_s\big] w_{s-1} \;+\; A_s \big(w_{s-1} - w_s\big).$$

For the first term, by convexity of $\|\cdot\|_*$ and Assumption 1,

$$\|\big[A_{s-1} - A_s\big]w_{s-1}\|_* \le \sum_i w_{s-1,i}\|\nabla f_i(x_{s-1}) - \nabla f_i(x_s)\|_* \le L\|x_s - x_{s-1}\| \le L\gamma R,$$

the last step because $x_s - x_{s-1} = \gamma D_{s-1}$ with $\|D_{s-1}\| \le R$; this holds *pathwise*, for every realization of the noise, and is free of K. For the second term, by convexity and Assumption 4,

$$\|A_s(w_{s-1} - w_s)\|_* \le \max_i \|\nabla f_i(x_s)\|_* \, \|w_s - w_{s-1}\|_1 \le M\|w_s - w_{s-1}\|_1.$$

The step $w_{s-1} \to w_s$ is the update of Lemma 4.7 with $\mathcal{F} = \mathcal{F}_{s-1}$, $w = w_{s-1}$, $x = x_{s-1}$, $D = D_{s-1}$ and noise $n_i = n_{s-1,i}$; the hypotheses of part (b) hold by Assumptions 3 and 4 and $\eta RM \le \frac{1}{2}$. Hence $\mathbb{E}\|w_s - w_{s-1}\|_1 \le \omega$, and

$$\mathbb{E}\|\Theta_s\|_* \le L\gamma R + M\omega =: B \qquad (s \ge 1),$$

while $\Theta_0 = 0$. Summing the geometric weights in (23),

$$\mathbb{E}\|\delta_t\|_* \le B\sum_{s=1}^{t}(1-\alpha)^{t-s+1} \le B\,\frac{1-\alpha}{\alpha} \le \frac{B}{\alpha}.$$

(ii) Noise. For $s < s'$, n_s is $\mathcal{F}_{s'}$-measurable and $\mathbb{E}[n_{s'} \mid \mathcal{F}_{s'}] = 0$, so $\mathbb{E}\langle n_s, n_{s'}\rangle = 0$: the summands of ν_t are orthogonal in L^2. Therefore, by (20),

$$\begin{aligned}\mathbb{E}\|\nu_t\|_2^2 &= \alpha^2\sum_{s=1}^{t}(1-\alpha)^{2(t-s)}\mathbb{E}\|n_s\|_2^2 + (1-\alpha)^{2t}\mathbb{E}\|n_0\|_2^2 \\ &\le \sigma^2\Big(\frac{\alpha^2}{1-(1-\alpha)^2} + (1-\alpha)^{2t}\Big) \le \sigma^2\big(\alpha + (1-\alpha)^{2t}\big),\end{aligned}$$

using $\alpha^2/(1-(1-\alpha)^2) = \alpha/(2-\alpha) \le \alpha$ for $\alpha \in (0,1]$. Converting with ζ and applying Jensen's inequality ($\mathbb{E}Y \le \sqrt{\mathbb{E}Y^2}$) and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$,

$$\mathbb{E}\|\nu_t\|_* \le \zeta\sqrt{\mathbb{E}\|\nu_t\|_2^2} \le \zeta\sigma\big(\sqrt{\alpha} + (1-\alpha)^t\big).$$

Combining (i) and (ii) with $\|\lambda_t\|_* \le \|\nu_t\|_* + \|\delta_t\|_*$ gives

$$\mathbb{E}\|\lambda_t\|_* \le \zeta\sigma\big(\sqrt{\alpha} + (1-\alpha)^t\big) + \frac{L\gamma R + M\omega}{\alpha} \qquad (t \ge 0), \tag{24}$$

which is the statement of Lemma 4.8. No maximal inequality over the K tasks has been used; K enters only through the second-order term of ω. □

D.3 One-step inequality and the rate

Lemma D.2 (Stochastic one-step inequality)**.** *Under Assumption 1, for every $w \in \Delta_K$ and every $t \ge 0$,*

$$\begin{aligned}F(x_{t+1})w \le F(x_t)w - \gamma R\|A_t w_t\|_* + \gamma\langle g_t, w - w_t\rangle + 2\gamma R\big(\|\lambda_{t-1}\|_* + \|\Theta_t\|_*\big) \\ - \gamma\Big\langle \textstyle\sum_i (w_i - w_{t,i})\, n_{t,i},\, D_t\Big\rangle + \tfrac{1}{2}L\gamma^2R^2,\end{aligned}$$

with the conventions $\lambda_{-1} = -A_0 w_0$, $\Theta_0 = 0$. The noise term is kept signed*: it is bounded in absolute value by $\gamma R\|\sum_i (w_i - w_{t,i})n_{t,i}\|_*$, but what we use is that its conditional mean given $\mathcal{F}_t$ is zero when w is deterministic.*

Proof. Weighting Lemma A.1 by w exactly as in the proof of Lemma 4.1 gives $F(x_{t+1})w \le F(x_t)w + \gamma\langle A_t w, D_t\rangle + \frac{1}{2}L\gamma^2R^2$. Write $A_t w = A_t w_t + A_t(w - w_t)$.

First part. For $t \ge 1$, $D_t \in \mathrm{lmo}(\mathcal{M}_{t-1})$, so by (2) and Hölder,

$$\langle A_t w_t, D_t\rangle = \langle \mathcal{M}_{t-1}, D_t\rangle + \langle A_t w_t - \mathcal{M}_{t-1}, D_t\rangle \le -R\|\mathcal{M}_{t-1}\|_* + R\|A_t w_t - \mathcal{M}_{t-1}\|_*.$$

By the triangle inequality $\|\mathcal{M}_{t-1}\|_* \geq \|A_t w_t\|_* - \|A_t w_t - \mathcal{M}_{t-1}\|_*$, hence $\langle A_t w_t, D_t\rangle \leq -R\|A_t w_t\|_* + 2R\|A_t w_t - \mathcal{M}_{t-1}\|_*$. Finally $A_t w_t - \mathcal{M}_{t-1} = -(\mathcal{M}_{t-1} - A_{t-1}w_{t-1}) - (A_{t-1}w_{t-1} - A_t w_t) = -\lambda_{t-1} - \Theta_t$, and the triangle inequality gives the term $2R(\|\lambda_{t-1}\|_* + \|\Theta_t\|_*)$. For $t = 0$, $\mathcal{M}_{-1} = 0$ and $D_0 \in B$ is arbitrary, so Hölder gives $\langle A_0 w_0, D_0\rangle \leq R\|A_0 w_0\|_* = -R\|A_0 w_0\|_* + 2R\|A_0 w_0\|_* = -R\|A_0 w_0\|_* + 2R(\|\lambda_{-1}\|_* + \|\Theta_0\|_*)$ under the stated conventions.

Second part. By definition, $g_{t,i} = \langle D_t, \nabla f_i(x_t; \xi_{t,i})\rangle = \langle D_t, \nabla f_i(x_t)\rangle + \langle D_t, n_{t,i}\rangle$, so

$$\langle A_t(w - w_t), D_t\rangle = \sum_i (w_i - w_{t,i})\langle \nabla f_i(x_t), D_t\rangle = \langle g_t, w - w_t\rangle - \Big\langle \sum_i (w_i - w_{t,i})\, n_{t,i},\, D_t \Big\rangle.$$

Multiplying both parts by γ and collecting terms gives the claim. □

Proof of Theorem 4.9. Throughout, $\alpha = T^{-1/2}$, $\gamma R = T^{-3/4}$ and $\eta R = T^{-3/4}/M^2$; the hypothesis $T \geq (2/M)^{4/3}$ is exactly $\eta R M \leq \frac{1}{2}$, so Lemma D.1(b), and with it Lemma 4.8, apply.

Step 1: the master inequality. Fix a deterministic $u \in \Delta_K$. Add Lemma D.2 with $w = u$ to γ/η times (6) (at $\tau = \kappa = 0$) with the same u; Lemma 4.2 is a deterministic statement about line 6 of Algorithm 2 and applies verbatim to the noisy scores, since its proof uses no property of g_t beyond finiteness. Summing over $t = 0, \ldots, T-1$, telescoping the divergences, using $\mathrm{KL}(u\|w_T) \geq 0$ and $F(x_0)u - F(x_T)u \leq \Delta$ (Assumption 2), and dividing by $\gamma R T$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \|A_t w_t\|_* \leq \frac{\Delta}{\gamma R T} + \frac{\mathrm{KL}(u\|w_0)}{\eta R T} + \frac{\eta}{2RT}\sum_t \|g_t\|_\infty^2 + \frac{L\gamma R}{2} + \frac{2}{T}\sum_{t=0}^{T-1}\big(\|\lambda_{t-1}\|_* + \|\Theta_t\|_*\big) - \frac{1}{RT}\sum_{t=0}^{T-1}\Big\langle \textstyle\sum_i (u_i - w_{t,i})\, n_{t,i},\, D_t\Big\rangle. \quad (25)$$

This holds pathwise. For the stationarity bound we take $u = u_K$, so that $\mathrm{KL}(u\|w_0) = 0$ and the second term vanishes; for the regret bound of Corollary D.3 a general u is kept and $\mathrm{KL}(u\|w_0) \leq \ln K$. We now take expectations term by term.

Step 2: the noise cross term vanishes. u is deterministic and w_t, D_t are $\mathcal{F}_t$-measurable, while $\mathbb{E}[n_{t,i} \mid \mathcal{F}_t] = 0$. By the tower property,

$$\mathbb{E}\Big\langle \sum_i (u_i - w_{t,i}) n_{t,i}, D_t\Big\rangle = \mathbb{E}\Big[\Big\langle \sum_i (u_i - w_{t,i})\, \mathbb{E}[n_{t,i} \mid \mathcal{F}_t],\, D_t\Big\rangle\Big] = 0.$$

This is exactly what evaluating the oracle at $\mathcal{M}_{t-1}$ rather than $\mathcal{M}_t$ buys: with $D_t = \mathrm{lmo}(\mathcal{M}_t)$ the direction would depend on ξ_t and this term would not vanish.

Step 3: the mirror-descent error term. By Hölder and Assumption 4, $\|g_t\|_\infty \leq R\max_i \|\nabla f_i(x_t; \xi_{t,i})\|_* \leq R\big(M + \zeta N_t\big)$ with $N_t := \max_i \|n_{t,i}\|_2$, so $\mathbb{E}\|g_t\|_\infty^2 \leq 2R^2\big(M^2 + \zeta^2 \mathbb{E}N_t^2\big) \leq 2R^2\big(M^2 + \zeta^2 K\sigma^2\big)$, using $\mathbb{E}N_t^2 \leq \sum_i \mathbb{E}\|n_{t,i}\|_2^2 \leq K\sigma^2$. Hence

$$\frac{\eta}{2RT}\sum_t \mathbb{E}\|g_t\|_\infty^2 \leq \eta R\big(M^2 + \zeta^2 K\sigma^2\big) = T^{-3/4}\Big(1 + \frac{\zeta^2 K\sigma^2}{M^2}\Big).$$

This is one of the two places where K appears polynomially; both are of lower order than $T^{-1/4}$.

Step 4: drift and noise terms. Let $B = L\gamma R + M\omega$ as in the proof of Lemma 4.8. Using $\|\lambda_{-1}\|_* = \|A_0 w_0\|_* \leq M$, $\Theta_0 = 0$, (24) for $t \geq 1$ and $\mathbb{E}\|\Theta_t\|_* \leq B$ for $t \geq 1$, together with $\sum_{t\geq 0}(1-\alpha)^t \leq 1/\alpha$,

$$\frac{2}{T}\sum_{t=0}^{T-1}\mathbb{E}\big(\|\lambda_{t-1}\|_* + \|\Theta_t\|_*\big) \leq \frac{2M}{T} + 2\zeta\sigma\sqrt{\alpha} + \frac{2\zeta\sigma}{\alpha T} + \frac{2B}{\alpha} + 2B.$$

With the stated step sizes,

$$M\omega = 16\, T^{-3/4}\Big(1 + \frac{\zeta\sigma}{M}\Big) + \frac{8K\zeta^2\sigma^2}{M^3}\, T^{-3/2},$$

$$B = L\, T^{-3/4} + 16\, T^{-3/4}\Big(1 + \frac{\zeta\sigma}{M}\Big) + \frac{8K\zeta^2\sigma^2}{M^3}\, T^{-3/2},$$

so that $2B/\alpha = 2\big(L + 16 + 16\zeta\sigma/M\big)T^{-1/4} + 16K\zeta^2\sigma^2M^{-3}T^{-1}$ and $2B \le 2(L + 16 + 16\zeta\sigma/M)T^{-3/4} + 16K\zeta^2\sigma^2M^{-3}T^{-3/2}$.

Step 5: collecting. Substituting into (25) after taking expectations,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|A_tw_t\|_* \le \frac{\Delta + 2L + 32 + 2\zeta\sigma\,(1 + 16/M)}{T^{1/4}} + \frac{2\zeta\sigma}{T^{1/2}} + \frac{1 + \frac{5}{2}L + 32(1 + \zeta\sigma/M) + \zeta^2K\sigma^2/M^2}{T^{3/4}} + \frac{2M + 32K\zeta^2\sigma^2/M^3}{T}, \tag{26}$$

where in the last term we used $T^{-3/2} \le T^{-1}$. This is (12); setting the right-hand side to ϵ yields $T = O(\epsilon^{-4})$. The number of objectives does not enter the $T^{-1/4}$ term (the comparator is uniform); the two terms carrying K polynomially are of order $T^{-3/4}$ (from the ℓ_∞ error term of mirror descent, Step 3) and T^{-1} (from the second-order part of the weight increment, Lemma D.1(b)).

Step 6: the CA distance. The inequality $\mathrm{CA}_t \le 2\|A_tw_t\|_*$ of Corollary C.1 is pathwise and holds for every measurable selection $w^\star(x_t)$, so averaging and taking expectations gives $\frac{1}{T}\sum_t \mathbb{E}\,\mathrm{CA}_t \le 2\times$(26).

Finally, note that with the deterministic choice $\eta R = \Theta(T^{-1/2}/M)$ the term $M\omega/\alpha$ would contain $16\eta RM^2/\alpha = \Theta(M)$, which does not vanish as $T \to \infty$; this is why the weight step size must be reduced to $\Theta(T^{-3/4}/M^2)$ in the stochastic setting. □

Corollary D.3 (Stochastic inner regret)**.** *Under the conditions of Theorem 4.9, for every $u \in \Delta_K$, $\frac{1}{T}\sum_t \mathbb{E}\|\nabla F(x_t)w_t\|_* \le \frac{1}{T}\sum_t \mathbb{E}\|\nabla F(x_t)u\|_*$ plus the right-hand side of* (12) *with Δ replaced by $M^2\ln K$; hence $\frac{1}{T}\sum_t \mathbb{E}\,\mathrm{gap}_t$ is bounded by the same quantity plus $\bar{V}_T := \min_{u\in\Delta_K}\mathbb{E}\frac{1}{T}\sum_t\big[\|\nabla F(x_t)u\|_* - \min_w\|\nabla F(x_t)w\|_*\big]$.*

Note $\bar{V}_T \ge \mathbb{E}V_T$: the comparator must be deterministic for the noise cross term to vanish, so the stochastic statement is against the best fixed weighting *in expectation*, not against the realized hindsight minimizer.

The linear-K terms in C_3, C_4 are two maxima over tasks of noise magnitudes bounded with second moments only, and are intrinsic there (noise of size $\sqrt{K}\sigma$ with probability $1/K$ makes some task an outlier at every step). A tail assumption removes them.

Corollary D.4 (Polylogarithmic K under sub-Gaussian noise)**.** *Replace the variance bound in Assumption 3 by the conditional tail bound $\Pr\big(\|\nabla f_i(x_t;\xi_t) - \nabla f_i(x_t)\|_2 > s \mid \mathcal{F}_t\big) \le 2e^{-s^2/(2\sigma^2)}$ for all $s \ge 0$ and i. Then Theorem 4.9 holds for $T^{3/2} \ge (8\zeta^2\sigma^2/M^4)\ln(4KT)$ with $\sigma \to 2\sigma$ in C_1, C_3, $C_2 = 4\zeta\sigma + 2M$, $\zeta^2K\sigma^2/M^2 \to 2\zeta^2\sigma^2(1 + \ln 2K)/M^2$ in C_3, and $C_4 = 4M$ (Appendix D): under these conditions every displayed constant depends on K only through $\ln K$, so $T = O(\epsilon^{-4})$ with a K-free constant for $K \le \exp(O(\epsilon^{-2}))$; Corollary D.3 holds with the same constants plus $M^2\ln K$. For a.s. bounded noise the constants of Theorem 4.9 are free of K.*

Proof of Corollary D.3. Fix a deterministic $u \in \Delta_K$ and write $\psi_t(w) = \|A_tw\|_*$ and $\bar{g}_{t,i} := \langle D_t, \nabla f_i(x_t)\rangle$ for the *noiseless* scores, so that $g_{t,i} = \bar{g}_{t,i} + \langle D_t, n_{t,i}\rangle$. Because $D_t = \mathrm{lmo}(\mathcal{M}_{t-1})$ is not an element of $\mathrm{lmo}(A_tw_t)$, Lemma A.3 does not apply directly; we argue from the definition of the dual norm instead. For any $v \in \Delta_K$, $\|D_t\| \le R$ gives $\psi_t(v) = \|A_tv\|_* \ge -\langle A_tv, D_t\rangle/R = -\langle\bar{g}_t, v\rangle/R$; at $v = u$ this is $\psi_t(u) \ge -\langle\bar{g}_t, u\rangle/R$. For w_t, the first part of the proof of Lemma D.2 gives $\langle A_tw_t, D_t\rangle \le -R\|A_tw_t\|_* + 2R(\|\lambda_{t-1}\|_* + \|\Theta_t\|_*)$, i.e. $\psi_t(w_t) \le -\langle\bar{g}_t, w_t\rangle/R + 2(\|\lambda_{t-1}\|_* + \|\Theta_t\|_*)$. Subtracting,

$$\begin{aligned}\psi_t(w_t) - \psi_t(u) &\le \frac{1}{R}\langle\bar{g}_t, u - w_t\rangle + 2\big(\|\lambda_{t-1}\|_* + \|\Theta_t\|_*\big)\\ &= \frac{1}{R}\langle g_t, u - w_t\rangle - \frac{1}{R}\Big\langle\sum_i (u_i - w_{t,i})n_{t,i}, D_t\Big\rangle + 2\big(\|\lambda_{t-1}\|_* + \|\Theta_t\|_*\big).\end{aligned}$$

Sum over t, divide by T, bound $\frac{1}{RT}\sum_t\langle g_t, u - w_t\rangle$ by $\frac{\ln K}{\eta RT} + \frac{\eta}{2RT}\sum_t\|g_t\|_\infty^2$ via (6), and take expectations. The signed noise term vanishes exactly as in Step 2 of the proof of Theorem 4.9; the remaining terms are precisely those of (25) without $\Delta/(\gamma RT)$ and $L\gamma R/2$ but with

$\mathrm{KL}(u\|w_0)/(\eta RT) \le \ln K/(\eta RT) = M^2 \ln K\, T^{-1/4}$ for the general comparator, and Steps 3–5 of that proof bound the rest by (26) with Δ removed and $2L$ (from $2B/\alpha$) retained. For the gap statement, fix a deterministic $u \in \Delta_K$ and write, as in the proof of Corollary 4.6, $\mathbb{E}\,\mathrm{gap}_t = \mathbb{E}[\psi_t(w_t) - \psi_t(u)] + \mathbb{E}[\psi_t(u) - \min_w \psi_t(w)]$; average over t, bound the first average by the regret bound just proved, and then take the minimum over deterministic u of the second average, which is $\bar{V}_T$. The comparator must be deterministic because Step 2 of the proof of Theorem 4.9 uses $\mathbb{E}[n_{t,i} \mid \mathcal{F}_t] = 0$ against a $\mathcal{F}_t$-measurable $u - w_t$; the realized hindsight minimizer $u(\omega) = \arg\min_u \frac{1}{T}\sum_t \psi_t(u)$ depends on the whole trajectory and is not admissible. Since $\mathbb{E}\min_u(\cdot) \le \min_u \mathbb{E}(\cdot)$, one has $\mathbb{E}V_T \le \bar{V}_T$, and the bound with $\mathbb{E}V_T$ in place of $\bar{V}_T$ is *not* established by this argument. □

Proof of Corollary D.4. Write $X_i := \|n_{t,i}\|_2$, so $\Pr(X_i > s \mid \mathcal{F}_t) \le 2e^{-s^2/(2\sigma^2)}$. Then $\mathbb{E}[X_i^2 \mid \mathcal{F}_t] = \int_0^\infty \Pr(X_i^2 > u)\, du \le \int_0^\infty 2e^{-u/(2\sigma^2)} du = 4\sigma^2$, so Assumption 3 holds with σ replaced by 2σ, and every step of the proofs of Lemma D.1, Lemma 4.8 and Theorem 4.9 goes through with that substitution except the two that bound a maximum over tasks. Those improve as follows.

The ℓ_∞ error term (Step 3). With $N_t = \max_i X_i$ and $u_0 := 2\sigma^2 \ln(2K)$,

$$\mathbb{E}[N_t^2 \mid \mathcal{F}_t] = \int_0^\infty \Pr(N_t^2 > u)\, du \le u_0 + \int_{u_0}^\infty 2Ke^{-u/(2\sigma^2)} du = u_0 + 4K\sigma^2 e^{-u_0/(2\sigma^2)} = 2\sigma^2\big(1 + \ln 2K\big),$$

by the union bound. Hence $\mathbb{E}\|g_t\|_\infty^2 \le 2R^2\big(M^2 + 2\zeta^2\sigma^2(1 + \ln 2K)\big)$ and the Step 3 term is $T^{-3/4}\big(1 + 2\zeta^2\sigma^2(1 + \ln 2K)/M^2\big)$.

The union bound in Lemma D.1(b). Since $|\nu_j| \le \eta R\zeta X_j$,

$$\Pr(\mathcal{A}^c \mid \mathcal{F}) \le \sum_j \Pr\Big(X_j > \frac{1}{2\eta R\zeta}\Big) \le 2K \exp\Big(-\frac{1}{8\eta^2 R^2\zeta^2\sigma^2}\Big),$$

so $\omega = 16\eta R(M + 2\zeta\sigma) + 4K\exp\big(-1/(8\eta^2R^2\zeta^2\sigma^2)\big)$, using $\mathbb{E}[X_i \mid \mathcal{F}] \le (\mathbb{E}[X_i^2 \mid \mathcal{F}])^{1/2} \le 2\sigma$ in the first term. With $\eta R = T^{-3/4}/M^2$ the exponent is $-M^4T^{3/2}/(8\zeta^2\sigma^2)$, and the hypothesis $T^{3/2} \ge (8\zeta^2\sigma^2/M^4)\ln(4KT)$ is exactly $4K\exp(-M^4T^{3/2}/(8\zeta^2\sigma^2)) \le 1/T$. Hence in the proof of Theorem 4.9, $M\omega \le 16T^{-3/4}(1 + 2\zeta\sigma/M) + M/T$, so $2B/\alpha \le 2(L + 16 + 32\zeta\sigma/M)T^{-1/4} + 2MT^{-1/2}$ and $2B \le 2(L + 16 + 32\zeta\sigma/M)T^{-3/4} + 2M/T$.

Collecting. Repeating Step 5 with $\sigma \to 2\sigma$ in the noise terms ($2\zeta\sigma\sqrt{\alpha} \to 4\zeta\sigma T^{-1/4}$, $2\zeta\sigma/(\alpha T) \to 4\zeta\sigma T^{-1/2}$) and the two replacements above gives the constants in the statement: $C_1 = \Delta + 2L + 32 + 4\zeta\sigma(1 + 16/M)$, $C_2 = 4\zeta\sigma + 2M$, $C_3 = 1 + \frac{5}{2}L + 32(1 + 2\zeta\sigma/M) + 2\zeta^2\sigma^2(1 + \ln 2K)/M^2$, $C_4 = 2M + 2M = 4M$. The requirement $C_3 \le 16C_1^3/\epsilon^2$ for a K-free iteration count now reads $\ln(2K) \lesssim \epsilon^{-2}$. For almost surely bounded noise $X_i \le \hat{\sigma}$ one has $N_t \le \hat{\sigma}$ and $\Pr(\mathcal{A}^c) = 0$ once $\eta R\zeta\hat{\sigma} \le \frac{1}{2}$, so no maximum over tasks is ever taken and the constants are those of Theorem 4.9 with $\sigma \to \hat{\sigma}$ and the K-terms deleted. The regret corollary adds $\mathrm{KL}(u\|w_0)/(\eta RT) \le M^2 \ln K\, T^{-1/4}$ as before. □

Remark D.1 (On the gradient-bound assumption). Assumption 4 is used in three places: to bound the w-component of the drift ($\|A_s(w_{s-1} - w_s)\|_* \le M\|w_s - w_{s-1}\|_1$), in Lemma D.1(b), and in Step 3 above. In the deterministic setting the analogous bound is *derived* in Lemma 4.3 by an induction on the function values along the trajectory. That induction does not transfer to Algorithm 2: the stochastic one-step inequality contains the random terms $\|\lambda_{t-1}\|_*$ and the noise cross term, which are controlled in expectation but not almost surely, so it yields a bound on $\mathbb{E}f_i(x_t)$ and not the almost-sure gradient bound that the drift argument requires. Removing the assumption would need a stopping-time argument in the manner of Zhang et al. (2025), which we leave aside. Note that under an almost-sure bound on the *stochastic* gradients, $\|\nabla f_i(x;\xi)\|_* \le \widehat{M}$ for all x, ξ, Assumption 4 holds with $M = \widehat{M}$ directly, since $\|\nabla f_i(x)\|_* = \|\mathbb{E}_\xi \nabla f_i(x;\xi)\|_* \le \mathbb{E}_\xi\|\nabla f_i(x;\xi)\|_*$ by Jensen; in that case the assumption is not additional.

E PER-TASK TRACKING VARIANT

For completeness we record the alternative referred to in Section 4.2; this is an upper-bound comparison of the two tracking schemes, not a convergence theorem. Track $m_{t,i} = (1-\alpha)m_{t-1,i} + \alpha\nabla f_i(x_t;\xi_{t,i})$ per task and set $\widehat{G}_t = \sum_i w_{t,i}m_{t,i}$, $D_t = \mathrm{lmo}(\widehat{G}_t)$, $g_{t,i} = \langle D_t, m_{t,i}\rangle$. Writing $\lambda_{t,i} = m_{t,i} - \nabla f_i(x_t)$ and $\Lambda_t = [\lambda_{t,1},\dots,\lambda_{t,K}]$, the recursion for each $\lambda_{t,i}$ has drift $\nabla f_i(x_{t-1}) - \nabla f_i(x_t)$ only, bounded by $L\gamma R$ deterministically, so no w-drift term appears and $\eta R = \Theta(T^{-1/2}/M)$ remains admissible, which gives the better constant $M\sqrt{\log K}$ in place of $M^2\log K$. The one-step inequality however involves $\|\Lambda_t w_t\|_* = \|\sum_i w_{t,i}\lambda_{t,i}\|_*$, and w_t is correlated with the past noise that produced $\lambda_{t,i}$: the weight update raises $w_{t,i}$ precisely when past scores $\langle D_s, n_{s,i}\rangle$ were positive, and those same noise vectors appear in $\lambda_{t,i}$. The Jensen argument of (20) bounds $\mathbb{E}\|\sum_i w_{t,i}\lambda_{t,i}\|^2$ by $\sum_i \mathbb{E}[w_{t,i}\|\lambda_{t,i}\|^2]$, which is not $\le \max_i \mathbb{E}\|\lambda_{t,i}\|^2$ when w_t and $\lambda_{t,i}$ are dependent. One is left with the pathwise bound $\|\Lambda_t w_t\|_* \le \max_i \|\lambda_{t,i}\|_*$ and a maximal inequality

$$\mathbb{E}\max_i \|\nu_{t,i}\|_2 \le \Big(\sum_i \mathbb{E}\|\nu_{t,i}\|_2^2\Big)^{1/2} \le \sqrt{K}\max_i \big(\mathbb{E}\|\nu_{t,i}\|_2^2\big)^{1/2},$$

so a factor $c_K = \sqrt{K}$ under Assumption 3 alone (improving to $O(\sqrt{\log K})$ if each $n_{t,i}$ is sub-Gaussian) multiplies $\zeta\sigma$ in the leading term. The two variants therefore trade a better M-dependence against a worse K-dependence, and only the blended form has a leading term free of K under bounded variance alone. Whether c_K can be removed for per-task tracking is open.

F ITERATION-WISE INNER GAP: PROOFS FOR SECTION 5

Throughout, $\|\cdot\|$ is an arbitrary norm, $A_t := \nabla F(x_t)$, and

$$f_t(w) := \|A_t w\|_*, \qquad r(w) := \mathrm{KL}(w\|u_K), \qquad \tilde{\psi}_t := f_t + \tau r,$$

$$p_t := \operatorname*{arg\,min}_{w\in\Delta_K^\kappa} \tilde{\psi}_t(w), \qquad \Phi_t := \mathrm{KL}(p_t\|w_t),$$

with $\Delta_K^\kappa = \{w \in \Delta_K : w_i \ge \kappa\}$, $\Lambda := \ln(1/\kappa)$, $\bar{\eta} := \eta R$, $\bar{\gamma} := \gamma R$, $\rho := \bar{\eta}\tau$, and $\mathrm{gap}_t := f_t(w_t) - \min_{\Delta_K} f_t$. For $\tau > 0$, $\tilde{\psi}_t$ is strictly convex (as r is) and continuous on the compact set Δ_K^κ, so p_t exists and is unique. We use three facts about f_t, valid in every norm along a trajectory with $\max_i \|\nabla f_i(x_t)\|_* \le M$:

(F1) f_t is convex and M-Lipschitz in $\|\cdot\|_1$ on Δ_K: $|f_t(v) - f_t(w)| \le \|A_t(v-w)\|_* \le \sum_i |v_i - w_i|\|\nabla f_i(x_t)\|_* \le M\|v-w\|_1$.

(F2) $\hat{g}_t := -g_t/R \in \partial f_t(w_t)$ (Lemma A.3), and $\|\hat{g}_t\|_\infty \le \max_i \|\nabla f_i(x_t)\|_* \le M$ by Hölder.

(F3) For any $w \in \Delta_K$ and x, x', $|\|\nabla F(x)w\|_* - \|\nabla F(x')w\|_*| \le \|(\nabla F(x) - \nabla F(x'))w\|_* \le \sum_i w_i\|\nabla f_i(x) - \nabla f_i(x')\|_* \le L\|x - x'\|$ (Assumption 1).

Lemma F.1 (Closed form of line 6)**.** *The minimizer in line 6 of Algorithm 1 is given by* (4)*:* $w_{t+1,i} = \max\{\kappa, c_i/Z_t\}$ *with* $c_i = w_{t,i}^{1/(1+\rho)} e^{\eta g_{t,i}/(1+\rho)}$ *and* $Z_t > 0$ *the unique normalizer for which* $\sum_i w_{t+1,i} = 1$ *(for* $\kappa < 1/K$*; if* $\kappa = 1/K$ *then* $w_{t+1} = u_K$*).*

Proof. The objective $h(w) = -\eta\langle g_t, w\rangle + \rho r(w) + \mathrm{KL}(w\|w_t)$ is strictly convex and differentiable on the relative interior, and the constraints are affine, so the KKT conditions are necessary and sufficient. With multiplier ν for $\sum_i w_i = 1$ and $\lambda_i \ge 0$ for $w_i \ge \kappa$, and using $\partial_i r(w) = \ln w_i + \ln K + 1$, $\partial_i \mathrm{KL}(w\|w_t) = \ln w_i - \ln w_{t,i} + 1$, stationarity reads $-\eta g_{t,i} + (1+\rho)\ln w_i + \rho(\ln K + 1) + 1 - \ln w_{t,i} - \nu - \lambda_i = 0$, i.e. $(1+\rho)\ln w_i = \ln w_{t,i} + \eta g_{t,i} + \mathrm{const} + \lambda_i$. If $\lambda_i = 0$ this gives $w_i = c_i/Z_t$ for a constant Z_t; if $\lambda_i > 0$ then $w_i = \kappa$, and $\lambda_i \ge 0$ is equivalent to $c_i/Z_t \le \kappa$. Hence $w_{t+1,i} = \max\{\kappa, c_i/Z_t\}$. The map $z \mapsto \sum_i \max\{\kappa, c_i/z\}$ is continuous, non-increasing, tends to $K\kappa < 1$ as $z \to \infty$ and to ∞ as $z \to 0$, and is strictly decreasing wherever its value exceeds $K\kappa$, so Z_t is unique. □

F.1 COMPUTING THE CLIPPED UPDATE

Equation (4) determines w_{t+1} implicitly through the normalizer Z_t, or equivalently through the set $S^\star := \{i : c_i/Z_t < \kappa\}$ of clipped coordinates. Given $S^\star$, Z_t is explicit: the unclipped coordinates must carry the mass $1 - |S^\star|\kappa$, so $Z_t = \sum_{i\notin S^\star} c_i/(1 - |S^\star|\kappa)$. The clipped set is found by the following recurrence, which is the "clip, renormalize, repeat" rule stated as a fixed-point iteration. Write $c = (c_i)_{i\in[K]}$ with $c_i = w_{t,i}^{1/(1+\rho)} e^{\eta g_{t,i}/(1+\rho)} > 0$ and, for $S \subseteq [K]$ with $|S|\kappa < 1$,

$$Z(S) := \frac{\sum_{i\notin S} c_i}{1 - |S|\kappa}. \tag{27}$$

Algorithm 3 Clipped renormalization (KL projection of the softmax output onto Δ_K^κ)

Require: $c \in \mathbb{R}^K_{>0}$; $\kappa \in [0, 1/K)$

$S_0 \leftarrow \emptyset$
for $k = 0, 1, 2, \dots$ **do**
  $Z_k \leftarrow Z(S_k)$ ▷ renormalize the unclipped coordinates, (27)
  $A_k \leftarrow \{i \notin S_k : \ c_i < \kappa Z_k\}$ ▷ coordinates that fall below the floor
  if $A_k = \emptyset$ **then return** $w_i = \kappa$ for $i \in S_k$, $\quad w_i = c_i/Z_k$ for $i \notin S_k$
  end if
  $S_{k+1} \leftarrow S_k \cup A_k$
end for

Lemma F.2 (Correctness and termination of Algorithm 3)**.** *Let $\kappa \in [0, 1/K)$ and $c \in \mathbb{R}^K_{>0}$, and let $Z^\star$ be the unique normalizer of Lemma F.1, with $S^\star = \{i : c_i/Z^\star < \kappa\}$. Along Algorithm 3:*

(i) $S_0 \subseteq S_1 \subseteq \cdots \subseteq S^\star$, and $Z_0 < Z_1 < \cdots \le Z^\star$ while the recurrence continues;

(ii) the recurrence returns after at most $|S^\star| + 1 \le K$ rounds, in the first round k with $A_k = \emptyset$; at that round $S_k = S^\star$ and $Z_k = Z^\star$, so the returned w is w_{t+1} of (4). (A coordinate with $c_i = \kappa Z^\star$ exactly may sit in either set; the returned vector is the same.)

Each round costs $O(K)$, so the update costs $O(K \min\{K, |S^\star| + 1\})$; sorting c once gives an $O(K \log K)$ implementation, since $S^\star$ is a lower set in the order of c. At $\kappa = 0$ the recurrence stops after one round with $S_0 = \emptyset$ and $Z_0 = \sum_i c_i$, i.e. the plain softmax.

Proof. Two facts about $Z(\cdot)$ are used. First, by the definition of $S^\star$ and $\sum_i \max\{\kappa, c_i/Z^\star\} = 1$,

$$\sum_{i\notin S^\star} c_i = Z^\star\big(1 - |S^\star|\kappa\big), \qquad c_i < \kappa Z^\star \ \ (i \in S^\star), \qquad c_i \ge \kappa Z^\star \ \ (i \notin S^\star). \tag{28}$$

Second, for $S \subseteq S^\star$,

$$\sum_{i\notin S} c_i = \sum_{i\notin S^\star} c_i + \sum_{i\in S^\star\setminus S} c_i \le Z^\star(1 - |S^\star|\kappa) + \kappa Z^\star\,(|S^\star| - |S|) = Z^\star(1 - |S|\kappa),$$

with strict inequality if $S \neq S^\star$; dividing by $1 - |S|\kappa > 0$ gives

$$Z(S) \le Z^\star \quad \text{for all } S \subseteq S^\star, \qquad \text{with equality iff } S = S^\star. \tag{29}$$

(i) Monotonicity. We show by induction that $S_k \subseteq S^\star$. This holds for $S_0 = \emptyset$. If $S_k \subseteq S^\star$ then $Z_k = Z(S_k) \le Z^\star$ by (29), so every $i \in A_k$ satisfies $c_i < \kappa Z_k \le \kappa Z^\star$, i.e. $i \in S^\star$; hence $S_{k+1} \subseteq S^\star$. The sets are nested by construction. For the normalizers, if $A_k \neq \emptyset$ then, using $c_i < \kappa Z_k$ for $i \in A_k$,

$$\sum_{i\notin S_{k+1}} c_i = Z_k(1 - |S_k|\kappa) - \sum_{i\in A_k} c_i > Z_k(1 - |S_k|\kappa) - |A_k|\kappa Z_k = Z_k\big(1 - |S_{k+1}|\kappa\big),$$

and $1 - |S_{k+1}|\kappa \ge 1 - |S^\star|\kappa > 0$, so $Z_{k+1} > Z_k$.

(ii) Termination and correctness. Since the S_k are nested subsets of $S^\star$ and strictly grow while $A_k \neq \emptyset$, the recurrence stops after at most $|S^\star| + 1$ rounds. At termination, $A_k = \emptyset$ means $c_i \geq \kappa Z_k$ for all $i \notin S_k$, while every $i \in S_k$ was added at some round $j < k$ with $c_i < \kappa Z_j \leq \kappa Z_k$ by (i). Hence the returned vector satisfies $w_i = \max\{\kappa, c_i/Z_k\}$ for every i, and it sums to 1 by the definition of $Z(S_k)$. It is therefore the water-filling vector of Lemma F.1 with normalizer Z_k; by the uniqueness of the normalizer proved there, $Z_k = Z^\star$, and consequently $S_k = S^\star$ and $w = w_{t+1}$.

For the complexity claim, note that $S^\star$ consists of the coordinates with the smallest c_i: if $c_i \leq c_j$ and $j \in S^\star$ then $c_i/Z^\star \leq c_j/Z^\star < \kappa$. After sorting, one can scan the candidates $S = \{\text{the } s \text{ smallest}\}$ for $s = 0, 1, \ldots$ and stop at the first s for which the $(s+1)$-st smallest c is at least $\kappa Z(S)$; by (29) and the argument above this is $S^\star$. □

At $\tau = \kappa = 0$ the update is a single softmax, and for $\kappa > 0$ the recurrence typically stops after one or two rounds, so the per-iteration cost of line 6 is $O(K)$ in practice and never more than $O(K^2)$; it is negligible against the K backward passes.

Remark F.1 (Why the unsquared objective). In Euclidean geometry, the squared inner objective $\frac{1}{2}\|\nabla F(x_t)w\|_2^2$ is smooth. The analyses of Zhang et al. (2025) exploit this structure for iteration-wise CA distance. Our oracle instead supplies a subgradient of the unsquared norm, matching the linear descent term in Lemma 4.1. In Euclidean geometry, the squared-objective gradient is $\|G_t\|_2$ times the unsquared gradient when $G_t \neq 0$; the latter involves normalization by $\|G_t\|_2$ and is ill-conditioned near a zero blend.

For general dual norms, nonsmoothness can also occur at nonzero blends: the ℓ_1 norm is nonsmooth at vectors with zero coordinates, and the nuclear norm at rank-deficient matrices. Squaring need not remove this nonsmoothness. Our analysis therefore uses only convexity, Lipschitz continuity, and the oracle subgradient identity. It incurs the slack $2\bar{\eta}^2 M^2$ per step in Lemma F.3, a Hölder-$1/2$ sensitivity bound, and the $1/\kappa$ factor in Lemma F.4. These bounds lead to the stated burn-in and total-horizon exponents; sharper analysis under additional norm smoothness assumptions remains open.

F.2 Contraction with slack

Lemma F.3 (Contraction toward the regularized target)**.** *Suppose* $\max_i \|\nabla f_i(x_t)\|_* \leq M$. *For every* $u \in \Delta_K^\kappa$,

$$(1+\rho)\,\mathrm{KL}(u\|w_{t+1}) \;\leq\; \mathrm{KL}(u\|w_t) + \bar{\eta}\big[\tilde{\psi}_t(u) - \tilde{\psi}_t(w_{t+1})\big] + 2\bar{\eta}^2 M^2.$$

In particular, since p_t *minimizes* $\tilde{\psi}_t$ *over* $\Delta_K^\kappa \ni w_{t+1}$,

$$\mathrm{KL}(p_t\|w_{t+1}) \;\leq\; \frac{\Phi_t + 2\bar{\eta}^2 M^2}{1+\rho}.$$

Proof. Start from (16) in the proof of Lemma 4.2, written with $\hat{g}_t = -g_t/R$, so that $\eta\langle g_t, u - w_{t+1}\rangle = -\bar{\eta}\langle \hat{g}_t, u - w_{t+1}\rangle$:

$$(1+\rho)\,\mathrm{KL}(u\|w_{t+1}) \leq \mathrm{KL}(u\|w_t) - \mathrm{KL}(w_{t+1}\|w_t) + \bar{\eta}\langle \hat{g}_t, u - w_{t+1}\rangle + \rho\big(r(u) - r(w_{t+1})\big).$$

For the inner product, split $u - w_{t+1} = (u - w_t) + (w_t - w_{t+1})$ and use (F2): the subgradient inequality gives $\langle \hat{g}_t, u - w_t\rangle \leq f_t(u) - f_t(w_t)$, and Hölder gives $\langle \hat{g}_t, w_t - w_{t+1}\rangle \leq M\|w_{t+1} - w_t\|_1$. Then, by (F1), $f_t(u) - f_t(w_t) = f_t(u) - f_t(w_{t+1}) + \big(f_t(w_{t+1}) - f_t(w_t)\big) \leq f_t(u) - f_t(w_{t+1}) + M\|w_{t+1} - w_t\|_1$. Hence

$$\begin{aligned}\bar{\eta}\langle \hat{g}_t, u - w_{t+1}\rangle &\leq \bar{\eta}\big(f_t(u) - f_t(w_{t+1})\big) + 2\bar{\eta}M\|w_{t+1} - w_t\|_1 \\ &\leq \bar{\eta}\big(f_t(u) - f_t(w_{t+1})\big) + 2\bar{\eta}^2 M^2 + \tfrac{1}{2}\|w_{t+1} - w_t\|_1^2,\end{aligned}$$

by Young's inequality $2ab \leq 2a^2 + \frac{1}{2}b^2$ with $a = \bar{\eta}M$, $b = \|w_{t+1} - w_t\|_1$. Pinsker's inequality bounds the last term by $\mathrm{KL}(w_{t+1}\|w_t)$, which cancels the $-\mathrm{KL}(w_{t+1}\|w_t)$ above. Substituting and using $\rho = \bar{\eta}\tau$,

$$(1+\rho)\,\mathrm{KL}(u\|w_{t+1}) \leq \mathrm{KL}(u\|w_t) + \bar{\eta}\big[f_t(u) + \tau r(u) - f_t(w_{t+1}) - \tau r(w_{t+1})\big] + 2\bar{\eta}^2 M^2,$$

which is the claim. The second statement follows from $\tilde{\psi}_t(p_t) \leq \tilde{\psi}_t(w_{t+1})$, as $w_{t+1} \in \Delta_K^\kappa$ by construction. □

The additive slack $2\bar{\eta}^2 M^2$ arises from our subgradient analysis of the potentially nonsmooth objective f_t. This argument does not establish contraction without slack. Euclidean squared-objective analyses exploit additional smoothness; such smoothness is not available for every dual norm.

F.3 THE MOVING TARGET

Lemma F.4 (Sensitivity of the regularized optimum)**.** *Let* $p(x) := \arg\min_{\Delta_K^\kappa}\{\|\nabla F(x)w\|_* + \tau r(w)\}$. *For any* x, x',

$$\|p(x) - p(x')\|_1 \;\le\; \sqrt{\frac{2L\|x - x'\|}{\tau}},$$

and hence, along the trajectory,

$$\chi^2(p_t\|p_{t+1}) := \sum_i \frac{(p_{t,i} - p_{t+1,i})^2}{p_{t+1,i}} \;\le\; \frac{2L\bar{\gamma}}{\tau\kappa}.$$

Proof. Write $p = p(x)$, $p' = p(x')$, $\tilde{\psi}_x = f_x + \tau r$ with $f_x = \|\nabla F(x)\cdot\|_*$. Since f_x is convex and finite, the first-order optimality condition for the convex problem $\min_{\Delta_K^\kappa}\tilde{\psi}_x$ provides $s \in \partial f_x(p)$ with $\langle s+\tau\nabla r(p), p'-p\rangle \ge 0$. Convexity of f_x and the Bregman identity $r(p') = r(p)+\langle\nabla r(p), p'-p\rangle + \mathrm{KL}(p'\|p)$ give

$$\tilde{\psi}_x(p') - \tilde{\psi}_x(p) \ge \langle s + \tau\nabla r(p), p' - p\rangle + \tau\,\mathrm{KL}(p'\|p) \ge \tau\,\mathrm{KL}(p'\|p) \ge \tfrac{\tau}{2}\|p' - p\|_1^2,$$

the last step by Pinsker. Symmetrically $\tilde{\psi}_{x'}(p) - \tilde{\psi}_{x'}(p') \ge \frac{\tau}{2}\|p - p'\|_1^2$. Adding and cancelling the identical τr terms,

$$\tau\|p - p'\|_1^2 \le \big[f_x(p') - f_{x'}(p')\big] - \big[f_x(p) - f_{x'}(p)\big] \le 2\sup_{w\in\Delta_K}|f_x(w) - f_{x'}(w)| \le 2L\|x - x'\|$$

by (F3), which gives the first claim. For the second, $\|x_{t+1} - x_t\| \le \bar{\gamma}$, and since $p_{t+1,i} \ge \kappa$, $\chi^2(p_t\|p_{t+1}) \le \kappa^{-1}\sum_i(p_{t,i} - p_{t+1,i})^2 \le \kappa^{-1}\|p_t - p_{t+1}\|_1^2$. □

This Hölder-$\frac{1}{2}$ bound uses only strong convexity and the uniform closeness (F3) of f_x and $f_{x'}$, so it holds in every norm. A route to stronger sensitivity estimates would require additional regularity. Even in Euclidean geometry, differentiating the unsquared objective introduces normalization by the blend norm near stationarity; general norms may also have nonzero nonsmooth points.

Lemma F.5 (KL perturbation of the first argument)**.** *Let* $p, p', q \in \Delta_K^\kappa$ *and* $\Lambda := \ln(1/\kappa)$. *Then*

$$\mathrm{KL}(p'\|q) \le \mathrm{KL}(p\|q) + \sqrt{(2+\Lambda)\,\mathrm{KL}(p'\|q)}\cdot\sqrt{\chi^2(p\|p')}.$$

Proof. Fix q and let $F(\cdot) := \mathrm{KL}(\cdot\|q)$, which is convex with $\nabla F(p')_i = \ln(p'_i/q_i)+1$. Convexity at p' gives $F(p') - F(p) \le \langle\nabla F(p'), p' - p\rangle = \sum_i \ln(p'_i/q_i)(p'_i - p_i)$, the constant dropping because $\sum_i(p'_i - p_i) = 0$. Cauchy–Schwarz with the splitting $\sqrt{p'_i}\cdot 1/\sqrt{p'_i}$ yields

$$\sum_i \ln\frac{p'_i}{q_i}(p'_i - p_i) \le \Big(\sum_i p'_i \ln^2\frac{p'_i}{q_i}\Big)^{1/2}\Big(\sum_i \frac{(p'_i - p_i)^2}{p'_i}\Big)^{1/2} = \sqrt{V(p'\|q)}\sqrt{\chi^2(p\|p')}.$$

It remains to show $V(p'\|q) \le (2+\Lambda)\,\mathrm{KL}(p'\|q)$. Put $r_i := p'_i/q_i$; since $p', q \in \Delta_K^\kappa$ we have $r_i \in [\kappa, 1/\kappa]$, so $|\ln r_i| \le \Lambda$. Writing $\psi(r) := r\ln^2 r$ and $d(r) := r\ln r - r + 1 \ge 0$, we have $V(p'\|q) = \sum_i q_i\psi(r_i)$ and, because $\sum_i q_i(r_i - 1) = 0$, also $\mathrm{KL}(p'\|q) = \sum_i q_i d(r_i)$. So it suffices that $\psi(r) \le (2+\Lambda)d(r)$ whenever $|\ln r| \le \Lambda$. Substituting $r = e^s$, this is

$$e^s s^2 \le (2+\Lambda)\,(se^s - e^s + 1) \qquad \text{for } |s| \le \Lambda, \tag{30}$$

where $se^s - e^s + 1 \ge 0$ with equality only at $s = 0$ (both sides of (30) vanish there). We prove the stronger statement $e^s s^2 \le (2 + \max(s,0))(se^s - e^s + 1)$ for all real s, which implies (30) since $\max(s,0) \le \Lambda$ on $[-\Lambda, \Lambda]$.

Case $s \ge 0$. Let $h(s) := (2+s)(se^s - e^s + 1) - e^s s^2$. Dividing by $e^s > 0$, $e^{-s}h(s) = (2+s)(s-1+e^{-s}) - s^2 = s - 2 + (2+s)e^{-s} =: k(s)$. Then $k(0) = 0$ and $k'(s) = 1-(1+s)e^{-s} \ge 0$ because $1 + s \le e^s$. Hence $k \ge 0$ on $[0,\infty)$, i.e. $h \ge 0$.

Case $s \le 0$. Put $u := -s \ge 0$. The claim $e^s s^2 \le 2(se^s - e^s + 1)$ reads $e^{-u}u^2 \le 2\big(1-(1+u)e^{-u}\big)$, i.e. $u^2 + 2u + 2 \le 2e^u$, which holds because $2e^u = 2 + 2u + u^2 + \sum_{j\ge 3} 2u^j/j! \ge u^2 + 2u + 2$.

Hence $e^s s^2 \le (2 + \max(s, 0))(se^s - e^s + 1)$ for all s, which gives (30) and therefore $V(p'\|q) \le (2 + \Lambda)\,\mathrm{KL}(p'\|q)$. □

F.4 The recursion, the bias, and the proof of Theorem 5.1

Lemma F.6 (Tracking recursion). *Assume $\rho = \bar\eta\tau \le 1$ and $\max_i \|\nabla f_i(x_t)\|_* \le M$. Then*

$$\Phi_{t+1} \;\le\; \Big(1 - \frac{\rho}{4}\Big)\Phi_t + 4\bar\eta^2 M^2 + \frac{2(2+\Lambda)}{\rho}\,\chi^2(p_t\|p_{t+1}).$$

Proof. Apply Lemma F.5 with $p = p_t$, $p' = p_{t+1}$, $q = w_{t+1}$ (all in Δ_K^κ: w_{t+1} by construction, p_t, p_{t+1} by definition): $\Phi_{t+1} \le \mathrm{KL}(p_t\|w_{t+1}) + \sqrt{(2+\Lambda)\Phi_{t+1}}\sqrt{\chi^2}$. Young's inequality $ab \le \frac{\theta}{2}a^2 + \frac{1}{2\theta}b^2$ with $a = \sqrt{\Phi_{t+1}}$, $b = \sqrt{(2+\Lambda)\chi^2}$, $\theta = \rho/2$ bounds the last term by $\frac{\rho}{4}\Phi_{t+1} + (2+\Lambda)\chi^2/\rho$. Using Lemma F.3 for $\mathrm{KL}(p_t\|w_{t+1})$ and rearranging,

$$\Phi_{t+1} \le \frac{1}{1-\rho/4}\Big[\frac{\Phi_t + 2\bar\eta^2 M^2}{1+\rho} + \frac{(2+\Lambda)\chi^2}{\rho}\Big].$$

For $\rho \in (0, 1]$, $(1-\rho/4)(1+\rho) = 1 + \frac{3}{4}\rho - \frac{1}{4}\rho^2 \ge 1 + \frac{1}{2}\rho$, so $[(1-\rho/4)(1+\rho)]^{-1} \le (1+\rho/2)^{-1} \le 1 - \rho/4$, and $(1-\rho/4)^{-1} \le 2$. Substituting gives the claim. □

Lemma F.7 (Bias of the regularized target). *Suppose $\max_i \|\nabla f_i(x_t)\|_* \le M$ and $\kappa \le 1/K$. Then $f_t(p_t) - \min_{\Delta_K} f_t \le 2MK\kappa + \tau \ln K$.*

Proof. Let $w_t^\star \in \arg\min_{\Delta_K} f_t$ and $\hat w := (1 - K\kappa)w_t^\star + K\kappa u_K$. Then $\hat w_i \ge \kappa$ and $\hat w \in \Delta_K$, so $\hat w \in \Delta_K^\kappa$; $\|\hat w - w_t^\star\|_1 = K\kappa\|u_K - w_t^\star\|_1 \le 2K\kappa$; and $r(\hat w) \le (1 - K\kappa)r(w_t^\star) \le \ln K$ by convexity of r, $r(u_K) = 0$ and $r \le \ln K$. Optimality of p_t and $r \ge 0$ give $f_t(p_t) \le \tilde\psi_t(p_t) \le \tilde\psi_t(\hat w) \le f_t(\hat w) + \tau \ln K \le f_t(w_t^\star) + M\|\hat w - w_t^\star\|_1 + \tau \ln K$, by (F1). □

Proof of Theorem 5.1. We first dispose of the trivial regime: if $\epsilon > 2M$ then $f_t(w) \le M < \epsilon/2$ for every $w \in \Delta_K$, so $\mathrm{gap}_t \le M < \epsilon$ and $\|\nabla F(x_t)w_t\|_* \le M < \epsilon$ for every t. Assume $\epsilon \le 2M$, so also $\beta \le 2M$. Then $\kappa = \beta/(4MK) \le 1/(2K) < 1/K$, so Δ_K^κ has non-empty relative interior; and

$$\rho = \bar\eta\tau = \frac{\epsilon^2\beta^2}{1536\,M^4 \ln^2 K} \le \frac{16}{1536 \ln^2 2} < 1,$$

so Lemma F.6 applies. Note $\Lambda = \ln(1/\kappa) = \ln(4MK/\beta)$.

(i) Tracking. Iterating Lemma F.6 with the uniform bound $\chi^2(p_t\|p_{t+1}) \le 2L\bar\gamma/(\tau\kappa)$ of Lemma F.4, from $\Phi_0 = \mathrm{KL}(p_0\|u_K) \le \ln K$, and summing the geometric series $\sum_{j\ge 0}(1-\rho/4)^j = 4/\rho$,

$$\Phi_t \le \Big(1 - \frac{\rho}{4}\Big)^t \ln K + \frac{16\bar\eta^2 M^2}{\rho} + \frac{8(2+\Lambda)}{\rho^2}\cdot\frac{2L\bar\gamma}{\tau\kappa} = \Big(1 - \frac{\rho}{4}\Big)^t \ln K + \frac{16\bar\eta M^2}{\tau} + \frac{16(2+\Lambda)L\bar\gamma}{\bar\eta^2\tau^3\kappa}. \tag{31}$$

We show each of the three terms is at most $\epsilon^2/(24M^2)$ for $t \ge T_0$. The second equals $16M^2 \cdot \frac{\epsilon^2\beta}{768M^4 \ln K}\cdot\frac{2\ln K}{\beta} = \epsilon^2/(24M^2)$ exactly. For the third, substitute $\bar\eta^2 = \epsilon^4\beta^2/(768^2 M^8 \ln^2 K)$, $\tau^3 = \beta^3/(8\ln^3 K)$ and $\kappa = \beta/(4MK)$, so that $\bar\eta^2\tau^3\kappa = \epsilon^4\beta^6/(9\cdot 2^{21} M^9 K \ln^5 K)$ (as $768^2\cdot 8\cdot 4 = 9\cdot 2^{16}\cdot 2^5$); then

$$\frac{16(2+\Lambda)L\bar\gamma}{\bar\eta^2\tau^3\kappa} = \frac{9\cdot 2^{25}\,(2+\Lambda)LM^9 K \ln^5 K}{\epsilon^4\beta^6}\,\bar\gamma = \frac{9\cdot 2^{25}\,\epsilon^2}{27\cdot 2^{28}M^2} = \frac{\epsilon^2}{24M^2}$$

exactly, by the choice of $\bar\gamma$. For the first, $(1-\rho/4)^t \le e^{-\rho t/4} \le \epsilon^2/(24M^2 \ln K)$ once $t \ge \frac{4}{\rho}\ln\frac{24M^2\ln K}{\epsilon^2}$, and $4/\rho = 6144M^4 \ln^2 K/(\epsilon^2\beta^2)$, so this holds for $t \ge T_0$. Hence $\Phi_t \le \epsilon^2/(8M^2)$ for $t \ge T_0$, and Pinsker gives $\|w_t - p_t\|_1 \le \sqrt{2\Phi_t} \le \epsilon/(2M)$. Lemma F.7 with $2MK\kappa = \beta/2$

and $\tau \ln K = \beta/2$ gives $\mathrm{gap}(p_t) \le \beta$. Finally, by (F1), $\mathrm{gap}_t = \big(f_t(w_t) - f_t(p_t)\big) + \mathrm{gap}(p_t) \le M\|w_t - p_t\|_1 + \beta \le \epsilon/2 + \beta$.

(ii) Progress. The bound $\frac{1}{T}\sum_{t<T} \|\nabla F(x_t) w_t\|_* \le \Delta/(\bar\gamma T) + \bar\eta M^2/2 + L\bar\gamma/2$ is (7) of Theorem 4.4, which holds for every $\tau \ge 0$, $\kappa \le 1/K$ and every γ, η. It remains to check the three terms at the stated parameters. $\Delta/(\bar\gamma T) \le \epsilon/4$ for $T \ge T_1 = \lceil 4\Delta/(\bar\gamma\epsilon)\rceil$. Next $\bar\eta M^2/2 = \epsilon^2\beta/(1536 M^2 \ln K) \le \epsilon/8$, because $\epsilon\beta/M^2 \le \epsilon^2/M^2 \le 4 \le 192 \ln 2 \le 192 \ln K$. Finally $L\bar\gamma/2 \le \epsilon/8$ because, using $(2+\Lambda)K \ge 1$, $\ln K \ge \ln 2$ and $\beta \le \epsilon \le 2M$, $L\bar\gamma = \epsilon^6\beta^6/(27 \cdot 2^{28}(2+\Lambda) M^{11} K \ln^5 K) \le \epsilon\,(\epsilon/M)^{11}/(27 \cdot 2^{28} \ln^5 2) \le \epsilon\, 2^{11}/(27 \cdot 2^{28} \ln^5 2) < \epsilon/4$. Hence the average is at most $\epsilon/4 + \epsilon/8 + \epsilon/8 = \epsilon/2$ for $T \ge T_1$.

(iii) Tail output. Let $T \ge \max\{2T_0, T_1\}$. By (ii), $\sum_{t=T_0}^{T-1} \|\nabla F(x_t) w_t\|_* \le \sum_{t=0}^{T-1} \|\nabla F(x_t) w_t\|_* \le T\epsilon/2$, the first inequality by non-negativity of the dropped terms. The minimum over the tail is at most its average, so $\|\nabla F(x_{\hat t}) w_{\hat t}\|_* \le (T\epsilon/2)/(T - T_0) \le \epsilon$, using $T - T_0 \ge T/2$. Since $\hat t \ge T_0$, part (i) gives $\mathrm{gap}_{\hat t} \le \epsilon/2 + \beta$, and $\min_w \|\nabla F(x_{\hat t}) w\|_* \le \|\nabla F(x_{\hat t}) w_{\hat t}\|_* \le \epsilon$. With $\beta = \epsilon/2$ the parameters reduce to $\tau = \epsilon/(4 \ln K)$, $\kappa = \epsilon/(8MK)$, $\bar\eta = \epsilon^3/(1536 M^4 \ln K)$, $\bar\gamma = \epsilon^{12}/(27 \cdot 2^{34}(2+\Lambda) L M^{11} K \ln^5 K)$ and $T_0 = \lceil (24576 M^4 \ln^2 K/\epsilon^4) \ln(24 M^2 \ln K/\epsilon^2)\rceil$, which gives the stated orders. □

Proof of Corollary 5.2. By the triangle inequality, $\mathrm{CA}_t \le \|\nabla F(x_t)(w_t - p_t)\|_* + \|\nabla F(x_t)(p_t - w^\star(x_t))\|_*$. The first term is at most $M\|w_t - p_t\|_1 \le \epsilon/2$ for $t \ge T_0$ by (F1) and Theorem 5.1(i). For the second, $w^\star(x_t)$ minimizes $\varphi_q \circ \nabla F(x_t)$ over Δ_K (the same minimizers as f_t), so the argument of Remark C.1 applied with w_t replaced by p_t gives $\|\nabla F(x_t)(p_t - w^\star(x_t))\|_*^q \le (q/\mu) M^{q-1} \mathrm{gap}(p_t) \le (q/\mu) M^{q-1}\beta$, which equals $(\epsilon/2)^q$ for $\beta = \mu\epsilon^q/(q 2^q M^{q-1})$. The orders of T_0 and T_1 follow by substituting this β into Theorem 5.1: in the Euclidean case $\beta = \epsilon^2/(8M)$, so $\epsilon^2\beta^2 = \epsilon^6/(64M^2)$ and $\epsilon^6\beta^6 = \epsilon^{18}/(2^{18} M^6)$. □

Remark F.2 (On the gradient bound in Theorem 5.1). Theorem 5.1 takes the trajectory gradient bound as a hypothesis, and we keep it as one: unlike Theorem 4.4 at $\tau = \kappa = 0$, we do not derive it. We record what a derivation would involve. The induction of Lemma 4.3 changes in the comparator: Lemma 4.3 uses $u = e_i$, which is not in Δ_K^κ, so one applies (6) at $\hat e_i = (1 - K\kappa)e_i + K\kappa u_K \in \Delta_K^\kappa$ and pays $\gamma\|g_s\|_\infty \|e_i - \hat e_i\|_1 \le 2\bar\gamma MK\kappa$ per step for the difference, together with the regularization term $\rho \ln K/\eta \cdot \gamma = \bar\gamma\tau \ln K$ per step from Lemma 4.2. Since $\tau \ln K + 2MK\kappa = \beta \le \epsilon$ and $T_1\bar\gamma \le 4\Delta/\epsilon + \bar\gamma$, over T_1 steps these add at most $T_1\bar\gamma\,\beta \le 4\Delta + \bar\gamma\epsilon$ to the function-value increase, so $f_i(x_t) - f_i^\star \le 5\Delta + C'$ with C' collecting $\bar\gamma\epsilon$ and the (much smaller) remaining terms of Lemma 4.3 at these step sizes, and Lemma A.2 converts this into $\|\nabla f_i(x_t)\|_* \le \sqrt{2L(5\Delta + C')}$, which is below the M of Lemma 4.3 ($\ge 4\sqrt{L\Delta}$) once $C' \le 3\Delta$. Verifying $C' \le 3\Delta$ requires relating the remaining terms, which involve M, ϵ and β but not Δ, to Δ through the explicit value of M in Lemma 4.3 and the horizon actually run, which is $\max\{2T_0, T_1\}$ rather than T_1. We have not carried this out, so the gradient bound remains an assumption of Theorem 5.1.

G Uniform convexity of ℓ_p and Schatten-p norms

Corollaries 4.6 and 5.2 convert a bound on the inner gap into a bound on the CA distance whenever the dual norm satisfies the following uniform convexity inequality of power type $q \ge 2$ with constant $\mu \in (0, 1]$: with $\varphi_q := \frac{1}{q}\|\cdot\|_*^q$,

$$\tfrac{1}{q}\|y\|_*^q \;\ge\; \tfrac{1}{q}\|y'\|_*^q + \langle \nabla\varphi_q(y'), y - y'\rangle + \tfrac{\mu}{q}\|y - y'\|_*^q \qquad \text{for all } y, y'. \tag{S2}$$

(The conversion itself is in Remark C.1.) This appendix records which ℓ_p and Schatten-p norms satisfy (S2), and with which (q, μ).

Proposition G.1. *Let $\|\cdot\|_*$ be the ℓ_p norm on $\mathbb{R}^m$ or the Schatten-p norm on $\mathbb{R}^{m\times n}$.*

(i) For $1 < p \le 2$, (S2) *holds with $q = 2$ and $\mu = p - 1$: the function $\psi_p := \frac{1}{2}\|\cdot\|_*^2$ is $(p-1)$-strongly convex with respect to $\|\cdot\|_*$.*

(ii) For $2 \le p < \infty$, (S2) *holds with $q = p$ and $\mu = 2^{1-p}$.*

(iii) For $p > 2$, (S2) *fails with $q = 2$ for every $\mu > 0$: the ratio $D_{\psi_p}(y, x)/\|y - x\|_*^2$ can be made arbitrarily small.*

For $p = 1$ (the ℓ_1 and nuclear norms) the unit ball is not strictly convex and no inequality of the form (S2) *can hold.*

Proof. (i) Theorem 1 of Ball et al. (1994) (optimal 2-uniform convexity) states that for $1 \le p \le 2$ and all X, Y,

$$\Big(\frac{\|X+Y\|_*^p + \|X-Y\|_*^p}{2}\Big)^{2/p} \;\ge\; \|X\|_*^2 + (p-1)\|Y\|_*^2, \tag{32}$$

both for ℓ_p (their case (a), functions in L_p) and for Schatten-p (their case (b), matrices in C_p). Put $X = \frac{a+b}{2}$ and $Y = \frac{a-b}{2}$, so that $X + Y = a$ and $X - Y = b$. Because $2/p \ge 1$, the map $s \mapsto s^{2/p}$ is convex on $[0, \infty)$ and Jensen's inequality gives $\big(\frac{\|a\|_*^p + \|b\|_*^p}{2}\big)^{2/p} \le \frac{1}{2}\|a\|_*^2 + \frac{1}{2}\|b\|_*^2$. Combining with (32),

$$\Big\|\frac{a+b}{2}\Big\|_*^2 + (p-1)\Big\|\frac{a-b}{2}\Big\|_*^2 \;\le\; \tfrac{1}{2}\|a\|_*^2 + \tfrac{1}{2}\|b\|_*^2 \qquad \text{for all } a, b.$$

Multiplying by $\frac{1}{2}$ gives $\psi_p(\frac{a+b}{2}) \le \frac{1}{2}\psi_p(a) + \frac{1}{2}\psi_p(b) - \frac{p-1}{8}\|a-b\|_*^2$, which is midpoint strong convexity with modulus $p - 1$; for continuous functions this upgrades to strong convexity by the usual dyadic-subdivision argument, and strong convexity of ψ_p is (S2) with $q = 2$, $\mu = p - 1$.

(ii) Let $\phi := \|\cdot\|_*^p$. Clarkson's inequality for $p \ge 2$ states that for all y, y',

$$\phi\Big(\frac{y+y'}{2}\Big) + \phi\Big(\frac{y-y'}{2}\Big) \le \frac{\phi(y) + \phi(y')}{2}. \tag{33}$$

For ℓ_p this follows coordinatewise from the scalar case, which is $X^{p/2} + Y^{p/2} \le (X+Y)^{p/2} \le \frac{1}{2}(|a|^p + |b|^p)$ for $X = (\frac{a+b}{2})^2$, $Y = (\frac{a-b}{2})^2$, $X+Y = \frac{1}{2}(a^2+b^2)$, using superadditivity of $s \mapsto s^{p/2}$ on $[0, \infty)$ (as $p/2 \ge 1$) and then convexity of $s \mapsto s^{p/2}$; for Schatten norms (33) is McCarthy's extension of Clarkson's inequality to trace classes (McCarthy, 1967). Now let $m := (y + y')/2$. Convexity of ϕ gives $\langle \nabla\phi(y'), m - y'\rangle \le \phi(m) - \phi(y')$, and $y - y' = 2(m - y')$, so

$$\begin{aligned} D_\phi(y, y') &:= \phi(y) - \phi(y') - \langle \nabla\phi(y'), y - y'\rangle \ge \phi(y) - \phi(y') - 2\big(\phi(m) - \phi(y')\big) \\ &= \phi(y) + \phi(y') - 2\phi(m) \ge 2\phi\Big(\frac{y-y'}{2}\Big) = 2^{1-p}\|y - y'\|_*^p, \end{aligned}$$

the last inequality being (33). Dividing by p gives (S2) for $\varphi_p = \phi/p$ with $q = p$ and $\mu = 2^{1-p}$.

(iii) Take $x = e_1$ and $y = e_1 + te_2$ with $t \downarrow 0$ (for Schatten norms, $x = E_{11}$, $y = E_{11} + tE_{22}$). Since $\nabla\psi_p(g) = \|g\|_p^{2-p}\big(\operatorname{sign}(g_i)|g_i|^{p-1}\big)_i$ we have $\nabla\psi_p(e_1) = e_1$, hence $\langle \nabla\psi_p(x), y - x\rangle = 0$ and $D_{\psi_p}(y, x) = \frac{1}{2}(1 + t^p)^{2/p} - \frac{1}{2} = \frac{1}{p}t^p + O(t^{2p})$ while $\|y - x\|_p^2 = t^2$, so the ratio is $\frac{1}{p}t^{p-2} + O(t^{2p-2}) \to 0$ for $p > 2$. Finally, if the unit ball is not strictly convex there are $y \ne y'$ on the unit sphere with $\|(y + y')/2\|_* = 1$, and (S2) applied at y and y' around the midpoint forces $\mu = 0$. □

The constant $\mu = 2^{1-p}$ in (ii) is not sharp (numerically the best constant for ℓ_p is about 0.59, 0.33 and 0.10 at $p = 3, 4, 6$, against 0.25, 0.125 and 0.031), but it is explicit. Statement (iii) is why the power type $q = p$, rather than 2, appears in Corollary 5.2 for $p > 2$: the unit sphere of ℓ_p is flatter than quadratic at the coordinate points.

H INEXACT ORACLES

The analysis uses the oracle only through $\|D_t\| \le R$ and $\langle G_t, D_t\rangle = -R\|G_t\|_*$. In the spectral case the polar factor is computed by a Newton–Schulz iteration and is only approximate. Suppose instead that for some $\vartheta \in [0, 1)$ and all t,

$$\|D_t\| \le (1 + \vartheta)R, \qquad \langle G_t, D_t\rangle \le -(1 - \vartheta)R\,\|G_t\|_*, \tag{34}$$

and that $\max_i \|\nabla f_i(x_t)\|_* \le M$ along the trajectory (we do not re-derive M here). Then $\|g_t\|_\infty \le (1 + \vartheta)RM$, and the two places where the oracle identity enters the deterministic analysis change as follows.

Proposition H.1. *Under Assumptions 1–2 and* (34)*, with* $\tau = \kappa = 0$*,* $\gamma R = \sqrt{2\Delta/(LT)}$ *and* $\eta R = \sqrt{2\ln K}/(M\sqrt{T})$*:*

$$\frac{1}{T}\sum_{t<T}\|\nabla F(x_t)w_t\|_* \le \frac{(1+\vartheta)^2}{1-\vartheta}\cdot\frac{\sqrt{2\Delta L}+M\sqrt{\ln K/2}}{\sqrt{T}},$$

$$\frac{1}{T}\sum_{t<T}\big[\|\nabla F(x_t)w_t\|_* - \|\nabla F(x_t)u\|_*\big] \le \frac{(1+\vartheta)^2}{1-\vartheta}M\sqrt{\frac{2\ln K}{T}} + \frac{2\vartheta M}{1-\vartheta} \qquad \textit{for every } u\in\Delta_K.$$

Proof. One-step inequality. The proof of Lemma 4.1 gives $F(x_{t+1})w \le F(x_t)w + \gamma\langle A_t w, D_t\rangle + \frac{1}{2}L\gamma^2\|D_t\|^2$; with (34), $\langle A_t w, D_t\rangle = \langle G_t, D_t\rangle + \langle g_t, w - w_t\rangle \le -(1-\vartheta)R\|G_t\|_* + \langle g_t, w-w_t\rangle$ and $\|D_t\|^2 \le (1+\vartheta)^2R^2$. Repeating the proof of Theorem 4.4 with $u = u_K$ and $\|g_t\|_\infty \le (1+\vartheta)RM$,

$$(1-\vartheta)\frac{1}{T}\sum_t\|G_t\|_* \le \frac{\Delta}{\gamma RT} + (1+\vartheta)^2\frac{\eta RM^2}{2} + (1+\vartheta)^2\frac{L\gamma R}{2} \le (1+\vartheta)^2\Big[\frac{\Delta}{\gamma RT} + \frac{\eta RM^2}{2} + \frac{L\gamma R}{2}\Big],$$

and the step sizes give the first claim.

Regret. Write $\psi_t(w) = \|A_t w\|_*$. For any $u\in\Delta_K$, Hölder gives $\|A_t u\|_*\|D_t\| \ge -\langle A_t u, D_t\rangle = -\langle g_t, u\rangle$, hence $\psi_t(u) \ge -\langle g_t, u\rangle/((1+\vartheta)R)$ (trivially if the right-hand side is negative). For w_t, (34) gives $\psi_t(w_t) = \|G_t\|_* \le -\langle G_t, D_t\rangle/((1-\vartheta)R) = -\langle g_t, w_t\rangle/((1-\vartheta)R)$. Subtracting, with $a := 1/((1-\vartheta)R)$ and $b := 1/((1+\vartheta)R)$,

$$\begin{aligned}\psi_t(w_t) - \psi_t(u) &\le -a\langle g_t, w_t\rangle + b\langle g_t, u\rangle = a\langle g_t, u - w_t\rangle - (a-b)\langle g_t, u\rangle \\ &\le \frac{\langle g_t, u - w_t\rangle}{(1-\vartheta)R} + (a-b)(1+\vartheta)RM,\end{aligned}$$

using $|\langle g_t, u\rangle| \le \|g_t\|_\infty \le (1+\vartheta)RM$. Since $(a-b)(1+\vartheta)RM = M(1+\vartheta)\big[\frac{1}{1-\vartheta} - \frac{1}{1+\vartheta}\big] = \frac{2\vartheta M}{1-\vartheta}$, summing and applying Lemma 4.2 as in the proof of Theorem 4.5, with $\|g_t\|_\infty^2 \le (1+\vartheta)^2R^2M^2$, gives the second claim. □

The multiplicative factor $(1+\vartheta)^2/(1-\vartheta)$ is controlled when ϑ is bounded away from one. The second inequality has an additive floor because of the mismatch between the upper and lower oracle constants. This is a limitation of that comparator argument, not an unavoidable floor for the average gap or regret along the descent trajectory. Indeed, if B_T denotes the first bound in Proposition H.1, (3) gives

$$\frac{1}{T}\sum_t \mathrm{gap}_t \le B_T,$$

$$\frac{1}{T}\left[\sum_t\|A_t w_t\|_* - \min_{u\in\Delta_K}\sum_t\|A_t u\|_*\right] \le B_T = O(T^{-1/2}).$$

Thus fixed relative oracle error satisfying (34) is already tolerated for deterministic stationarity and average inner optimality.

We do not extend this proposition to the stochastic or iteration-wise results. There, oracle error must be propagated through the tracking and subgradient arguments. Moreover, the error parameters of a practical Newton–Schulz implementation depend on its polynomial, normalization, coefficient schedule, numerical precision, and input spectrum. We have not established a uniform $\vartheta < 1$ for the five-step quintic implementation used in the experiments, nor a geometric error-decay claim for that particular implementation. Relating it to the oracle assumptions and extending the stochastic and iteration-wise guarantees remain open.

I Experimental details

I.1 Cityscapes and NYU-v2

Protocol. We follow the protocol of Navon et al. (2022); Liu et al. (2023): an MTAN model on a SegNet backbone, with 200 epochs, and batch sizes 8 on Cityscapes and 2 on NYU-v2. The standard

baselines use Adam with learning rate 10^{-4}, which is multiplied by 0.5 after epoch 100. Our method applies the oracle direction directly with step size γ, while MOON uses the released Muon+AdamW optimizer. Reported metrics are averaged over the final 10 epochs with one seed; these epochs are correlated and do not estimate variability across independent runs. The baseline rows are re-runs with the same protocol. Cityscapes has $K = 2$ tasks (7-class semantic segmentation, depth); NYU-v2 has $K = 3$ (13-class segmentation, depth, surface normals).

Metrics. Segmentation: mean IoU and pixel accuracy (higher is better). Depth: absolute and relative error (lower is better). Normals: mean and median angular error (lower is better) and the fraction of pixels within $11.25°$, $22.5°$ and $30°$ (higher is better). The aggregate is

$$\Delta m\% = \frac{100}{n} \sum_{j=1}^{n} (-1)^{s_j} \frac{M_j - M_{\mathrm{STL},j}}{M_{\mathrm{STL},j}},$$

over the n metrics, with $s_j = 1$ if higher is better and 0 otherwise. For each dataset, M_{STL} is the corresponding single-task baseline reported in the STL row of the same table. Thus, the STL reference is $(74.11, 93.44, 0.0137, 38.22)$ on Cityscapes and $(38.77, 63.98, 0.6350, 0.2654, 25.10, 19.06, 30.12, 57.66, 69.53)$ on NYU-v2. Accordingly, the STL row has $\Delta m = 0$, and all other Δm values are computed relative to these references.

Our method. We use Algorithm 2 with the spectral oracle on convolutional and linear weights (4D kernels flattened over their last three dimensions) and the Euclidean oracle on the remaining parameters. Each matrix-valued parameter tensor is treated as a unit-radius spectral-norm block, $\rho_\ell = 1$, without width- or kernel-dependent rescaling, while all scalar and vector parameters are concatenated into a single unit-radius Euclidean block. The outer model step size controls the overall update magnitude. We use seed 42 for all dense-task experiments, with momentum $\alpha = 0.1$, $\tau = 0$, and $\kappa = 0$ fixed throughout.

Hyperparameter tuning uses the same non-training benchmark split on which the dense-task results are reported: the official 500-image validation split for Cityscapes and the 654-image evaluation (test) split for NYU-v2. Thus, no additional held-out validation split is used for these dense-task experiments. Short-horizon segmentation performance is used as the primary tuning criterion, subject to maintaining joint-task performance, while the final late-training schedule is selected using the final-window Δm. For Cityscapes, we search $\gamma \in \{10^{-3}, 5 \times 10^{-3}\}$ and $\eta \in \{10^{-4}, 5 \times 10^{-4}\}$, selecting $\gamma = 5 \times 10^{-3}$ and $\eta = 10^{-4}$. The final model-step schedule uses 5×10^{-3} for epochs 1–100, 2.5×10^{-3} for epochs 101–150, and 1.25×10^{-3} for epochs 151–200. For NYU-v2, we search $\gamma \in \{2.5 \times 10^{-3}, 5 \times 10^{-3}\}$ and $\eta \in \{5 \times 10^{-5}, 10^{-4}\}$, selecting $\gamma = 2.5 \times 10^{-3}$ and $\eta = 10^{-4}$; the model step is reduced to 1.25×10^{-3} from epoch 101 onward. Standard baselines use their reference hyperparameters without an additional tuning grid.

I.2 MULTIMNIST

Protocol. Following (Zhou et al., 2026), we use MultiMNIST, where each example contains two MNIST digits and the two tasks are to classify the left and right digits. The digits are concatenated horizontally and, with probability 0.5, overlapped by up to 30% of their width. The resulting images are resized to 28×43. We generate 10,000 training examples from MNIST-train and 1,000 test examples from MNIST-test. We use a small Vision Transformer with patch size 4 and embedding dimension 16, consisting of a class token, learned positional embeddings, and 3 transformer encoder layers, each with 2 attention heads and a feed-forward dimension of 64. No dropout is used, and the shared encoder is followed by one linear classification head for each task. All methods in Table 4 are evaluated on the same generated dataset and trained for 100 epochs with batch size 256.

Our method. Algorithm 2 with the unregularized, unclipped specialization of the weight update in Algorithm 1. At each iteration, we compute per-task gradients for all trainable model parameters and form their weighted aggregate. The oracle comes from the previous momentum buffer. For every parameter tensor with at least two dimensions, we flatten all but the first dimension and apply five Newton–Schulz steps. If the resulting matrix is taller than wide, we first transpose it. Writing G for this (possibly transposed) matrix, the iteration is carried out in `float32`, initialized as $X_0 =$

$G/(\|G\|_F + 10^{-7})$, and uses

$$X_{r+1} = aX_r + b(X_rX_r^\top)X_r + c(X_rX_r^\top)^2X_r, \qquad (a, b, c) = (3.4445, -4.7750, 2.0315),$$

for five steps. Each matrix-valued parameter tensor is treated as a unit-radius spectral-norm block, without width- or kernel-dependent rescaling, while all scalar and vector parameters are concatenated into a single unit-radius Euclidean block. The blended momentum buffer is initialized at zero; its first update uses $\alpha_0 = 1$, and subsequent updates use $\alpha = 0.1$. We use model step size $\gamma = 10^{-3}$, weight step $\eta = 5 \times 10^{-5}$, $\tau = 0$, and $\kappa = 0$.

MOON. We use (Zhou et al., 2026)'s released implementation: the task logits are updated with Adam at learning rate 10^{-4} and weight decay 10^{-3}, Muon uses Nesterov momentum, and tasks are scored using the previous optimizer update. The data, model, number of epochs, batch size and numerical model learning rate are shared with our method. Momentum conventions, logit optimizers, weight decay, and auxiliary parameter updates differ; a common numerical learning rate need not imply a common effective update size. The comparison therefore evaluates complete implementations rather than isolating a single algorithmic component.

Other baselines. MGDA and FAMO use Adam, as in MOON's experiments; "+ Muon" replaces Adam on the encoder matrices by Muon while keeping the baseline's task weights, and Muon (LS) uses Muon with equal weights. In all MultiMNIST runs the model learning rate is 10^{-3}; MGDA uses its minimum-norm task weights, FAMO uses weight learning rate 0.025 and $\gamma = 0.01$, and the Muon variants use momentum 0.95, five Newton–Schulz steps, RMS matching 0.2, and weight decay 10^{-3}, with auxiliary Adam parameters using $(\beta_1, \beta_2) = (0.9, 0.95)$.

Tuning and reporting. All hyperparameter selection is performed on a fixed 10% validation split drawn from MNIST-train using split seed 2026. The weight-step sweep evaluates $\eta \in \{10^{-5}, 5 \times 10^{-5}, 10^{-4}, 5 \times 10^{-4}, 10^{-3}\}$ (Table 5), together with the other candidate configurations and schedules considered during tuning. Final candidates are evaluated over seeds 42, 43, and 44 using the mean validation accuracy over epochs 96–100. After all hyperparameters are frozen, the selected configuration is evaluated on MNIST-test.

Ablations. Each ablation changes one component of our method and keeps the others, including the step sizes tuned for the spectral configuration, on three (42, 43, and 44) seeds. These are fixed-hyperparameter substitutions, not comparisons of separately tuned variants. In particular, the oracle results do not establish superiority over a retuned Euclidean oracle, and the weight results do not establish equivalence of entropic and projected updates. For the oracle ablations, the sign variant uses the unit-radius ℓ_∞ oracle $\mathrm{sign}(M)$, while the ℓ_2 variant uses the unit-radius Euclidean oracle $M/\|M\|_2$ on the full parameter vector. The spectral configuration follows the product geometry of the main method: each matrix-valued parameter tensor forms a unit-radius spectral-norm block, while all scalar and vector parameters are concatenated into a single unit-radius Euclidean block. No additional RMS matching or 0.2 rescaling is applied to these oracle variants.

Weights: "projected" replaces the entropic step by the Euclidean projection of $w_t + \eta g_t$ onto Δ_K^κ. *Momentum*: "per-task" keeps one buffer $m_{t,i}$ per task, evaluates the oracle at $\sum_i w_{t,i} m_{t,i}$ and scores with $\langle D_t, m_{t,i}\rangle$ (Appendix E); "none" evaluates the oracle at the current blended minibatch gradient. *Step* η: the grid above, all other settings fixed.

η	10^{-5}	5×10^{-5}	10^{-4}	5×10^{-4}	10^{-3}
Average accuracy (%)	90.20	**90.54**	90.08	89.69	89.53
Final inner gap	1.97	0.90	0.84	0.29	1.04

Table 5: MultiMNIST: sensitivity of our method to the weight step η (spectral oracle, entropic weights, blended momentum), test accuracy averaged over three seeds. The main configuration, $\eta = 5 \times 10^{-5}$, was selected on the validation split.

Clipping. Table 6 varies the floor κ with all other settings of the main configuration fixed ($\eta = 5 \times 10^{-5}$, $\tau = 0$). For $\kappa \leq 0.48$ the results coincide with the unclipped run to the reported precision,

κ	0	0.48	0.49	0.495	0.4975
Avg.↑	90.54	90.54	90.73	**91.11**	90.96
Gap↓	**0.90**	**0.90**	1.17	1.22	1.65

Table 6: Ablation of the clipping parameter κ on MultiMNIST: test accuracy and final inner gap, averaged over three seeds.

so the floor essentially never binds: with this weight step the weights remain within $[0.48, 0.52]$. As κ approaches $1/K = 0.5$ the floor forces the weights toward uniform; the mean accuracy rises slightly, to 91.11% at $\kappa = 0.495$, while the empirical inner gap grows from 0.90 to 1.65. Since $\kappa \to 1/K$ reduces the method to linear scalarization with the same oracle and momentum, and this configuration reaches a similar accuracy, the margin over the baselines in Table 4 cannot be attributed to the adaptive weighting at $K = 2$.

Empirical inner-gap diagnostic. At the end of training, we compute per-task gradients of all trainable parameters on a fixed probe consisting of the first 256 training examples and report the simplex Frank–Wolfe gap of

$$f(w) = \left\| \sum_i w_i \nabla \ell_i \right\|_*$$

at the final task weights. Specifically, writing $G(w) = \sum_i w_i \nabla \ell_i$ and letting $P(G(w))$ denote the exact geometry-specific polar direction, we report

$$\text{gap} = \langle G(w), P(G(w)) \rangle - \min_i \langle \nabla \ell_i, P(G(w)) \rangle .$$

For the spectral configuration, the exact diagnostic uses the product geometry of the method: each matrix-valued parameter block has unit spectral-norm radius, whose dual contribution is the nuclear norm, while all scalar and vector parameters form a single unit-ℓ_2 block. The sign oracle corresponds to a unit ℓ_∞ ball over the full parameter vector and hence an entrywise ℓ_1 dual norm. The ℓ_2 ablation uses one unit Euclidean ball over the full parameter vector, so its dual is the global ℓ_2 norm, rather than a sum of per-layer Frobenius norms. Because the underlying geometries differ, the reported gaps for different oracle variants are not on a common numerical scale. This is a fixed-minibatch diagnostic rather than the exact full-dataset inner optimization error appearing in the theoretical analysis. All reported main configurations use $\tau = \kappa = 0$ and therefore do not directly test the regularized iteration-wise tracking guarantee.

Code. An anonymous implementation of all methods and experiments is available at `https://anonymous.4open.science/r/LMO-MOO-D648/`.