BEYOND ROBUST AGGREGATION: CERTIFIED MULTI-OBJECTIVE DESCENT UNDER BYZANTINE ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Robust aggregation limits Byzantine influence but does not ensure reliable decisions in distributed multi-objective optimization: residual task-gradient errors can distort the task weights and direction selected by the multiple-gradient descent algorithm (MGDA) or falsely signal Pareto stationarity. We propose Certified Robust MGDA (CR-MGDA), which converts task-wise error bounds into a necessary-and-sufficient common-descent test for every honest gradient consistent with these bounds. When a regularized MGDA proposal fails certification, a convex recovery problem searches for an alternative certified direction; if recovery fails, CR-MGDA returns a computable upper bound on the honest Pareto-stationarity gap. We establish non-asymptotic stationarity bounds for nonconvex objectives and weighted objective-gap bounds for strongly convex objectives. For stochastic reports, local geometric median-of-means estimation supports trajectory-wide high-probability certificates and expected convergence without assuming independence between the reports and the resulting MGDA decisions. Experiments under multiple Byzantine attacks show that CR-MGDA improves descent safety while maintaining optimization progress.

1 INTRODUCTION

Distributed learning trains a shared model from data held by multiple clients or institutions (McMahan et al., 2017; Kairouz et al., 2021). Many such problems involve several objectives: different tasks or application requirements define different losses, and improving one loss may worsen another (Li et al., 2020b; Yang et al., 2023). Multi-objective optimization (MOO) therefore seeks Pareto-stationary points, at which no direction decreases all objectives to first order.

The multiple-gradient descent algorithm (MGDA) is a standard method for smooth MOO (Fliege & Svaiter, 2000; Désidéri, 2012). Rather than fixing scalarization weights in advance, MGDA chooses the minimum-norm convex combination of the task gradients. In a distributed setting where all N clients report gradients for all S tasks, mean aggregation forms $\bar{G}^t = \sum_{i\in[N]} G_i^t/N = [\bar{g}_1^t, \dots, \bar{g}_S^t]$ from the client matrices $G_i^t \in \mathbb{R}^{d\times S}$. The server solves the quadratic program (QP)

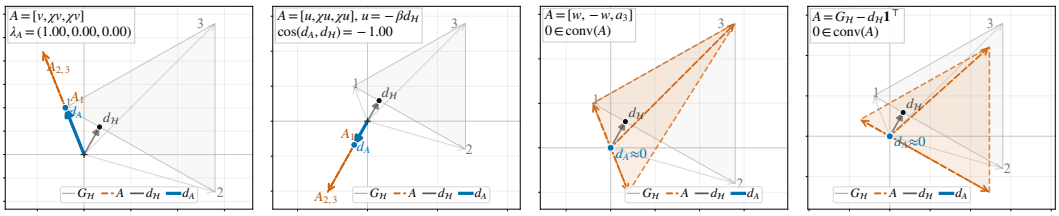
$$\bar{\lambda}^t \in \arg\min_{\lambda\in\Delta_S:=\{\lambda\in\mathbb{R}_+^S:\mathbf{1}^\top\lambda=1\}} \|\bar{G}^t\lambda\|^2, \qquad \bar{d}^t = \bar{G}^t\bar{\lambda}^t, \tag{1}$$

to select the minimum-norm point in the convex hull of the mean task gradients. Its optimality condition gives $\langle \bar{g}_s^t, \bar{d}^t\rangle \geq \|\bar{d}^t\|^2$ for every task, so $-\bar{d}^t$ is a common-descent direction. MGDA updates $x^{t+1} = x^t - \eta\bar{d}^t$ when $\bar{d}^t \neq \mathbf{0}$, with $\|\bar{d}^t\|$ measuring the Pareto-stationarity gap associated with the aggregated task gradients.

Byzantine clients may submit arbitrary reports, corrupting the aggregated task gradients (Blanchard et al., 2017; Xie et al., 2018). Unlike a single-objective gradient step, MGDA determines the task weights from the input gradients and then computes the update direction through the nonlinear mapping in equation 1. This dependence complicates the effect of gradient errors: even unbiased task-gradient estimates need not yield an unbiased MGDA direction (Fernando et al., 2023; Zhou et al., 2022; 2026). Under Byzantine attacks, residual errors after robust aggregation can therefore destroy common descent and make the resulting Pareto-stationarity assessment unreliable.

Figure 1 shows Byzantine attacks on the task-weight selection and convex-hull geometry specific to distributed MGDA. These attacks concentrate weights on a selected task, reverse the honest MGDA

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.



(a) Task-weight hijacking (b) Descent-direction reversal (c) False Pareto stationarity (d) Inlier geometry attack

Figure 1: Two-dimensional geometry of Byzantine attacks specific to distributed MGDA. In (a) and (b), the other task directions are replaced by scaled copies of an attacker-chosen direction, forcing the QP toward the target column. In (c), the opposite columns w and $-w$ place the origin in $\mathrm{conv}(A)$. In (d), translating every honest column by $-d_{\mathcal{H}}$ moves the honest MGDA point to the origin and produces the same zero-gap outcome while preserving the shape of the convex hull.

direction, or produce a zero observed gap away from honest Pareto stationarity. The last construction quantifies the perturbation sufficient for false stationarity. Let $G_{\mathcal{H}}$ denote the honest task-gradient matrix, $\lambda_{\mathcal{H}}$ an MGDA weight vector, and $d_{\mathcal{H}} = G_{\mathcal{H}}\lambda_{\mathcal{H}}$. The translated matrix $A = G_{\mathcal{H}} - d_{\mathcal{H}}\mathbf{1}^\top$ satisfies $A\lambda_{\mathcal{H}} = \mathbf{0}$ and $\|A - G_{\mathcal{H}}\|_F = \sqrt{S}\|d_{\mathcal{H}}\|$. The perturbation used to make the observed gap zero thus decreases with the honest Pareto-stationarity gap.

We propose **Certified Robust MGDA** (CR-MGDA), which combines robust task-gradient estimates with task-wise error bounds to certify common descent for the honest objectives. CR-MGDA first tests a regularized MGDA proposal and, if certification fails, solves a convex recovery problem for an alternative certified direction; if recovery does not meet its acceptance criterion, the method returns an upper bound on the honest Pareto-stationarity gap. For stochastic reports, local geometric median-of-means (GMOM) provides high-probability task-wise error bounds for the same certification and recovery framework. Our main contributions are summarized as follows:

- We characterize the vulnerability of distributed MGDA under Byzantine reports. In contrast to single-objective methods, distributed MGDA determines its task weights, update direction, and stationarity test from the entire task-gradient matrix. We show how adversarial perturbations can hijack the task weights, reverse the descent direction, or induce false stationarity.
- We propose CR-MGDA to certify MGDA decisions under task-wise gradient uncertainty. Using task-wise error balls, CR-MGDA applies an exact common-descent test to the regularized MGDA proposal and searches the same convex-combination family for an alternative certified direction when the proposal fails. If recovery does not produce an accepted direction, CR-MGDA returns an explicit upper bound on the honest Pareto-stationarity gap. For stochastic reports, local GMOM provides high-probability radii with more favorable failure-probability dependence than sample-mean bounds under the same sampling budget.
- We establish full-gradient and stochastic optimization guarantees for both non-convex and strongly convex objectives. In the non-convex setting, the non-asymptotic stationarity bounds cover proposal updates, recovery updates, and rounds paused after recovery failure. Under strong convexity and a common minimizer, we derive weighted objective-gap bounds for runs that update at every round. For stochastic reports, failure-probability budgets give trajectory-wide high-probability certificates and convergence bounds, while a separate mean-square analysis yields expected convergence without assuming independence between the reports and the resulting MGDA decisions.
- We evaluate CR-MGDA on three learning benchmarks under MGDA-specific and standard Byzantine attacks, and quantify how radius calibration affects error coverage and update availability. With empirical radii, we observe no unsafe CR-MGDA updates in the UNSW-NB15 safety evaluations, while robust-aggregation baselines still do under some attacks; recovery sustains progress after proposal failure, and local GMOM improves performance under the tested heavy-tailed noise.

2 Related Work

Multi-objective and multi-task learning. MGDA constructs a common-descent direction from multiple objective gradients (Fliege & Svaiter, 2000; Désidéri, 2012), and this formulation under-

lies multi-task learning methods that manage conflicts among task gradients (Sener & Koltun, 2018; Liu et al., 2021). Related work considers both deterministic gradient errors and stochastic sampling errors. Under bounded gradient uncertainty, Peitz & Dellnitz (2018) derive an error-aware condition for common descent. Under stochastic sampling, the weights depend on the sampled task gradients, so the resulting multi-gradient can be biased even when the individual gradient estimates are unbiased (Liu & Vicente, 2024; Zhou et al., 2022; Fernando et al., 2023). MoCo mitigates this bias through gradient correction (Fernando et al., 2023), while STIMULUS uses recursive gradient estimation to improve convergence and sample complexity (Liu et al., 2025). Unlike these settings, we consider task-gradient reports that may be replaced arbitrarily by Byzantine clients. We use post-aggregation error bounds to certify common descent, recover an alternative direction when certification fails, or bound the honest Pareto-stationarity gap.

Distributed multi-objective optimization. Federated MOO considers both trade-offs among client objectives and multiple tasks distributed across clients. FedMGDA+ treats client losses as separate objectives and promotes fairness and robustness to certain malicious manipulations through constrained combination weights and gradient normalization (Hu et al., 2022). Shen et al. (2025) balance global performance and individual client fairness. For task objectives distributed across clients, FMGDA and FSMGDA support local updates (Yang et al., 2023), FedCMOO reduces communication through compressed gradient (Askin et al., 2025), and DREAM corrects client drift and inter-task aggregation bias (Zhou et al., 2026). Unlike these methods, we allow Byzantine clients to replace task-gradient reports arbitrarily and study common descent for the honest objectives.

Byzantine-robust distributed learning. Robust aggregation methods use distance-based selection, coordinate-wise statistics, and geometric medians to limit the influence of Byzantine reports (Blanchard et al., 2017; El-Mhamdi et al., 2018; Yin et al., 2018; Minsker, 2015; Pillutla et al., 2022). Subsequent work develops attacks that remain within the variation of honest updates (Baruch et al., 2019), handles data heterogeneity (Karimireddy et al., 2022; Liu et al., 2023), and quantifies the accuracy cost of robust aggregation even without attacks (Yang et al., 2025). In distributed multi-task learning, Byzantine-resilient methods use locally evaluated losses to filter unreliable neighbors while learning distinct but related models (Li et al., 2020a). In our setting, robust aggregation is applied task-wise before MGDA, but its error bounds do not by themselves certify common descent or the validity of the resulting Pareto-stationarity assessment.

3 Certified Robust Distributed MGDA

We first formulate the MOO problem over the honest clients and the Byzantine attack model. We then present CR-MGDA for full-gradient reports and extend the framework to stochastic reports. Detailed proofs and derivations are provided in Appendices C and D.

3.1 Problem Formulation

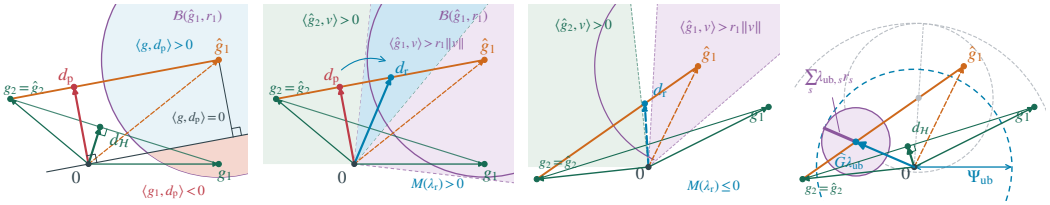
We consider N clients and S task objectives. Client i participates in $\mathcal{S}_i \subseteq [S]$, and $\mathcal{R}_s = \{i : s \in \mathcal{S}_i\} \subseteq [N]$ denotes the clients assigned to task s. Let $\mathcal{B} \subset [N]$ be the Byzantine set, and let $\mathcal{H} := [N] \setminus \mathcal{B}$ be the honest set. Since Byzantine reports are arbitrary and need not correspond to any local loss, we formulate the target as an unconstrained MOO problem over the honest clients:

$$\min_{x\in\mathbb{R}^d} F^{\mathcal{H}}(x) = [f_1^{\mathcal{H}}(x), \dots, f_S^{\mathcal{H}}(x)], \quad f_s^{\mathcal{H}}(x) = \frac{\sum_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i} f_{s,i}(x)}{\alpha_s^{\mathcal{H}} := \sum_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i}}, \tag{2}$$

where $\alpha_{s,i} \geq 0$ and $\sum_{i\in\mathcal{R}_s} \alpha_{s,i} = 1$ are task-wise client weights. We set $\alpha_{s,i}$ as the proportion of training samples held by client i among all clients assigned to task s. The honest objectives in equation 2 define the optimization target, although the server does not know $\mathcal{H}$.

Byzantine reports. At each round t, the server broadcasts x^t, and each client returns one report for every task in which it participates. Under full-gradient reporting, an honest client $i \in \mathcal{R}_s \cap \mathcal{H}$ submits $u_{s,i}^t = \nabla f_{s,i}(x^t)$, whereas a Byzantine client $i \in \mathcal{R}_s \cap \mathcal{B}$ may submit an arbitrary vector $u_{s,i}^{\star,t} \in \mathbb{R}^d$, possibly varying across rounds and tasks (Wu et al., 2020; Zuo et al., 2025). This model includes standard Byzantine attacks such as scaling, sign flipping, and noise injection, as well as the MGDA-specific attacks illustrated in Figure 1. We make the following assumptions.

Assumption 3.1 (**Task-wise honest majority**). *Honest weight $\alpha_s^{\mathcal{H}} > 1/2$ for every task $s \in [S]$.*



(a) Descent certification (b) Successful recovery (c) Recovery failure (d) Stationarity bound

Figure 2: CR-MGDA certifies a proposal, invokes recovery if needed, and bounds the honest Pareto-stationarity gap when recovery fails.

Assumption 3.2 (Smooth and lower-bounded objectives). *Every honest local objective $f_{s,i}$ is differentiable. For each $s \in [S]$, $f_s^{\mathcal{H}}$ has a finite lower bound $f_s^{\mathcal{H},\inf}$ and is L_s-smooth: $\|\nabla f_s^{\mathcal{H}}(x) - \nabla f_s^{\mathcal{H}}(y)\| \le L_s\|x-y\|$ for all $x, y \in \mathbb{R}^d$. Set $L := \max_{s\in[S]} L_s$.*

Assumption 3.3 (Bounded gradient heterogeneity). *For each task $s \in [S]$, there exists $\zeta_s \ge 0$ such that $\|\nabla f_{s,i}(x) - \nabla f_s^{\mathcal{H}}(x)\| \le \zeta_s$, $\forall x,\ i \in \mathcal{R}_s \cap \mathcal{H}$.*

Assumption 3.1 is the task-wise version of the standard honest-majority condition in Byzantine-robust aggregation (Blanchard et al., 2017; Xie et al., 2018; Yin et al., 2018). Given the potential conflict among objectives, we use the standard notions of Pareto optimality and stationarity:

Definition 3.4. *For $F^{\mathcal{H}}$, x dominates y if $f_s^{\mathcal{H}}(x) \le f_s^{\mathcal{H}}(y)$ for every $s \in [S]$, with a strict inequality for at least one task. A point x is Pareto optimal if no point dominates it, and Pareto-stationary if no common descent direction $d \in \mathbb{R}^d$ satisfies $\nabla f_s^{\mathcal{H}}(x)^\top d < 0$ for every $s \in [S]$.*

For equation 2, let $G_{\mathcal{H}}(x) := [\nabla f_1^{\mathcal{H}}(x), \dots, \nabla f_S^{\mathcal{H}}(x)]$, and write $G_{\mathcal{H}}^t := G_{\mathcal{H}}(x^t)$. The honest Pareto-stationarity gap is $\Psi(x) := \min_{\lambda\in\Delta_S} \|G_{\mathcal{H}}(x)\lambda\|$; thus $\Psi(x) = 0$ characterizes honest Pareto stationarity. The inlier-geometry attack in Figure 1 shows that gradient errors at the scale of $\Psi(x)$ can drive the observed MGDA gap to zero. Because $G_{\mathcal{H}}(x)$ is unobserved, the server must account for these errors when assessing both common descent and Pareto stationarity.

3.2 CR-MGDA: Certified Robust Pareto Decisions

Task-wise robust aggregation. For each task s, the central server aggregates the submitted reports $U_s^t := \{u_{s,i}^t : i \in \mathcal{R}_s\}$. We use the task-wise weighted geometric median (TWGM) as the default aggregator, computing an ϵ-approximate solution with Weiszfeld-type iterations (Aftab et al., 2015):

$$\hat{g}_s^t = \mathrm{TWGM}(U_s^t, \epsilon), \quad \sum_{i\in\mathcal{R}_s} \alpha_{s,i}\|\hat{g}_s^t - u_{s,i}^t\| \le \inf_{y\in\mathbb{R}^d} \sum_{i\in\mathcal{R}_s} \alpha_{s,i}\|y - u_{s,i}^t\| + \epsilon. \tag{3}$$

The weights are the task-wise data proportions used in equation 2, accounting for task participation and client data imbalance. The resulting matrix $\widehat{G}^t = [\hat{g}_1^t, \dots, \hat{g}_S^t]$ estimates the honest task gradients. The following lemma bounds each task's estimation error.

Lemma 3.5 (Task-wise TWGM error)**.** *Under Assumptions 3.1 and 3.3, if $\hat{g}_s^t$ is an ϵ-approximate TWGM computed from equation 3, then, for every task s and round t,*

$$\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_s := \frac{2\alpha_s^{\mathcal{H}}\zeta_s + \epsilon}{2\alpha_s^{\mathcal{H}} - 1}. \tag{4}$$

The radius increases with the heterogeneity ζ_s and approximation tolerance ϵ, and decreases as the honest weight $\alpha_s^{\mathcal{H}}$ moves away from $1/2$. It remains valid for arbitrary Byzantine report values and is therefore independent of the attack magnitude. Any known lower bound on $\alpha_s^{\mathcal{H}}$ above $1/2$ and upper bound on ζ_s therefore yield a valid, possibly conservative radius without identifying $\mathcal{H}$. Each radius bounds the distance from the estimate to the honest task gradient and defines the uncertainty ball used for certification and recovery. The certification and recovery procedures below require only valid task-wise error bounds; Appendix C.4 derives such bounds for several alternative aggregators.

Regularized MGDA proposal. Given the matrix $\widehat{G}^t$, CR-MGDA constructs a candidate direction using the MGDA subproblem in equation 1, with a quadratic penalty on concentrated task weights

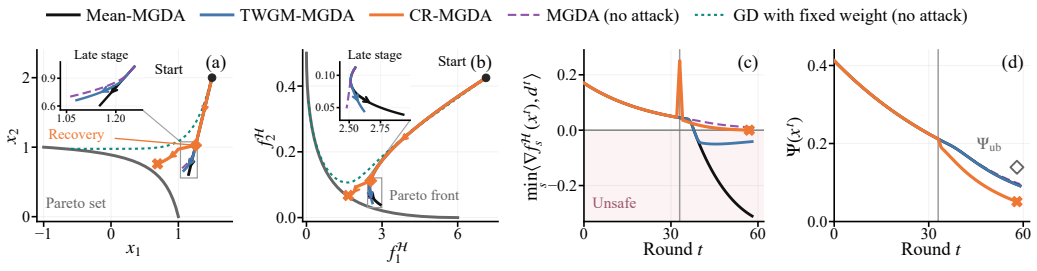


Figure 3: Full-gradient quadratic example with two objectives and five heterogeneous clients, including one Byzantine client that adds a fixed vector to its task 1 gradient report. TWGM-MGDA mitigates the attack relative to Mean-MGDA in the parameter and objective trajectories (a,b), but still takes updates with negative honest descent margins (c). CR-MGDA accepts only certified updates and, when recovery fails, stops and returns a stationarity-gap upper bound (d); in this instance, it reaches a smaller honest Pareto-stationarity gap than attack-free MGDA over the same horizon.

(Chen et al., 2024; Zhou et al., 2026):

$$\lambda_{\mathrm{p}}^t \in \arg\min_{\lambda \in \Delta_S} \left\{ \|\widehat{G}^t \lambda\|^2 + \frac{\gamma}{2}\|\lambda\|^2 \right\}, \quad d_{\mathrm{p}}^t = \widehat{G}^t \lambda_{\mathrm{p}}^t, \tag{5}$$

where $\gamma \geq 0$ controls the strength of the weight penalty. For $\gamma > 0$, the penalty makes the subproblem γ-strongly convex in λ, ensuring a unique solution. When $\gamma = 0$, equation 5 reduces to unregularized MGDA. The proposal is then tested for certified common descent.

Descent certificate. By Lemma 3.5, each honest task gradient lies in an uncertainty ball centered at $\hat{g}_s^t$. A direction computed from the centers of these balls may still fail to provide first-order descent for an honest objective, as illustrated in Figure 2a. CR-MGDA therefore requires common descent for every task gradient in the corresponding uncertainty ball. Proposition 3.6 gives an exact test.

Proposition 3.6 (Certified descent margin). *Let $v^t \in \mathbb{R}^d$ be a nonzero server-side candidate direction. Then $-v^t$ is a common-descent direction for every task-gradient matrix consistent with the error bounds in equation 4 if and only if $\langle \hat{g}_s^t, v^t \rangle > r_s \|v^t\|$ for every $s \in [S]$.*

Geometrically, every task-gradient uncertainty ball must lie entirely within the open half-space $\{g : \langle g, v^t \rangle > 0\}$. The test is exact for these balls: if it fails, at least one ball contains a gradient for which $-v^t$ is not a descent direction, as illustrated in Figure 2a. This does not necessarily imply that the proposal increases an honest objective, but common descent cannot be guaranteed from the current error bounds. Larger radii include more possible gradients and therefore make the certificate harder to satisfy. Peitz & Dellnitz (2018) express their inexact-descent condition as lower bounds on the combination weights, whereas Proposition 3.6 directly tests any candidate direction. For the proposal $d_{\mathrm{p}}^t = \widehat{G}^t \lambda_{\mathrm{p}}^t$ from equation 5, the task-wise certified margin satisfies

$$\langle \hat{g}_s^t, d_{\mathrm{p}}^t \rangle - r_s \|d_{\mathrm{p}}^t\| \geq \|d_{\mathrm{p}}^t\|^2 - \gamma/8 - r_s \|d_{\mathrm{p}}^t\|, \quad \forall s \in [S]. \tag{6}$$

Positivity of this lower bound for every task is sufficient to certify the proposal.

Recovery. Failure of the proposal certificate does not rule out a certified direction elsewhere in the same convex-combination family. CR-MGDA therefore searches this family for an alternative direction satisfying Proposition 3.6 before rejecting the update. For $\lambda \in \Delta_S$, define the certified margin $M_t(\lambda) := \min_{s \in [S]} \{ \langle \hat{g}_s^t, \widehat{G}^t \lambda \rangle - r_s \|\widehat{G}^t \lambda\| \}$. By Proposition 3.6, $M_t(\lambda) > 0$ certifies descent along $-\widehat{G}^t \lambda$. We control recovery progress through

$$\max_{\lambda \in \Delta_S, q, m} \; m - \kappa q^2 \quad \text{s.t. } \|\widehat{G}^t \lambda\| \leq q, \quad m + r_s q \leq \langle \hat{g}_s^t, \widehat{G}^t \lambda \rangle, \quad \forall s \in [S], \tag{7}$$

where $\kappa \in [0, 1/2]$. Setting $\kappa = 0$ gives the maximum-margin second-order cone program (SOCP) and the broadest acceptance region. With positive κ, the problem remains convex and conic-representable while enforcing sufficient progress. Let $(\lambda_{\mathrm{r}}^t, q_{\mathrm{r}}^t, m_{\mathrm{r}}^t)$ be an optimizer and set $d_{\mathrm{r}}^t = \widehat{G}^t \lambda_{\mathrm{r}}^t$. If $m_{\mathrm{r}}^t - \kappa (q_{\mathrm{r}}^t)^2 > 0$, then $m_{\mathrm{r}}^t > 0$ and $-d_{\mathrm{r}}^t$ is a certified common-descent direction. Figure 2b illustrates a recovered direction in the intersection of the task-wise certified regions. Whenever either the proposal or recovery is accepted, we denote the selected pair by (λ^t, d^t) and

update $x^{t+1} = x^t - \eta d^t$. If $m_{\mathrm{r}}^t - \kappa(q_{\mathrm{r}}^t)^2 \le 0$, no direction generated by a convex combination of the estimated task gradients has certified margin greater than κ times its squared norm. For $\kappa = 0$, this means that no direction in this family has a positive certified margin, as illustrated in Figure 2c.

Stationarity certificate. When recovery fails, CR-MGDA rejects the update and computes an upper bound on the honest Pareto-stationarity gap. Specifically, $m_{\mathrm{r}}^t - \kappa(q_{\mathrm{r}}^t)^2 \le 0$ implies

$$\Psi(x^t) \le \Psi_{\mathrm{ub}}^t := \min_{\lambda\in\Delta_S} \left[\|\widehat{G}^t\lambda\| + \sum\nolimits_{s\in[S]} \lambda_s r_s \right] \le \frac{2-\kappa}{1-\kappa} r, \quad r := \max_{s\in[S]} r_s. \tag{8}$$

For any fixed $\lambda \in \Delta_S$, the error bounds place the honest gradient combination in a ball centered at $\widehat{G}^t\lambda$ with radius $\sum_{s\in[S]} \lambda_s r_s$. The distance of the center from the origin plus this radius therefore upper-bounds the honest Pareto-stationarity gap. Minimizing over λ yields Ψ_{ub}^t, as illustrated in Figure 2d. Recovery failure further bounds the norm of the minimum-norm point in the estimated gradient convex hull by $r/(1-\kappa)$. Using its combination weights in Ψ_{ub}^t adds at most r, which gives the last inequality in equation 8; for $\kappa = 0$, the bound reduces to $2r$.

The server may terminate and return Ψ_{ub}^t, or pause with $x^{t+1} = x^t$ and reevaluate the certificate when new reports become available. Updates resume only when a proposal or recovery direction passes the descent certificate. Figure 3 illustrates this full-gradient procedure using the analytic radius from Lemma 3.5; Appendix B.1 specifies the construction and analyzes the resulting trajectory.

3.3 Probabilistic Certification with Stochastic Reports

Sampling introduces additional uncertainty into the aggregated task gradients, so the error balls used for certification must cover the honest task gradients with high probability.

Assumption 3.7 (**Stochastic gradients**). *Let $\mathcal{F}_t$ be the pre-report history. Conditional on $\mathcal{F}_t$, samples within each honest client-task report are independent. For every $s \in [S]$, $i \in \mathcal{R}_s \cap \mathcal{H}$, and sample ξ, $\mathbb{E}[\nabla f_{s,i}(x^t,\xi) \mid \mathcal{F}_t] = \nabla f_{s,i}(x^t)$ and $\mathbb{E}[\|\nabla f_{s,i}(x^t,\xi) - \nabla f_{s,i}(x^t)\|^2 \mid \mathcal{F}_t] \le \sigma^2$.*

Local reports. The finite-variance assumption permits heavy-tailed stochastic gradient noise. We use local GMOM to obtain high-probability control of the sampling error in this setting, followed by server-side TWGM to limit Byzantine contamination. Each honest client collects B samples in $\Xi_{s,i}^t$ and forms $K_{s,t}$ disjoint blocks $\Xi_{s,i,k}^t$ of size $b_{s,t} = \lfloor B/K_{s,t} \rfloor$, assuming $K_{s,t} \le B/2$. Define

$$\bar{u}_{s,i,k}^t = \frac{1}{b_{s,t}} \sum\nolimits_{\xi\in\Xi_{s,i,k}^t} \nabla f_{s,i}(x^t,\xi), \qquad \tilde{u}_{s,i}^t \in \arg\min_{u\in\mathbb{R}^d} \sum\nolimits_{k=1}^{K_{s,t}} \|u - \bar{u}_{s,i,k}^t\|.$$

The server applies TWGM across clients to obtain $\tilde{g}_s^t = \mathrm{TWGM}(\{\tilde{u}_{s,i}^t : i \in \mathcal{R}_s\}, \epsilon)$ for each task and forms $\widetilde{G}^t = [\tilde{g}_1^t, \dots, \tilde{g}_S^t]$. Although each block mean is unbiased, the GMOM report need not be. For a failure-probability budget $\delta_{s,t} \in (0,1)$, the following lemma provides a high-probability radius for certification and a conditional mean-square error bound for the convergence analysis.

Lemma 3.8 (Stochastic TWGM error). *Suppose Assumptions 3.1, 3.3, and 3.7 hold. Fix s, t, and let $K_{s,t} \le B/2$ satisfy $|\mathcal{R}_s|\exp(-K_{s,t}/3.5) \le \delta_{s,t}$. Then*

$$\Pr\Big(\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_{s,t}^{\mathrm{hp}} \mid \mathcal{F}_t \Big) \ge 1 - \delta_{s,t}, \quad r_{s,t}^{\mathrm{hp}} := \frac{2\alpha_s^{\mathcal{H}}(\zeta_s + \tau_{s,t}) + \epsilon}{2\alpha_s^{\mathcal{H}} - 1}, \quad \tau_{s,t} := \frac{11\sqrt{5}\,\sigma}{6\sqrt{b_{s,t}}},$$

$$\mathbb{E}[\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \mid \mathcal{F}_t] \le (R_{s,t}^{\mathrm{ms}})^2 := \frac{8(\alpha_s^{\mathcal{H}})^2(\zeta_s + 2\sigma/\sqrt{b_{s,t}})^2 + 2\epsilon^2}{(2\alpha_s^{\mathcal{H}} - 1)^2}.$$

The high-probability radius has the same form as the full-gradient radius in Lemma 3.5, with the additional sampling term $\tau_{s,t}$; in particular, $r_{s,t}^{\mathrm{hp}} \to r_s$ as $b_{s,t} \to \infty$. Following the block-count scaling of Minsker (2015, Corollary 4.1), one admissible choice is $K_{s,t} = 1 + \lfloor 3.5\log(|\mathcal{R}_s|/\delta_{s,t}) \rfloor \le B/2$. At fixed B, smaller failure budgets require more blocks and thus enlarge the sampling terms.

Remark 3.9. *For the same sampling budget B, the sample mean is unbiased and has conditional mean-square error at most σ^2/B. After TWGM aggregation, a second-moment bound and Markov's inequality yield a certification radius scaling as $1/\sqrt{\delta_{s,t}}$, with other parameters fixed. Under the same finite-variance assumption, the GMOM sampling term instead has an $\mathcal{O}(\sqrt{\log(1/\delta_{s,t})})$ dependence within the feasible block-count range. Figure 4 compares the radii obtained at different*

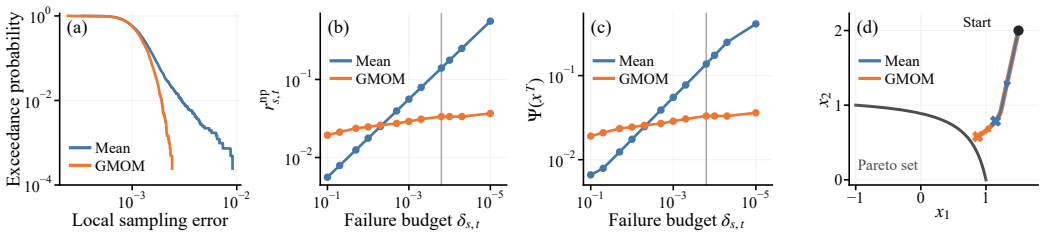


Figure 4: Stochastic quadratic example with homogeneous honest clients and symmetric heavy-tailed noise. Both CR-MGDA variants use the same sampling budget and TWGM aggregation. GMOM produces a lighter upper tail for the local estimation error (a), while its smaller sufficient radii at high confidence (b) allow further certified progress toward Pareto stationarity (c,d).

prescribed confidence levels and the resulting optimization trajectories under heavy-tailed noise at the same sampling budget. Appendix B.1 specifies the stochastic setting and analyzes the results.

Probabilistic certificates. We apply the certification procedure of Section 3.2 with $(\hat{g}_s^t, \widehat{G}^t, r_s)$ replaced by $(\tilde{g}_s^t, \widetilde{G}^t, r_{s,t}^{\mathrm{hp}})$, and denote the resulting upper bound in equation 8 by $\Psi_{\mathrm{ub,hp}}^t$. Conditional on $\mathcal{F}_t$, all task-wise error balls hold simultaneously with probability at least $1 - \sum_{s\in[S]} \delta_{s,t}$. On this event, every accepted direction d^t makes $-d^t$ a common-descent direction for all honest objectives, while failed recovery implies equation 8 with the same substitutions. The direction d^t may be computed from the same reports used for certification because the error bounds hold simultaneously for all tasks. If the failure budgets are chosen in advance so that $\sum_{t=0}^{T-1} \sum_{s\in[S]} \delta_{s,t} \le \delta_{\mathrm{tot}}$, a union bound gives both guarantees over all T rounds with probability at least $1 - \delta_{\mathrm{tot}}$.

4 Convergence Analysis

We bound the average squared honest Pareto-stationarity gap for non-convex objectives and the weighted objective gap under strong convexity. Proofs are provided in Appendix E.

4.1 Non-convex Objectives

Full-gradient reports. A positive certified margin guarantees common descent. For $\kappa > 0$, the quadratic penalty and acceptance rule in equation 7 further relate the recovery margin to the squared direction norm, yielding the progress bound below.

Theorem 4.1 (Non-convex convergence)**.** *Suppose Assumptions 3.1, 3.2, and 3.3 hold. At each round $t < T$, CR-MGDA either executes an accepted proposal or recovery update, or sets $x^{t+1} = x^t$ after failed recovery. For a fixed $\kappa \in (0, 1/2]$ and a constant step size $0 < \eta \le \kappa/L$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \Psi(x^t)^2 \le \frac{4\Delta_0}{\eta\kappa T} + \left(\frac{4}{\kappa} + 2\right) r^2 + \frac{\gamma}{2\kappa}, \quad \Delta_0 := \max_{s\in[S]} [f_s^{\mathcal{H}}(x^0) - f_s^{\mathcal{H},\inf}].$$

For fixed η and κ, the average squared honest Pareto-stationarity gap decreases at rate $\mathcal{O}(1/T)$ up to an $\mathcal{O}(r^2 + \gamma)$ residual due to gradient estimation and regularization. Rounds with failed recovery are included: by equation 8, they satisfy $\Psi(x^t)^2 \le [(2-\kappa)/(1-\kappa)]^2 r^2$, which is covered by the displayed residual term for $\kappa \in (0, 1/2]$. For $\kappa = 0$, accepted recovery directions remain certified common-descent directions, but the fixed-step rate above requires $\kappa > 0$; Appendix E.1 gives a simple example showing why a positive margin alone is insufficient. For proposal-only iterations, the sharper bound in Appendix E.1 holds under $0 < \eta \le 1/L$ and requires no recovery condition.

Stochastic reports. Under the failure-budget condition in Section 3.3, all task-gradient errors are bounded by their prescribed radii throughout T rounds with probability at least $1 - \delta_{\mathrm{tot}}$. On this event, the convergence bound in Theorem 4.1 holds with r replaced by $\bar{r}_{\mathrm{hp}} := \max_{t<T, s\in[S]} r_{s,t}^{\mathrm{hp}}$, under the same step-size condition.

We also bound the expected squared honest Pareto-stationarity gap directly using the conditional mean-square bounds in Lemma 3.8.

Theorem 4.2 (Expected non-convex convergence). *Suppose Assumptions 3.1, 3.2, 3.3, and 3.7 hold, and the deterministic local-GMOM block counts $K_{s,t}$ satisfy Lemma 3.8. Let CR-MGDA use the same update rule as in Theorem 4.1 and $R_{\mathrm{ms}}^2 := \max_{0\le t<T} \sum_{s\in[S]} (R_{s,t}^{\mathrm{ms}})^2$. For a fixed $\kappa \in (0, 1/2]$ and the step size $0 < \eta \le \kappa/L$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\Psi(x^t)^2] \le \frac{8\Delta_0}{\eta\kappa T} + \left(\frac{8}{\kappa^2} + 2 + \frac{6}{(1-\kappa)^2}\right) R_{\mathrm{ms}}^2 + \frac{\gamma}{\kappa}.$$

For fixed η and κ, the expected bound retains an $\mathcal{O}(1/T)$ term and an $\mathcal{O}(R_{\mathrm{ms}}^2 + \gamma)$ residual while allowing pauses after failed recovery. The task weights, selected direction, and decision to pause may all depend on the same reports used to estimate gradients; no independence between these choices and the estimation errors is required. If an update is executed at every round almost surely, the coefficient of R_{ms}^2 improves to $8/\kappa^2 + 2$.

4.2 Strongly Convex Objectives

Full-gradient reports. For strongly convex honest objectives sharing a common minimizer, we bound the weighted objective gap along runs that execute an update at every round.

Assumption 4.3 (Strong convexity). *For every task $s \in [S]$ and all $x, y \in \mathbb{R}^d$, the honest objective is μ-strongly convex: $f_s^{\mathcal{H}}(y) \ge f_s^{\mathcal{H}}(x) + \langle \nabla f_s^{\mathcal{H}}(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$. The honest objectives share a minimizer $x^\star \in \bigcap_{s\in[S]} \arg\min_{x\in\mathbb{R}^d} f_s^{\mathcal{H}}(x)$. There is a domain $\mathcal{D}$ containing all iterates and a constant $G_{\max} < \infty$ such that $\|\nabla f_{s,i}(x)\| \le G_{\max}$ for every $s \in [S]$, $i \in \mathcal{R}_s \cap \mathcal{H}$, and $x \in \mathcal{D}$.*

Theorem 4.4 (Strongly convex convergence). *Suppose Assumptions 3.1, 3.3, and 4.3 hold. Let $0 < \eta < 2/\mu$ and set $w_t := (1 - \eta\mu/2)^{-(t+1)}$. Suppose CR-MGDA executes an update at every round $t < T$. Let $\lambda^t \in \Delta_S$ denote the weights of the executed direction. Then*

$$\frac{\sum_{t=0}^{T-1} w_t \sum_{s\in[S]} \lambda_s^t \big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^\star)\big)}{\sum_{t=0}^{T-1} w_t} \le \frac{\mu\|x^0 - x^\star\|^2}{4((1-\eta\mu/2)^{-T} - 1)} + \frac{r^2}{\mu} + \eta(G_{\max}^2 + r^2).$$

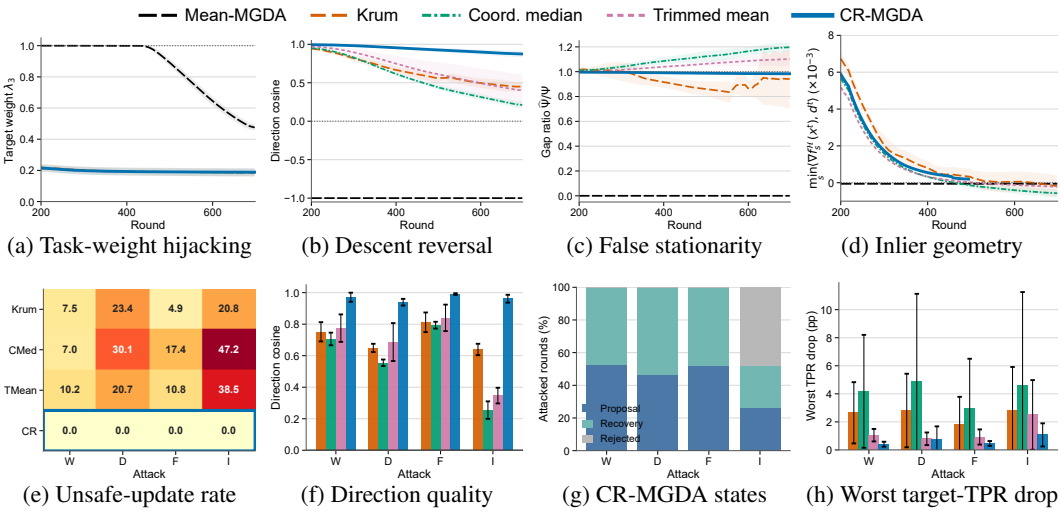
The bound controls a time-weighted average of the task-weighted objective gaps. The initial-distance term decays geometrically with T, while the residual depends on both the estimation error and the step size. Since all tasks share $x^\star$, each task-weighted gap is at least $(\mu/2)\|x^t - x^\star\|^2$, so the result also bounds the weighted average squared distance to $x^\star$. The proof uses only the simplex-combination structure of the executed direction, covering both proposal and recovery updates without requiring $\kappa > 0$ or an explicit γ term.

Stochastic reports. Under stochastic reports, the expected weighted gap satisfies the same bound with r^2 replaced by R_{ms}^2. This counterpart assumes the conditions of Theorem 4.4, Assumption 3.7, and deterministically chosen block counts satisfying Lemma 3.8; CR-MGDA must execute an update almost surely at every round.

5 Experiments

We evaluate CR-MGDA on UNSW-NB15 (Moustafa & Slay, 2015) using four class-conditional objectives: Normal, DoS, Exploits, and Reconnaissance. The data are partitioned across 10 clients with moderate heterogeneity, including 2 Byzantine clients, and all methods use the same multilayer perceptron. We report means and standard deviations over four seeds. The controlled quadratic examples use theoretically valid radii and evaluate formal certification, whereas the learning benchmarks calibrate practical radii from clean client reports and evaluate observed safety and optimization progress. The full setup and additional results on CelebA and CIFAR-100 appear in Appendix B.

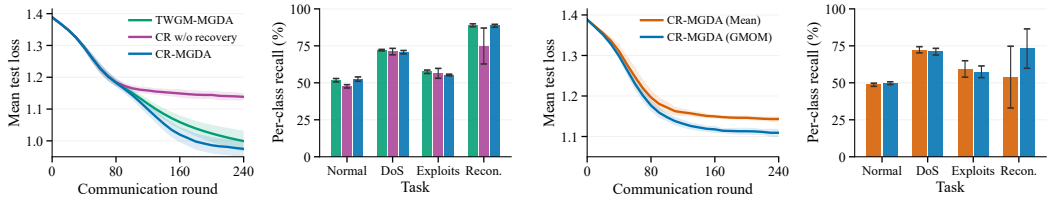
Using full-gradient local reports, we apply the four MGDA-specific attacks illustrated in Figure 1 and compare CR-MGDA with unprotected Mean-MGDA and three robust-aggregation baselines: Krum, coordinate median, and trimmed mean. Figures 5a–5d show that all robust methods substantially attenuate the attack-specific distortions observed with Mean-MGDA. The update-level summaries in Figures 5e and 5f, however, show that robustness alone does not ensure safe MGDA decisions. We call an executed update unsafe when its minimum honest first-order descent margin is



(a) Task-weight hijacking (b) Descent reversal (c) False stationarity (d) Inlier geometry

(e) Unsafe-update rate (f) Direction quality (g) CR-MGDA states (h) Worst target-TPR drop

Figure 5: Full-gradient results under four Byzantine attacks.



(a) Ablation: loss (b) Ablation: recall (c) Mean/GMOM: loss (d) Mean/GMOM: recall

Figure 6: Stochastic-gradient results under descent reversal.

negative. The three aggregation baselines still execute unsafe updates, whereas CR-MGDA executes none and remains closely aligned with the honest MGDA direction. Figure 5g shows that recovery supplies updates when the proposal cannot be certified, with some updates rejected under the inlier-geometry attack. Figure 5h shows small target-TPR drops for CR-MGDA. Appendix B.2.2 reports all five diagnostics for each attack.

For stochastic reports, we consider a descent-reversal attack targeting Reconnaissance, using TWGM aggregation and a sampling budget of $B = 256$. Under mini-batch sampling with GMOM local reports, Figures 6a and 6b compare TWGM-MGDA without certification, CR-MGDA without recovery, which skips failed proposals, and full CR-MGDA. The certification-only variant plateaus early, whereas recovery allows CR-MGDA to continue optimizing, reach a lower loss, and retain per-task recall comparable to TWGM-MGDA. Under symmetrized finite-variance Pareto noise, Figures 6c and 6d directly compare local sample means and GMOM within CR-MGDA. With the same sampling budget, GMOM yields a lower average test loss and higher Reconnaissance recall.

6 Conclusion

Robust aggregation alone does not ensure reliable MGDA decisions under Byzantine reports. CR-MGDA certifies common descent, searches for an alternative certified direction when the MGDA proposal fails, and returns an upper bound on the honest Pareto-stationarity gap when recovery fails. We establish convergence guarantees for non-convex and strongly convex objectives under full-gradient and stochastic reports, while experiments on three benchmarks show improved descent safety and sustained optimization progress under Byzantine attacks. Future work will seek tighter data-dependent radii and more efficient recovery rules to reduce conservatism.

REFERENCES

Khurrum Aftab, Richard Hartley, and Jochen Trumpf. Generalized Weiszfeld algorithms for L_q optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4):728–745, 2015. doi: 10.1109/TPAMI.2014.2353625.

Baris Askin, Pranay Sharma, Gauri Joshi, and Carlee Joe-Wong. Federated communication-efficient multi-objective optimization. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 4627–4635. PMLR, 2025.

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8632–8642, 2019.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 118–128. Curran Associates, Inc., 2017.

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(193):1–53, 2024.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (MGDA) for multiobjective optimization. *Comptes Rendus Mathematique*, 350(5–6):313–318, 2012. ISSN 1631-073X. doi: 10.1016/j.crma.2012.03.014.

El-Mahdi El-Mhamdi, Rachid Guerraoui, and Sébastien Rouault. The hidden vulnerability of distributed learning in Byzantium. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3521–3530. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/mhamdi18a.html.

Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Byzantine machine learning made easy by resilient averaging of momentums. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6246–6283. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/farhadkhani22a.html.

Heshan Devaka Fernando, Han Shen, Miao Liu, Subhajit Chaudhury, Keerthiram Murugesan, and Tianyi Chen. Mitigating gradient bias in multi-objective learning: A provably convergent approach. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dLAYGdKTi2.

Jörg Fliege and Benar Fux Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research*, 51(3):479–494, 2000. doi: 10.1007/s001860000043.

Zeou Hu, Kiarash Shaloudegi, Guojun Zhang, and Yaoliang Yu. Federated learning meets multi-objective optimization. *IEEE Transactions on Network Science and Engineering*, 9(4):2039–2051, 2022. doi: 10.1109/TNSE.2022.3169117.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021. doi: 10.1561/2200000083.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Byzantine-robust learning on heterogeneous datasets via bucketing. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=jXKKDEi5vJt.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Jiani Li, Waseem Abbas, and Xenofon Koutsoukos. Byzantine resilient distributed multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18215–18225. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/d37eb50d868361ea729bb4147eb3c1d8-Abstract.html.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=ByexElSYDr.

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*, volume 34, pp. 18878–18890, 2021.

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*, 339(3):1119–1148, 2024. doi: 10.1007/s10479-021-04033-z.

Yuchen Liu, Chen Chen, Lingjuan Lyu, Fangzhao Wu, Sai Wu, and Gang Chen. Byzantine-robust learning on heterogeneous data via gradient splitting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21404–21425. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23d.html.

Zhuqing Liu, Chaosheng Dong, Michinari Momma, Simone Shao, Shaoyuan Xu, Yan Gao, Haibo Yang, and Jia Liu. STIMULUS: Achieving fast convergence and low sample complexity in stochastic multi-objective learning. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 2711–2747. PMLR, 2025. URL https://proceedings.mlr.press/v286/liu25d.html.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4): 2308–2335, 2015. doi: 10.3150/14-BEJ645.

Nour Moustafa and Jill Slay. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 Military Communications and Information Systems Conference (MilCIS)*, pp. 1–6. IEEE, 2015. doi: 10.1109/MILCIS.2015.7348942.

Sebastian Peitz and Michael Dellnitz. Gradient-based multiobjective optimization with uncertainties. In *NEO 2016: Results of the Numerical and Evolutionary Optimization Workshop NEO 2016 and the NEO Cities 2016 Workshop Held on September 20–24, 2016 in Tlalnepantla, Mexico*, volume 731 of *Studies in Computational Intelligence*, pp. 159–182. Springer, Cham, 2018. doi: 10.1007/978-3-319-64063-1_7.

Krishna Pillutla, Sham M. Kakade, and Zaid Harchaoui. Robust aggregation for federated learning. *IEEE Transactions on Signal Processing*, 70:1142–1154, 2022. doi: 10.1109/TSP.2022.3153135.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 525–536. Curran Associates, Inc., 2018.

Yuhao Shen, Wei Xi, Yunyun Cai, Yuwei Fan, He Yang, and Jizhong Zhao. Multi-objective federated learning: Balancing global performance and individual fairness. *Future Generation Computer Systems*, 162:107468, 2025. ISSN 0167-739X. doi: 10.1016/j.future.2024.07.046.

Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B. Giannakis. Federated variance-reduced stochastic gradient descent with robustness to Byzantine attacks. *IEEE Transactions on Signal Processing*, 68:4583–4596, 2020. doi: 10.1109/TSP.2020.3012952.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Generalized Byzantine-tolerant SGD. *arXiv preprint arXiv:1802.10116*, 2018. doi: 10.48550/arXiv.1802.10116.

Haibo Yang, Zhuqing Liu, Jia Liu, Chaosheng Dong, and Michinari Momma. Federated multi-objective learning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 39602–39625. Curran Associates, Inc., 2023.

Yi-Rui Yang, Chang-Wei Shi, and Wu-Jun Li. On the tension between Byzantine robustness and no-attack accuracy in distributed learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71051–71072. PMLR, 2025. URL https://proceedings.mlr.press/v267/yang25aa.html.

Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5650–5659. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/yin18a.html.

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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 38103–38115. Curran Associates, Inc., 2022.

Yuan Zhou, Yidan Ou, and Xinli Shi. DREAM: A unified framework for drift-corrected federated multi-objective learning. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=G8EfZTGGto.

Shiyuan Zuo, Xingrun Yan, Rongfei Fan, Han Hu, Hangguan Shan, Tony Q. S. Quek, and Puning Zhao. Federated learning resilient to Byzantine attacks and data heterogeneity. *IEEE Transactions on Mobile Computing*, 24(10):10729–10742, 2025. doi: 10.1109/TMC.2025.3571058.

LLM USAGE

We used ChatGPT for language editing, literature search, technical discussion, and assistance with the derivation of some supporting lemmas, and OpenAI Codex for LaTeX revisions and assistance with experiment implementation and execution. The authors checked the suggested references against the original publications and independently verified the mathematical arguments, experimental code, and reported results.

REPRODUCIBILITY STATEMENT

Algorithm 1 gives the full CR-MGDA procedure. Appendix B describes the datasets, models, training protocols, attacks, radius calibration, and additional results. Appendices C–E provide the complete proofs.

A ALGORITHM OVERVIEW

Algorithm 1 summarizes CR-MGDA using full-gradient notation.

Algorithm 1 CR-MGDA

Require: Initial model x^0, reporting horizon T, step size η, regularization parameter γ, recovery parameter $\kappa \in [0, 1/2]$, TWGM tolerance ϵ, task-wise client weights $\{\alpha_{s,i}\}$, error radii $\{r_s\}$, and a policy to terminate or pause after recovery failure.

1: **for** $t = 0, \dots, T-1$ **do**
2: Broadcast x^t and collect task-wise reports $\{u^t_{s,i}\}_{i \in \mathcal{R}_s}$.
3: For each task s, compute an ϵ-approximate minimizer

$$\hat{g}^t_s \approx \arg\min_{y \in \mathbb{R}^d} \sum\nolimits_{i \in \mathcal{R}_s} \alpha_{s,i} \|y - u^t_{s,i}\|.$$

4: Form $\widehat{G}^t = [\hat{g}^t_1, \dots, \hat{g}^t_S]$ and compute

$$\lambda^t_{\mathrm{p}} \in \arg\min_{\lambda \in \Delta_S} \left\{ \|\widehat{G}^t \lambda\|^2 + \frac{\gamma}{2} \|\lambda\|^2 \right\}, \qquad d^t_{\mathrm{p}} = \widehat{G}^t \lambda^t_{\mathrm{p}}.$$

5: **if** $\langle \hat{g}^t_s, d^t_{\mathrm{p}} \rangle > r_s \|d^t_{\mathrm{p}}\|$ for every $s \in [S]$ **then**
6: Set $(\lambda^t, d^t) = (\lambda^t_{\mathrm{p}}, d^t_{\mathrm{p}})$.
7: **else**
8: Compute an optimizer $(\lambda^t_{\mathrm{r}}, q^t_{\mathrm{r}}, m^t_{\mathrm{r}})$ of

$$\max_{\lambda \in \Delta_S,\, q,\, m} \quad m - \kappa q^2 \qquad \text{s.t.} \quad \|\widehat{G}^t \lambda\| \le q, \quad m + r_s q \le \langle \hat{g}^t_s, \widehat{G}^t \lambda \rangle, \quad \forall s \in [S].$$

9: **if** $m^t_{\mathrm{r}} - \kappa (q^t_{\mathrm{r}})^2 > 0$ **then**
10: Set $(\lambda^t, d^t) = (\lambda^t_{\mathrm{r}}, \widehat{G}^t \lambda^t_{\mathrm{r}})$.
11: **else**
12: Compute and report the stationarity-gap bound

$$\Psi^t_{\mathrm{ub}} := \min_{\lambda \in \Delta_S} \left[\|\widehat{G}^t \lambda\| + \sum\nolimits_{s \in [S]} \lambda_s r_s \right].$$

13: **if** the policy selects termination **then**
14: **return** x^t and Ψ^t_{ub}.
15: **end if**
16: Set $x^{t+1} = x^t$ and **continue** to the next reporting round.
17: **end if**
18: **end if**
19: Update $x^{t+1} = x^t - \eta d^t$.
20: **end for**
21: **return** x^T.

For stochastic reporting, honest clients form local GMOM reports as in Section 3.3. The server uses $(\tilde{g}_s^t, \widetilde{G}^t, r_{s,t}^{\mathrm{hp}})$ in place of $(\hat{g}_s^t, \widehat{G}^t, r_s)$, and reports $\Psi_{\mathrm{ub,hp}}^t$ upon recovery failure.

B Experimental Details and Additional Results

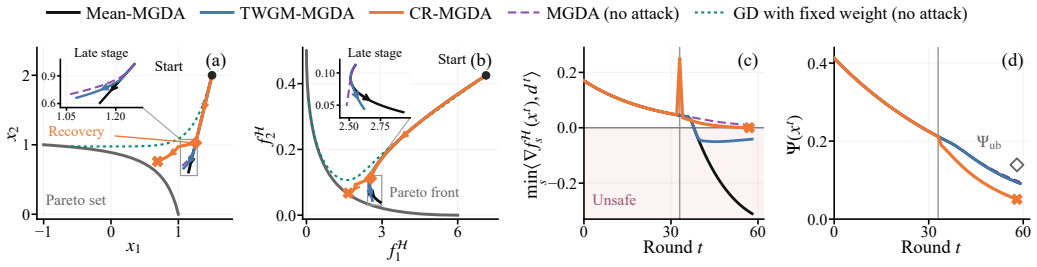
B.1 Quadratic Illustrations

Problem setting. Figures 3 and 4 use the quadratic objectives

$$f_1^{\mathcal{H}}(x) = \frac{1}{2}\left[(x_1+1)^2 + 8(x_2-1)^2\right], \qquad f_2^{\mathcal{H}}(x) = 0.1\left[(x_1-1)^2 + x_2^2\right], \quad x = (x_1, x_2).$$

All runs start from $x^0 = (1.5, 2)$ and use step size $\eta = 0.1$, MGDA regularization $\gamma = 2 \times 10^{-3}$, and maximum-margin recovery ($\kappa = 0$). The full-gradient setting uses the analytic radius from Lemma 3.5 to isolate residual aggregation error under a Byzantine report. The stochastic setting keeps the server-side aggregator fixed and uses the high-probability radius from Lemma 3.8 to isolate local sampling error.

Full-gradient setting. We use five equally weighted clients, four honest and one Byzantine, and let $u = (1,1)/\sqrt{2}$. For task 1, two honest reports are $\nabla f_1^{\mathcal{H}}(x) - 0.5u$ and two are $\nabla f_1^{\mathcal{H}}(x) + 0.5u$; all honest task-2 reports equal $\nabla f_2^{\mathcal{H}}(x)$. Thus, the honest mean equals the task gradient and the heterogeneity bounds are $(\zeta_1, \zeta_2) = (0.5, 0)$. The Byzantine client reports $\nabla f_1^{\mathcal{H}}(x) + 24u$ for task 1 and the exact task-2 gradient. Since all task 1 reports lie on the same affine line, the TWGM estimate differs from the honest task gradient by exactly $0.5u$, compared with $4.8u$ for the client mean. Substituting $\alpha_1^{\mathcal{H}} = \alpha_2^{\mathcal{H}} = 4/5$, $(\zeta_1, \zeta_2) = (0.5, 0)$, and $\epsilon = 0$ into equation 4 gives $(r_1, r_2) = (4/3, 0)$. We fix these radii throughout optimization and compare Mean-MGDA, TWGM-MGDA, and CR-MGDA under the same client and attack configuration.



Reproduction of Figure 3. Mean-MGDA, TWGM-MGDA, and CR-MGDA use the same client and attack configuration; CR-MGDA applies certification and recovery to the TWGM estimates. Crosses mark the CR-MGDA stopping point.

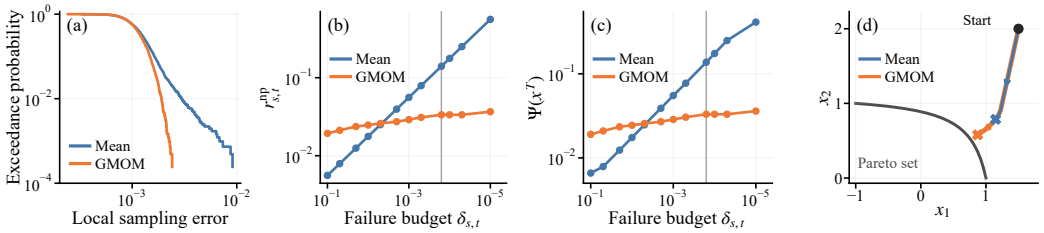
Mean-MGDA deviates substantially under the Byzantine report in Figure 3, whereas TWGM-MGDA removes most of this error. The residual error still makes the TWGM-MGDA direction unsafe from round 38: the task 1 objective begins to increase in panel (b) as its honest descent margin becomes negative in panel (c). CR-MGDA accepts the proposal for 33 rounds. Once the proposal first fails certification, recovery is invoked immediately and finds an alternative direction with a positive honest descent margin. The first recovery update appears as a larger jump in the parameter and objective trajectories in panels (a,b), but decreases both objectives and moves the iterate toward the Pareto set. CR-MGDA then takes 25 certified recovery steps before the optimized margin becomes nonpositive. At the resulting stopping point, panel (d) reports a true honest Pareto-stationarity gap of 0.0509 and an upper bound $\Psi_{\mathrm{ub}} = 0.1391$.

Stochastic setting. We retain the same objectives and initialization but use five honest, homogeneous clients, so $\alpha_s^{\mathcal{H}} = 1$, $\zeta_s = 0$, and $\epsilon = 0$ for both tasks, and sampling is the only source of error. Each sample gradient is perturbed along u by symmetric Student-t noise with $\nu = 2.5$ degrees of freedom, standardized to unit variance and scaled by $\sigma = 0.01$. Each client uses $B = 256$ samples

per task to form either a sample mean or a local GMOM report, followed by server-side TWGM. For task–round failure budget $\delta_{s,t}$, the sample-mean radius is

$$r_{s,t}^{\mathrm{mean}} = \sqrt{\frac{8\sigma^2}{B\delta_{s,t}}},$$

while the GMOM radius follows Lemma 3.8 with $K_{s,t} = 1 + \lfloor 3.5 \log(5/\delta_{s,t}) \rfloor$ blocks of size $b_{s,t} = \lfloor B/K_{s,t} \rfloor$. For panels (a,d), we use $T = 157$ and $\delta_{s,t} = 0.05/(2T)$, which allocates failure probability 0.05 across both tasks and all rounds. The corresponding radii are $r_{s,t}^{\mathrm{mean}} = 0.1401$ and $r_{s,t}^{\mathrm{GMOM}} = 0.0335$. Results are averaged over 40 paired runs, with the two local estimators using the same samples within each pair; after the first failed recovery, the iterate is held fixed for plotting.



Reproduction of Figure 4. Mean and GMOM use the same sampling budget and server-side TWGM. Panel (a) shows the empirical tail of the maximum local sampling error over clients and tasks from 4096 independent draws, and panel (b) shows the corresponding sufficient radii. Panel (c) reports the mean final stationarity gap and 10th–90th percentile range over 40 paired runs, while panel (d) shows their trajectories at the selected failure budget. The gray vertical lines in panels (b,c) mark the budget used in panels (a,d).

Panel (a) of Figure 4 shows the empirical tail of the maximum local sampling error over clients and tasks; GMOM has the shorter upper tail. This distinction matters at small failure budgets: the sample-mean radius scales as $1/\sqrt{\delta_{s,t}}$, while the GMOM sampling term scales as $\sqrt{\log(1/\delta_{s,t})}$ within the feasible block-count range. At the marked budget, the sample-mean runs execute about 48 updates on average and have a mean final stationarity gap of 0.1374, whereas the GMOM runs execute about 96 updates and reach 0.0331. The ordering reverses at moderate failure budgets, where the sample mean yields both a smaller sufficient radius and a smaller final gap, as shown in panels (b,c).

B.2 UNSW-NB15 Experiments

Data and training. UNSW-NB15 is a network-intrusion benchmark containing normal traffic and several attack categories. We use Normal, DoS, Exploits, and Reconnaissance traffic, with one class-conditional loss per category, to obtain four objectives. The training examples for each objective are partitioned independently across ten clients using a Dirichlet distribution with concentration $\alpha = 3$. Two clients are Byzantine, with $\sum_{i \in \mathcal{B}} \alpha_{s,i} \leq 0.2$ for every objective s. All comparisons use four matched runs; within each run, they share the same client partition.

The classifier is a two-layer ReLU MLP with hidden widths 128 and 64. All methods use step size $\eta = 0.1$ and MGDA regularization $\gamma = 10^{-3}$. The full-gradient experiments begin from a common clean MGDA checkpoint obtained after $T_{\mathrm{warm}} = 200$ warm-up rounds and continue for $T = 500$ rounds under scheduled attacks or a matched clean condition. The stochastic experiments instead run for $T = 240$ rounds from a common random initialization. Each client uses $B = 256$ samples to form either a sample-mean or GMOM report, with samples matched across variants; the heavy-tailed setting applies independent random signs to centered finite-variance Pareto weights and adds the resulting noise to the sampled gradients.

Methods and radii. We compare Mean-MGDA, Krum, coordinate median, trimmed mean, and CR-MGDA, applying each aggregation rule task-wise. CR-MGDA uses TWGM estimates with certification and recovery, with $\kappa = 0.1$ throughout. If neither the proposal nor the recovered

Table 1: Target matrices and resulting MGDA solutions for the four attacks illustrated in Figure 1.

Byzantine Attack	Target Matrix $A \in \mathbb{R}^{d\times S}$	Weight $\lambda_A \in \arg\min_{\lambda\in\Delta_S} \|A\lambda\|^2$	Direction $d_A = A\lambda_A$
Task-weight hijacking	$A = [v, \chi v, \ldots, \chi v]$	$\lambda_A = [1, 0, \ldots, 0]$	$d_A = v$
Descent-direction reversal	$A = [u := -\beta d_{\mathcal{H}}, \chi u, \ldots, \chi u]$	$\lambda_A = [1, 0, \ldots, 0]$	$d_A = -\beta d_{\mathcal{H}}$
False Pareto stationarity	$A = [w, -w, a_3, \ldots, a_S]$	$\lambda_A = [0.5, 0.5, \ldots, 0]$	$d_A = \mathbf{0}$
Inlier geometry attack	$A = G_{\mathcal{H}} - d_{\mathcal{H}}\mathbf{1}^\top$	$\lambda_A = \lambda_{\mathcal{H}}$	$d_A = \mathbf{0}$

Note. Here $\chi > 1$ and $\beta > 0$, with $v \neq \mathbf{0}$ and $d_{\mathcal{H}} \neq \mathbf{0}$ where applicable. The displayed λ_A denotes one minimizer; minimizers need not be unique. In the last row, $A\lambda_{\mathcal{H}} = \mathbf{0}$.

direction is certified, the update is skipped and new reports are collected in the next round. Under standard mini-batch sampling without injected heavy-tailed noise, the ablation compares TWGM-MGDA, CR-MGDA without recovery, and full CR-MGDA. The heavy-tailed experiment fixes the server-side method and varies only the local estimator.

Unlike the controlled illustrations, the UNSW-NB15 experiments use radii calibrated from clean client reports. Each bundle provides a task-wise calibration value based on the discrepancy between the TWGM estimate and the clean reports, together with their across-client spread. We use $N_{\mathrm{cal}} = 128$ clean bundles at two reference model points. The full-gradient runs use fixed task-wise radii obtained from these bundles. For stochastic reports, the calibrated scale is combined with the current weighted median residual ρ_s^t, so $r_s^t = c_s\rho_s^t$ follows the report dispersion observed at the server. Section B.2.2 compares the analytic and empirical radii, including their coverage and resulting update decisions.

Attacks and metrics. In the full-gradient experiment, the four MGDA-specific attacks illustrated in Figure 1 are applied every third round. Table 1 gives their target matrices and the resulting MGDA solutions. Task-weight hijacking and descent reversal target Reconnaissance, false stationarity targets Normal, and inlier geometry is task-agnostic. The stochastic experiment applies descent reversal to Reconnaissance every third round from initialization. Attack construction and offline diagnostics may use honest gradients, whereas CR-MGDA receives only the aggregated reports.

An executed update is unsafe if its minimum honest descent margin, $\min_s\langle\nabla f_s^{\mathcal{H}}(x^t), d^t\rangle$, is negative; the unsafe-update rate is computed over executed attacked rounds. We also measure cosine similarity to the honest minimum-norm direction and the stationarity-gap ratio $\widehat{\Psi}^t/\Psi(x^t)$, where $\widehat{\Psi}^t := \min_{\lambda\in\Delta_S}\|\widehat{G}^t\lambda\|$ is the observed gap. The target-class true-positive rate (TPR) is the Reconnaissance TPR at a 1% false-positive rate (FPR), and its worst drop is the largest decrease from the matched clean run. Predictive metrics comprise the mean class-conditional test loss, accuracy, macro F1, and worst-class recall.

B.2.1 Detailed Full-Gradient Results

Figure 7 reports all five diagnostics for the full-gradient runs summarized in Figure 5, rather than the single attack-specific diagnostic selected for each column in the main text.

Each attack produces its intended distortion under Mean-MGDA. The three aggregation baselines reduce these effects, but some method–attack pairs still execute updates with negative honest descent margins. As summarized in Figure 5e, these failures depend on both the attack and the aggregation rule. Every attacked update executed by CR-MGDA has a positive honest descent margin, and its direction remains more closely aligned with the honest MGDA direction than those of the aggregation baselines.

Only inlier geometry causes CR-MGDA to reject updates on attacked rounds, reducing its overall update rate to 83.9%. The direction-related curves omit these rejected rounds and therefore end early, although evaluation and certified updates on nonattacked rounds continue. Under CR-MGDA, Reconnaissance TPR shows no systematic degradation across the four attacks. Because several baseline bands are wide, we treat this metric only as a check that safety filtering does not severely degrade target-class performance.

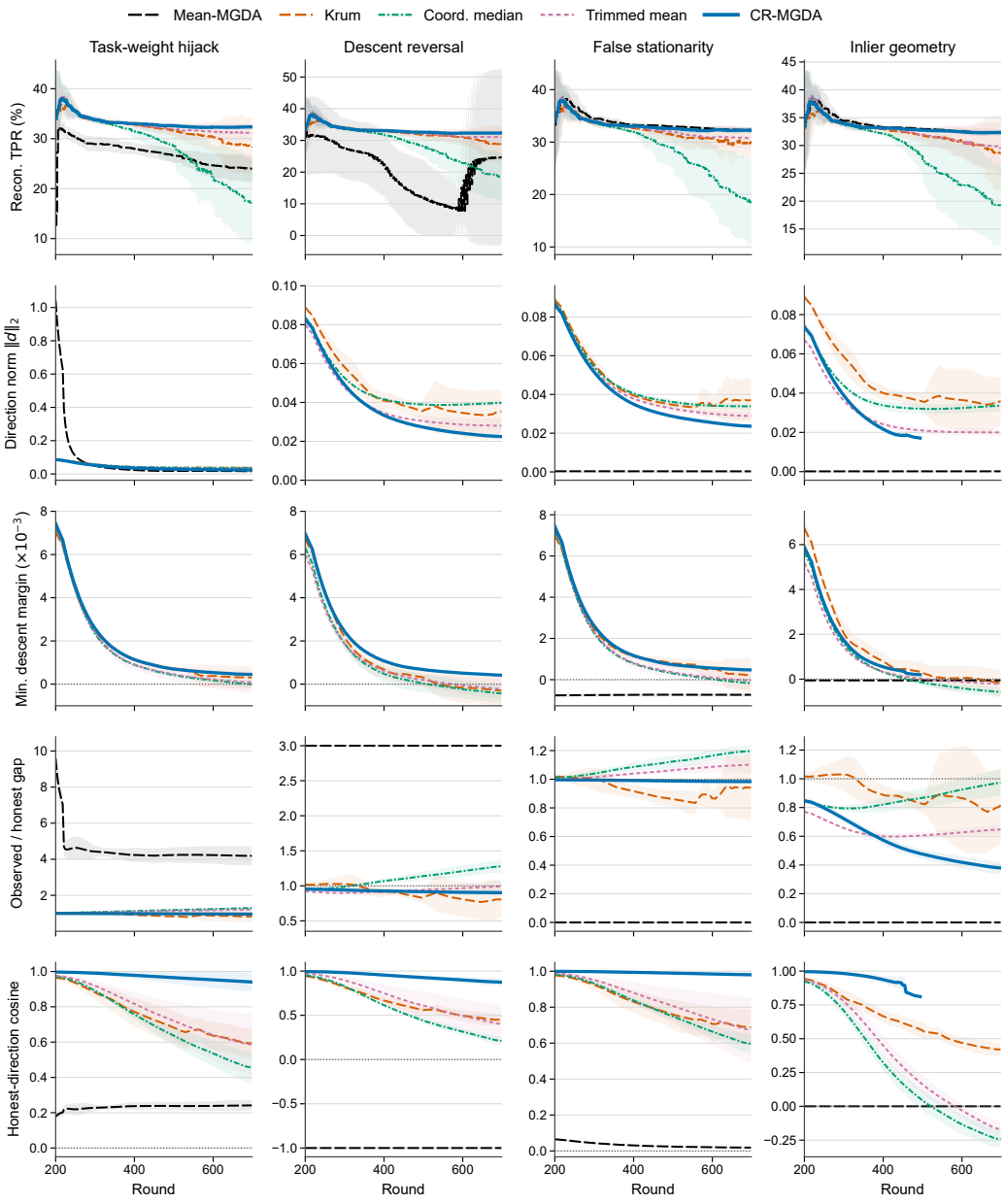


Figure 7: Full-gradient diagnostics for the same runs as Figure 5. Columns denote the four attacks; rows report Reconnaissance TPR, direction norm, minimum honest descent margin, stationarity-gap ratio, and honest-direction cosine. Curves show means over four matched runs, with bands indicating one standard deviation.

B.2.2 Radius Choice: Coverage and Update Availability

Table 2 compares the empirical and analytic radii on UNSW-NB15. For each task and round, we record whether the radius covers the realized aggregation error $e_s^t = \|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|$, and whether CR-MGDA accepts the proposal, uses recovery, or rejects the round. The honest aggregate gradients are used only for this evaluation.

For the stochastic experiments, the Empirical rows correspond to the runs in Figure 6. The Analytic rows recompute the decisions at the same model states and from the same reports using the heterogeneity term in the analytic bound. They omit the stochastic sampling term, whose population

quantities are not known exactly in these experiments. These rows therefore do not provide formal stochastic certificates; the complete high-probability radii would be larger. For the full-gradient comparison, no sampling term is needed. We run 240 rounds using the same residual-scaled calibration, with c_s calibrated at two clean reference models and then fixed throughout training. At the same states and from the same reports, the Analytic row recomputes the decisions using the valid radius from Lemma 3.5.

Table 2: Effect of the radius choice on error coverage and CR-MGDA decisions for the stochastic runs in Figure 6 and a matched full-gradient comparison. Mean radius, error, and task coverage are computed over task–round pairs; all-task coverage and decision fractions are computed over rounds.

Reporting condition	Radius	Mean radius	Mean error	Coverage (task/all)	Proposal	Recovery	Rejected
Mini-batch GMOM	Empirical	0.109	0.080	83.2% / 44.3%	37.0%	39.1%	24.0%
	Analytic	0.338	0.080	100.0% / 100.0%	8.1%	19.4%	72.5%
Heavy-tailed mean	Empirical	0.234	0.138	94.7% / 80.7%	11.1%	26.5%	62.4%
	Analytic	0.258	0.138	93.9% / 77.6%	9.1%	21.8%	69.2%
Heavy-tailed GMOM	Empirical	0.207	0.133	91.3% / 69.3%	17.4%	27.1%	55.5%
	Analytic	0.286	0.133	97.3% / 89.6%	8.2%	19.0%	72.8%
Full gradient	Empirical	0.021	0.015	88.2% / 61.9%	100.0%	0.0%	0.0%
	Analytic	0.347	0.015	100.0% / 100.0%	8.4%	19.5%	72.1%

In all three stochastic settings, the empirical radii reject fewer rounds than the analytic comparison. The largest difference occurs for mini-batch GMOM: the rejection rate decreases from 72.5% to 24.0%, while task-level coverage decreases from 100.0% to 83.2%. This lower rejection rate is consistent with the continued optimization in Figure 6. Including the omitted sampling term would further increase the analytic radius.

The full-gradient comparison shows the same effect more clearly. The valid analytic radius has mean 0.347, more than twenty times the mean realized error of 0.015. It covers every task and round but rejects 72.1% of the rounds. The empirical radius has mean 0.021, covers 88.2% of the task–round pairs, and accepts every proposal. Across all empirical-radius runs, 1,027 of 2,478 accepted rounds contain at least one uncovered task, but none of these rounds has a negative honest descent margin. This observation does not certify those rounds; it shows only that incomplete coverage need not imply an unsafe update.

B.3 Additional Learning Benchmarks

Common protocol. For both benchmarks, we distribute the training data IID across five clients, one of which is Byzantine. Each client forms task-wise GMOM reports using $K = 8$ blocks, and the server applies TWGM before solving the regularized MGDA subproblem. We normalize task gradients to unit norm, set $\gamma = 10^{-3}$, and use maximum-margin recovery ($\kappa = 0$). We use a clean warm-up period to calibrate fixed task-wise radii. We then compare clean training with sign flipping ($g \mapsto -g$), relative Gaussian noise ($g \mapsto g + \xi$, where $\xi \sim \mathcal{N}(0, 0.25\|g\|^2 I/d)$), and zero-gradient suppression ($g \mapsto 0$). Attacks occur every third post-warm-up round; all results use four matched runs, with evaluation every ten rounds.

B.3.1 CelebA

We use eight binary facial attributes from CelebA (Liu et al., 2015) as objectives: Heavy Makeup, High Cheekbones, Male, Brown Hair, Mouth Slightly Open, Wavy Hair, Wearing Necklace, and Wearing Lipstick. Images are resized to 64×64 and normalized to $[-1, 1]$ per channel, and evaluation uses a fixed subset of 5,000 images from the official validation split. The shared model has three 3×3 convolutional layers of widths 16, 32, and 64 with ReLU activations, followed by a 128-dimensional representation and a linear attribute head. We use blocks of size $b = 16$, giving $B = 128$ examples per local report, with $\eta = 0.5$, $T = 1{,}000$, and a 200-round warm-up.

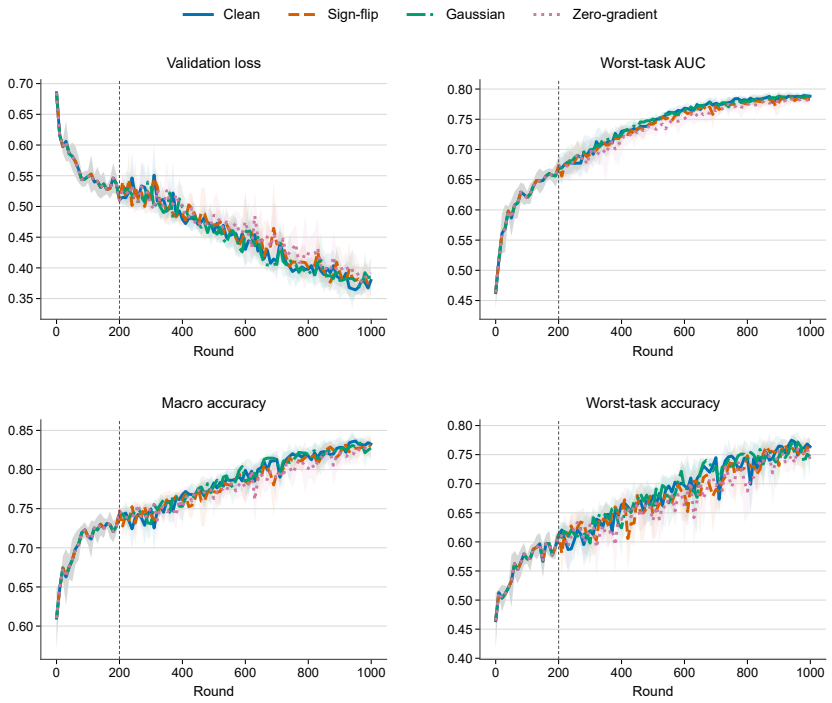


Figure 8: CelebA convergence under standard report-level attacks over four matched runs. Curves and bands show the mean and one standard deviation, respectively; the dashed line ends the 200-round warm-up, after which attacks occur every third round. Metrics are evaluated every ten rounds.

Table 3: Late-stage CelebA performance and CR-MGDA update rates over four matched runs. Values are means and sample standard deviations; performance is first averaged over the final ten evaluations within each run. Update rates cover all post-warm-up rounds or attacked rounds only.

Condition	Val. loss ↓	Macro AUC ↑	Worst AUC ↑	Update rate	Attacked-round update rate
Clean	0.374 ± 0.009	0.878 ± 0.006	0.787 ± 0.004	$80.0 \pm 10.6\%$	N/A
Sign-flip	0.381 ± 0.013	0.874 ± 0.007	0.784 ± 0.006	$73.8 \pm 12.1\%$	$64.3 \pm 13.2\%$
Gaussian	0.381 ± 0.021	0.877 ± 0.010	0.787 ± 0.007	$80.3 \pm 15.5\%$	$80.5 \pm 16.4\%$
Zero-gradient	0.392 ± 0.015	0.869 ± 0.011	0.782 ± 0.007	$65.1 \pm 12.9\%$	$40.6 \pm 9.8\%$

Zero-gradient suppression most strongly limits updating on CelebA: CR-MGDA accepts 40.6% of attacked rounds, compared with 64.3% under sign flipping and 80.5% under Gaussian noise. Despite this lower acceptance rate, the validation loss continues to decrease, and the mean paired reduction in worst-task AUC is 0.0055 ± 0.0046. Performance under the other two attacks remains close to the matched clean runs.

B.3.2 CIFAR-100

CIFAR-100 (Krizhevsky, 2009) groups its 100 fine classes into 20 superclasses, each of which defines a five-way classification objective. We use the official training and test splits, normalize each channel using training-set statistics, and apply no data augmentation. The shared model has three 3×3 convolutional layers of widths 16, 32, and 32 with ReLU activations, followed by a 128-dimensional representation and a five-way linear head for each superclass. Here $b = 4$ and $B = 32$; we set $\eta = 0.35$, $T = 400$, and use the first 100 rounds for radius calibration.

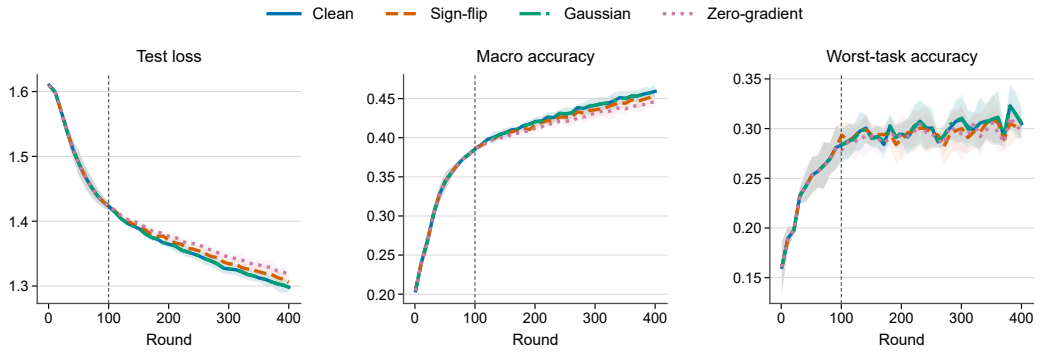


Figure 9: CIFAR-100 convergence under standard report-level attacks over four matched runs. Curves and bands show the mean and one standard deviation, respectively; the dashed line ends the 100-round warm-up, after which attacks occur every third round.

Table 4: Late-stage CIFAR-100 performance and CR-MGDA update rates over four matched runs. Values are means and sample standard deviations; performance is first averaged over the final ten evaluations within each run. Update rates cover all post-warm-up rounds or attacked rounds only.

Condition	Test loss ↓	Macro acc. ↑	Worst acc. ↑	Update rate	Attacked-round update rate
Clean	1.311 ± 0.009	0.452 ± 0.008	0.307 ± 0.017	$100.0 \pm 0.0\%$	N/A
Sign-flip	1.320 ± 0.009	0.446 ± 0.008	0.301 ± 0.013	$100.0 \pm 0.0\%$	$100.0 \pm 0.0\%$
Gaussian	1.312 ± 0.009	0.451 ± 0.008	0.307 ± 0.017	$100.0 \pm 0.0\%$	$100.0 \pm 0.0\%$
Zero-gradient	1.331 ± 0.008	0.439 ± 0.008	0.299 ± 0.011	$76.0 \pm 2.1\%$	$28.0 \pm 6.3\%$

Sign flipping and Gaussian noise trigger no rejected updates in these runs, and their test metrics remain close to the clean condition. Under zero-gradient suppression, the update rate falls to 28.0% on attacked rounds but remains 76.0% overall, with a mean paired reduction in macro accuracy of 0.0123 ± 0.0015. Most accepted updates in this condition therefore occur between attacked rounds.

C Proofs and Derivations for Section 3.2

C.1 Proof of Lemma 3.5

Lemma 3.5 (Task-wise TWGM error). *Under Assumptions 3.1 and 3.3, if $\hat{g}_s^t$ is an ϵ-approximate TWGM computed from equation 3, then, for every task s and round t,*

$$\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_s := \frac{2\alpha_s^{\mathcal{H}}\zeta_s + \epsilon}{2\alpha_s^{\mathcal{H}} - 1}. \tag{4}$$

Proof. Fix a task s and round t. Comparing the TWGM objective at $\hat{g}_s^t$ with its value at the honest task gradient, ϵ-optimality gives

$$\sum\nolimits_{i\in\mathcal{R}_s} \alpha_{s,i}\|\hat{g}_s^t - u_{s,i}^t\| \le \sum\nolimits_{i\in\mathcal{R}_s} \alpha_{s,i}\|\nabla f_s^{\mathcal{H}}(x^t) - u_{s,i}^t\| + \epsilon.$$

For every honest client $i \in \mathcal{R}_s \cap \mathcal{H}$, $u_{s,i}^t = \nabla f_{s,i}(x^t)$; hence Assumption 3.3 implies

$$\|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| = \|\nabla f_{s,i}(x^t) - \nabla f_s^{\mathcal{H}}(x^t)\| \le \zeta_s.$$

Consequently, the triangle inequality gives

$$\begin{aligned}
&\|\hat{g}_s^t - u_{s,i}^t\| - \|\nabla f_s^{\mathcal{H}}(x^t) - u_{s,i}^t\| \\
&\quad \ge \|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| - 2\|\nabla f_s^{\mathcal{H}}(x^t) - u_{s,i}^t\| \\
&\quad \ge \|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| - 2\zeta_s.
\end{aligned}$$

For every Byzantine client $i \in \mathcal{R}_s \setminus \mathcal{H}$, no bound on $u_{s,i}^t$ is needed: the reverse triangle inequality gives

$$\|\hat{g}_s^t - u_{s,i}^t\| - \|\nabla f_s^{\mathcal{H}}(x^t) - u_{s,i}^t\| \geq -\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|.$$

Splitting the weighted sum into its honest and Byzantine parts now yields

$$\begin{aligned}
&\sum\nolimits_{i\in\mathcal{R}_s} \alpha_{s,i}\big(\|\hat{g}_s^t - u_{s,i}^t\| - \|\nabla f_s^{\mathcal{H}}(x^t) - u_{s,i}^t\|\big) \\
&\quad\geq \sum\nolimits_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i}\big(\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| - 2\zeta_s\big) - \sum\nolimits_{i\in\mathcal{R}_s\setminus\mathcal{H}} \alpha_{s,i}\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \\
&\quad= \alpha_s^{\mathcal{H}}\big(\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| - 2\zeta_s\big) - (1-\alpha_s^{\mathcal{H}})\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \\
&\quad= (2\alpha_s^{\mathcal{H}} - 1)\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| - 2\alpha_s^{\mathcal{H}}\zeta_s.
\end{aligned}$$

Combining this lower bound with ϵ-optimality gives

$$(2\alpha_s^{\mathcal{H}} - 1)\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \leq 2\alpha_s^{\mathcal{H}}\zeta_s + \epsilon.$$

Finally, $\alpha_s^{\mathcal{H}} > 1/2$ makes the coefficient on the left positive, so division by it yields equation 4. □

C.2 Proof of Proposition 3.6 and Derivation of Equation 6

Proposition 3.6 (Certified descent margin). *Let $v^t \in \mathbb{R}^d$ be a nonzero server-side candidate direction. Then $-v^t$ is a common-descent direction for every task-gradient matrix consistent with the error bounds in equation 4 if and only if $\langle \hat{g}_s^t, v^t\rangle > r_s\|v^t\|$ for every $s \in [S]$.*

Proof. Sufficiency. For each task s, let g_s be any gradient satisfying $\|g_s - \hat{g}_s^t\| \leq r_s$. Cauchy–Schwarz gives

$$\begin{aligned}
\langle g_s, v^t\rangle &= \langle \hat{g}_s^t, v^t\rangle + \langle g_s - \hat{g}_s^t, v^t\rangle \\
&\geq \langle \hat{g}_s^t, v^t\rangle - \|g_s - \hat{g}_s^t\|\|v^t\| \\
&\geq \langle \hat{g}_s^t, v^t\rangle - r_s\|v^t\|.
\end{aligned}$$

If the stated test holds for every task, then $\langle g_s, -v^t\rangle < 0$ for every choice of gradients in these balls. In particular, Lemma 3.5 places each honest task gradient in its corresponding ball, so $-v^t$ provides common descent for the honest objectives.

Necessity. For nonzero v^t, the smallest margin over the task-wise error ball is

$$\min\nolimits_{\|g_s - \hat{g}_s^t\| \leq r_s} \langle g_s, v^t\rangle = \langle \hat{g}_s^t, v^t\rangle - r_s\|v^t\|.$$

The preceding inequality supplies the lower bound, which is attained at $g_s = \hat{g}_s^t - r_s v^t/\|v^t\|$. Indeed, this choice satisfies

$$\|g_s - \hat{g}_s^t\| = r_s, \qquad \langle g_s, v^t\rangle = \langle \hat{g}_s^t, v^t\rangle - r_s\frac{\|v^t\|^2}{\|v^t\|} = \langle \hat{g}_s^t, v^t\rangle - r_s\|v^t\|.$$

If the test fails for any task, this gradient has a nonpositive inner product with v^t, so $-v^t$ is not a strict descent direction for that task. Thus the test is necessary for common descent over every task-gradient matrix consistent with the error balls. □

Proposal margin bound. For the regularized direction $d_{\mathrm{p}}^t = \widehat{G}^t\lambda_{\mathrm{p}}^t$ from equation 5, we now derive

$$\langle \hat{g}_s^t, d_{\mathrm{p}}^t\rangle - r_s\|d_{\mathrm{p}}^t\| \geq \|d_{\mathrm{p}}^t\|^2 - \gamma/8 - r_s\|d_{\mathrm{p}}^t\|, \qquad \forall s \in [S]. \tag{6}$$

Proof. For the regularized MGDA proposal, the objective in equation 5 is differentiable and convex. Its minimizer λ_{p}^t therefore satisfies the first-order optimality condition on Δ_S. Testing this condition against the canonical vertex e_s gives

$$0 \leq \langle 2(\widehat{G}^t)^\top d_{\mathrm{p}}^t + \gamma\lambda_{\mathrm{p}}^t, e_s - \lambda_{\mathrm{p}}^t\rangle.$$

Using $\widehat{G}^t e_s = \hat{g}_s^t$ and $\widehat{G}^t\lambda_{\mathrm{p}}^t = d_{\mathrm{p}}^t$, the inner product expands as

$$\begin{aligned}
0 &\leq 2\langle d_{\mathrm{p}}^t, \widehat{G}^t e_s\rangle - 2\langle d_{\mathrm{p}}^t, \widehat{G}^t\lambda_{\mathrm{p}}^t\rangle + \gamma\langle\lambda_{\mathrm{p}}^t, e_s\rangle - \gamma\|\lambda_{\mathrm{p}}^t\|^2 \\
&= 2\langle \hat{g}_s^t, d_{\mathrm{p}}^t\rangle - 2\|d_{\mathrm{p}}^t\|^2 + \gamma\big((\lambda_{\mathrm{p}}^t)_s - \|\lambda_{\mathrm{p}}^t\|^2\big).
\end{aligned}$$

Rearranging and dividing by 2 gives

$$\langle \hat{g}_s^t, d_{\mathrm{p}}^t \rangle \geq \|d_{\mathrm{p}}^t\|^2 - \frac{\gamma}{2}\big((\lambda_{\mathrm{p}}^t)_s - \|\lambda_{\mathrm{p}}^t\|^2\big).$$

Since $\lambda_{\mathrm{p}}^t \in \Delta_S$, its coordinates lie in $[0,1]$, and $\|\lambda_{\mathrm{p}}^t\|^2 = \sum_{j\in[S]}((\lambda_{\mathrm{p}}^t)_j)^2 \geq ((\lambda_{\mathrm{p}}^t)_s)^2$. Thus

$$\begin{aligned}(\lambda_{\mathrm{p}}^t)_s - \|\lambda_{\mathrm{p}}^t\|^2 &\leq (\lambda_{\mathrm{p}}^t)_s - ((\lambda_{\mathrm{p}}^t)_s)^2 \\ &= \frac{1}{4} - \left((\lambda_{\mathrm{p}}^t)_s - \frac{1}{2}\right)^2 \leq \frac{1}{4}.\end{aligned}$$

Together with $\gamma \geq 0$, this yields

$$\langle \hat{g}_s^t, d_{\mathrm{p}}^t \rangle \geq \|d_{\mathrm{p}}^t\|^2 - \gamma/8, \quad \forall s \in [S].$$

Subtracting $r_s\|d_{\mathrm{p}}^t\|$ gives equation 6. □

C.3 Derivation of the Stationarity Bound in Equation 8

For any feasible (λ, q, m) in equation 7, nonnegativity of the radii and $q \geq \|\widehat{G}^t\lambda\|$ imply

$$m - \kappa q^2 \leq \min_{s\in[S]}\big[\langle \hat{g}_s^t, \widehat{G}^t\lambda\rangle - r_s q\big] - \kappa q^2 \leq M_t(\lambda) - \kappa\|\widehat{G}^t\lambda\|^2.$$

Conversely, for any $\lambda \in \Delta_S$, choosing $q = \|\widehat{G}^t\lambda\|$ and $m = M_t(\lambda)$ gives a feasible solution. Thus equation 7 has optimal value $\max_{\lambda\in\Delta_S}[M_t(\lambda) - \kappa\|\widehat{G}^t\lambda\|^2]$. For $\kappa = 0$, this is the maximum certified margin.

Under the task-wise error bounds in equation 4, if the optimal objective in equation 7 is nonpositive, then

$$\Psi(x^t) \leq \Psi_{\mathrm{ub}}^t := \min_{\lambda\in\Delta_S}\left[\|\widehat{G}^t\lambda\| + \sum\nolimits_{s\in[S]}\lambda_s r_s\right] \leq \frac{2-\kappa}{1-\kappa}r, \tag{8}$$

where $r = \max_{s\in[S]} r_s$. The first inequality holds at every round; only the final bound requires recovery failure.

Proof. Validity of the upper bound. For every $\lambda \in \Delta_S$, the triangle inequality and the task-wise error bounds give

$$\begin{aligned}\|G_{\mathcal{H}}^t\lambda\| &= \left\|\widehat{G}^t\lambda + \sum\nolimits_{s\in[S]}\lambda_s(\nabla f_s^{\mathcal{H}}(x^t) - \hat{g}_s^t)\right\| \\ &\leq \|\widehat{G}^t\lambda\| + \sum\nolimits_{s\in[S]}\lambda_s\|\nabla f_s^{\mathcal{H}}(x^t) - \hat{g}_s^t\| \\ &\leq \|\widehat{G}^t\lambda\| + \sum\nolimits_{s\in[S]}\lambda_s r_s.\end{aligned}$$

Since $\Psi(x^t) \leq \|G_{\mathcal{H}}^t\lambda\|$ for every $\lambda \in \Delta_S$, minimizing the right-hand side over the simplex yields $\Psi(x^t) \leq \Psi_{\mathrm{ub}}^t$.

Bound after recovery failure. Now suppose the optimal recovery objective is nonpositive. Let $\lambda_0^t \in \arg\min_{\lambda\in\Delta_S}\|\widehat{G}^t\lambda\|$ and set $d_0^t = \widehat{G}^t\lambda_0^t$. The minimum-norm point d_0^t is the projection of the origin onto $\mathrm{conv}(\widehat{G}^t)$. Since each $\hat{g}_s^t$ belongs to this convex hull, projection optimality gives

$$\langle \hat{g}_s^t - d_0^t, d_0^t\rangle \geq 0, \quad \text{and hence} \quad \langle \hat{g}_s^t, d_0^t\rangle \geq \|d_0^t\|^2, \qquad \forall s \in [S].$$

Substituting the definition of M_t and using $r_s \leq r$, we obtain

$$\begin{aligned}M_t(\lambda_0^t) &= \min_{s\in[S]}\big[\langle \hat{g}_s^t, d_0^t\rangle - r_s\|d_0^t\|\big] \\ &\geq \|d_0^t\|^2 - r\|d_0^t\|.\end{aligned}$$

If $\|d_0^t\| > r/(1-\kappa)$, then

$$M_t(\lambda_0^t) - \kappa\|d_0^t\|^2 \geq (1-\kappa)\|d_0^t\|^2 - r\|d_0^t\| > 0.$$

Taking $\lambda = \lambda_0^t$, $q = \|d_0^t\|$, and $m = M_t(\lambda_0^t)$ satisfies every constraint of equation 7 and gives this positive objective value, contradicting recovery failure. Therefore $\|d_0^t\| \leq r/(1-\kappa)$. Evaluating the stationarity bound at λ_0^t and using $\sum_{s\in[S]}(\lambda_0^t)_s = 1$, we conclude that

$$\Psi_{\mathrm{ub}}^t \leq \|d_0^t\| + \sum\nolimits_{s\in[S]}(\lambda_0^t)_s r_s \leq \frac{r}{1-\kappa} + r = \frac{2-\kappa}{1-\kappa}r.$$

□

C.4 Alternative Robust Aggregation Radii

We derive task-wise error radii for Multi-Krum, minimum-diameter averaging (MDA), coordinate-wise median, and coordinate-wise trimmed mean under the full-gradient reporting model of Section 3.1 and Assumption 3.3. Table 5 summarizes the resulting bounds alongside the TWGM radius in Lemma 3.5. For each rule, $\hat{g}_s^t$ denotes its output and r_s its error radius. Fix a task s and round t. Since $\mathcal{B}$ is the global Byzantine set, task s contains at most $|\mathcal{B}|$ Byzantine reports. For Multi-Krum, MDA, and trimmed mean, we take $|\mathcal{B}|$ as a given count parameter. A known task-specific upper bound on $|\mathcal{B} \cap \mathcal{R}_s|$ may replace $|\mathcal{B}|$ consistently in the aggregation rules, conditions, and bounds below.

Table 5: Task-wise error radii under full-gradient reporting and Assumption 3.3. Multi-Krum selects $1 \le m_s \le |\mathcal{R}_s| - |\mathcal{B}|$ reports, with $m_s = 1$ giving Krum; trimmed mean removes $|\mathcal{B}|$ values at each end of each coordinate.

Aggregator	**Conditions**	**Error radius**
TWGM	$\alpha_s^{\mathcal{H}} > 1/2$	$\dfrac{2\alpha_s^{\mathcal{H}}\zeta_s + \epsilon}{2\alpha_s^{\mathcal{H}} - 1}$
Multi-Krum / Krum	$\lvert\mathcal{R}_s\rvert \ge 2\lvert\mathcal{B}\rvert + 3$	$\left(1 + 2\sqrt{\dfrac{\lvert\mathcal{R}_s\rvert - \lvert\mathcal{B}\rvert - 2}{\lvert\mathcal{R}_s\rvert - 2\lvert\mathcal{B}\rvert - 1}}\right)\zeta_s$
MDA	$\lvert\mathcal{R}_s\rvert > 2\lvert\mathcal{B}\rvert$	$\left(1 + \dfrac{2\lvert\mathcal{B}\rvert}{\lvert\mathcal{R}_s\rvert - \lvert\mathcal{B}\rvert}\right)\zeta_s$
Coordinate-wise median	$\lvert\mathcal{R}_s\rvert > 2\lvert\mathcal{B}\rvert$	$\sqrt{d}\,\zeta_s$
Coordinate-wise trimmed mean	$\lvert\mathcal{R}_s\rvert > 2\lvert\mathcal{B}\rvert$	$\sqrt{d}\,\zeta_s$

C.4.1 Multi-Krum and Krum

Multi-Krum first scores each report by its squared distances to a local neighborhood. Under $|\mathcal{R}_s| \ge 2|\mathcal{B}| + 3$, define

$$\nu_s := |\mathcal{R}_s| - |\mathcal{B}| - 2.$$

For each $i \in \mathcal{R}_s$, let $\mathcal{N}_s(i)$ contain the ν_s reports closest to $u_{s,i}^t$, excluding i, and set

$$\mathrm{score}_s(i) := \sum\nolimits_{j \in \mathcal{N}_s(i)} \|u_{s,i}^t - u_{s,j}^t\|^2.$$

For a prescribed integer m_s with $1 \le m_s \le |\mathcal{R}_s| - |\mathcal{B}|$, Multi-Krum selects a set $\mathcal{I}_s$ of the m_s smallest-score reports and returns their average (Blanchard et al., 2017):

$$\hat{g}_s^t := \frac{1}{|\mathcal{I}_s|} \sum\nolimits_{i \in \mathcal{I}_s} u_{s,i}^t.$$

When $m_s = 1$, this rule reduces to Krum. Ties in the neighborhoods and scores are broken arbitrarily; the bound below holds for any such tie-breaking rule.

Lemma C.1 (Task-wise Multi-Krum radius)**.** *Suppose Assumption 3.3 holds and $|\mathcal{R}_s| \ge 2|\mathcal{B}| + 3$. Let $\mathcal{I}_s \subseteq \mathcal{R}_s$ consist of the m_s smallest-score indices, with $1 \le m_s = |\mathcal{I}_s| \le |\mathcal{R}_s| - |\mathcal{B}|$. Then*

$$\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_s := \left(1 + 2\sqrt{\frac{|\mathcal{R}_s| - |\mathcal{B}| - 2}{|\mathcal{R}_s| - 2|\mathcal{B}| - 1}}\right)\zeta_s. \tag{9}$$

Proof. Fix an honest $i \in \mathcal{R}_s \cap \mathcal{H}$. By Assumption 3.3 and the triangle inequality, every other honest report j satisfies

$$\|u_{s,i}^t - u_{s,j}^t\| \le \|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| + \|u_{s,j}^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le 2\zeta_s.$$

There are at least $|\mathcal{R}_s| - |\mathcal{B}| - 1 > \nu_s$ other honest reports. Since $\mathcal{N}_s(i)$ contains the ν_s closest reports, its sum of squared distances is no larger than that of any ν_s other reports. Comparing with ν_s honest reports therefore gives

$$\mathrm{score}_s(i) \le 4\nu_s\zeta_s^2.$$

There are at least $|\mathcal{R}_s| - |\mathcal{B}|$ honest reports with this score bound. Because $m_s \le |\mathcal{R}_s| - |\mathcal{B}|$, the m_s-th smallest score is at most $4\nu_s\zeta_s^2$. Hence, every selected index $i \in \mathcal{I}_s$ satisfies the same bound.

If a selected report is honest, its error is at most ζ_s. Now consider a selected Byzantine report i. The desired bound is immediate if $\|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le \zeta_s$. Otherwise, since i is itself Byzantine and is excluded from $\mathcal{N}_s(i)$, at most $|\mathcal{B}| - 1$ of its neighbors are Byzantine. Thus the neighborhood contains at least

$$\nu_s - (|\mathcal{B}| - 1) = |\mathcal{R}_s| - 2|\mathcal{B}| - 1$$

honest reports. For each such honest neighbor j, the reverse triangle inequality gives

$$\|u_{s,i}^t - u_{s,j}^t\| \ge \|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| - \zeta_s.$$

The right-hand side is positive in the case under consideration. Squaring and summing over the honest neighbors gives

$$\begin{aligned}\mathrm{score}_s(i) &\ge \sum\nolimits_{j \in \mathcal{N}_s(i) \cap \mathcal{H}} \|u_{s,i}^t - u_{s,j}^t\|^2 \\ &\ge (|\mathcal{R}_s| - 2|\mathcal{B}| - 1)\big(\|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| - \zeta_s\big)^2.\end{aligned}$$

Combining this lower bound with $\mathrm{score}_s(i) \le 4\nu_s\zeta_s^2$ and using $|\mathcal{R}_s| - 2|\mathcal{B}| - 1 > 0$, we obtain

$$\big(\|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| - \zeta_s\big)^2 \le \frac{4\nu_s\zeta_s^2}{|\mathcal{R}_s| - 2|\mathcal{B}| - 1}.$$

Taking square roots and substituting $\nu_s = |\mathcal{R}_s| - |\mathcal{B}| - 2$ yields

$$\|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le \left(1 + 2\sqrt{\frac{|\mathcal{R}_s| - |\mathcal{B}| - 2}{|\mathcal{R}_s| - 2|\mathcal{B}| - 1}}\right)\zeta_s = r_s.$$

Thus, every selected report satisfies the radius bound. Convexity of the norm then gives

$$\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le \frac{1}{|\mathcal{I}_s|}\sum\nolimits_{i \in \mathcal{I}_s} \|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_s.$$

□

C.4.2 Minimum-Diameter Averaging

MDA selects a subset $\mathcal{I}_s \subseteq \mathcal{R}_s$ of $|\mathcal{R}_s| - |\mathcal{B}|$ reports whose maximum pairwise Euclidean distance is smallest among all subsets of this size, and returns their average (Farhadkhani et al., 2022):

$$\hat{g}_s^t := \frac{1}{|\mathcal{I}_s|}\sum\nolimits_{i \in \mathcal{I}_s} u_{s,i}^t.$$

Ties are broken arbitrarily; the bound below is independent of the tie-breaking rule.

Lemma C.2 (Task-wise MDA radius)**.** *Suppose Assumption 3.3 holds and $|\mathcal{R}_s| > 2|\mathcal{B}|$. If $\hat{g}_s^t$ is the MDA output, then*

$$\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_s := \left(1 + \frac{2|\mathcal{B}|}{|\mathcal{R}_s| - |\mathcal{B}|}\right)\zeta_s.$$

Proof. By Assumption 3.3 and the triangle inequality, the distance between any two honest reports is at most $2\zeta_s$. There are at least $|\mathcal{R}_s| - |\mathcal{B}|$ honest reports, so an all-honest subset of the size used by MDA has diameter at most $2\zeta_s$. The defining minimum-diameter property therefore gives

$$\max_{i,j \in \mathcal{I}_s} \|u_{s,i}^t - u_{s,j}^t\| \le 2\zeta_s.$$

Moreover, $|\mathcal{I}_s| = |\mathcal{R}_s| - |\mathcal{B}| > |\mathcal{B}|$, so $\mathcal{I}_s$ contains an honest report. For each selected Byzantine report i, choose any $j \in \mathcal{I}_s \cap \mathcal{H}$. The diameter bound and Assumption 3.3 give

$$\begin{aligned}\|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| &\le \|u_{s,i}^t - u_{s,j}^t\| + \|u_{s,j}^t - \nabla f_s^{\mathcal{H}}(x^t)\| \\ &\le 2\zeta_s + \zeta_s = 3\zeta_s.\end{aligned}$$

Selected honest reports remain within ζ_s of the same target. Since $|\mathcal{I}_s \cap \mathcal{H}| + |\mathcal{I}_s \cap \mathcal{B}| = |\mathcal{I}_s|$, convexity of the norm yields

$$\begin{aligned}\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| &\le \frac{\sum_{i\in\mathcal{I}_s\cap\mathcal{H}} \|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| + \sum_{i\in\mathcal{I}_s\cap\mathcal{B}} \|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\|}{|\mathcal{I}_s|} \\ &\le \frac{|\mathcal{I}_s \cap \mathcal{H}|\zeta_s + 3|\mathcal{I}_s \cap \mathcal{B}|\zeta_s}{|\mathcal{I}_s|} \\ &= \left(1 + \frac{2|\mathcal{I}_s \cap \mathcal{B}|}{|\mathcal{I}_s|}\right)\zeta_s \\ &\le \left(1 + \frac{2|\mathcal{B}|}{|\mathcal{R}_s| - |\mathcal{B}|}\right)\zeta_s = r_s.\end{aligned}$$

□

C.4.3 Coordinate-Wise Median and Trimmed Mean

The next two rules aggregate each coordinate independently (Yin et al., 2018). The coordinate-wise median sets its kth coordinate to any median of the submitted values $\{u_{s,i,k}^t \mid i \in \mathcal{R}_s\}$. The coordinate-wise $|\mathcal{B}|$-trimmed mean instead sorts these values, removes the $|\mathcal{B}|$ smallest and $|\mathcal{B}|$ largest, and averages the remainder.

By Assumption 3.3, every honest coordinate satisfies

$$|u_{s,i,k}^t - (\nabla f_s^{\mathcal{H}}(x^t))_k| \le \|u_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le \zeta_s,$$

and hence

$$u_{s,i,k}^t \in [(\nabla f_s^{\mathcal{H}}(x^t))_k - \zeta_s, (\nabla f_s^{\mathcal{H}}(x^t))_k + \zeta_s], \qquad i \in \mathcal{R}_s \cap \mathcal{H}, \quad k \in [d]. \tag{10}$$

Controlling the coordinates separately therefore introduces a conservative $\sqrt{d}$ factor when the final error is measured in the Euclidean geometry used by MGDA.

Lemma C.3 (Task-wise radius for coordinate-wise aggregation)**.** *Suppose Assumption 3.3 holds and $|\mathcal{R}_s| > 2|\mathcal{B}|$. If $\hat{g}_s^t$ is either the coordinate-wise median or the coordinate-wise $|\mathcal{B}|$-trimmed mean, then*

$$\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_s := \sqrt{d}\,\zeta_s. \tag{11}$$

Proof. Fix a coordinate $k \in [d]$. The number of honest reports satisfies

$$|\mathcal{R}_s \cap \mathcal{H}| = |\mathcal{R}_s| - |\mathcal{R}_s \setminus \mathcal{H}| \ge |\mathcal{R}_s| - |\mathcal{B}| > \frac{|\mathcal{R}_s|}{2}.$$

Hence, more than half of the submitted values lie in the interval in equation 10, so any median also lies in this interval.

For the trimmed mean, every value below this interval is Byzantine. There are at most $|\mathcal{B}|$ such values, so all of them are removed among the $|\mathcal{B}|$ smallest values. Similarly, all values above the interval are removed among the $|\mathcal{B}|$ largest values. The condition $|\mathcal{R}_s| > 2|\mathcal{B}|$ ensures that the retained set is nonempty. All retained values lie in the same interval, and so does their average. Thus, for either rule,

$$|\hat{g}_{s,k}^t - (\nabla f_s^{\mathcal{H}}(x^t))_k| \le \zeta_s, \qquad k \in [d].$$

Summing the squared coordinate errors gives

$$\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 = \sum\nolimits_{k=1}^{d} |\hat{g}_{s,k}^t - (\nabla f_s^{\mathcal{H}}(x^t))_k|^2 \le d\zeta_s^2,$$

which proves equation 11. □

TWGM uses the same task-wise client weights as the honest objectives and provides a dimension-independent radius under Assumption 3.1. Its radius also accounts explicitly for the aggregation tolerance ϵ. A conservative upper bound on ζ_s and a lower bound on $\alpha_s^{\mathcal{H}}$ strictly above $1/2$ can therefore be substituted into equation 4 to obtain a valid radius for certification. Under their respective assumptions, the alternative radii above can replace the TWGM radius in the full-gradient certificates and convergence analysis.

D Proofs and Derivations for Section 3.3

This section proves the stochastic error bounds and probabilistic certificates in Section 3.3, including their extension to the full reporting trajectory.

D.1 Local GMOM Stability

We first state the geometric fact used in the GMOM analysis. It is the Euclidean specialization of Lemma 2.1 of Minsker (2015); we include the proof for completeness. The statement is deterministic: once the block averages are fixed, it uses no independence or distributional assumption.

Lemma D.1 (Local GMOM stability)**.** *Fix s, t and $i \in \mathcal{R}_s \cap \mathcal{H}$, and let*

$$\tilde{u}_{s,i}^t \in \arg\min_{u \in \mathbb{R}^d} \sum_{k=1}^{K_{s,t}} \|u - \bar{u}_{s,i,k}^t\|$$

be the geometric median of the block averages. Call block k accurate if

$$\|\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t)\| \le \frac{\sqrt{10}\sigma}{\sqrt{b_{s,t}}},$$

and inaccurate otherwise. If at least $11K_{s,t}/18$ blocks are accurate, then

$$\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\| \le \frac{11\sqrt{5}\,\sigma}{6\sqrt{b_{s,t}}} = \tau_{s,t}.$$

Equivalently, if the GMOM bound above fails, then more than $7K_{s,t}/18$ blocks must be inaccurate.

Proof. If $\sigma = 0$, all accurate block averages equal the client's full gradient. Since they form a strict majority, moving away from this gradient increases their total distance by more than the remaining blocks can decrease it. The geometric median therefore equals the full gradient, and the claim holds.

For the remainder of the proof, suppose $\sigma > 0$. We show that a report violating the claimed bound can be moved toward the client's full gradient to decrease the sum of distances, contradicting geometric-median optimality. Suppose that at least $11K_{s,t}/18$ blocks are accurate but that the claimed GMOM bound fails. Define the unit vector pointing from the GMOM report toward the client's full gradient by

$$v := \frac{\nabla f_{s,i}(x^t) - \tilde{u}_{s,i}^t}{\|\nabla f_{s,i}(x^t) - \tilde{u}_{s,i}^t\|}.$$

Geometry of accurate blocks. We first compare the direction of an accurate block with v, both viewed from $\tilde{u}_{s,i}^t$. For such a block, the assumed failure and the definition of accuracy give

$$\frac{\|\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t)\|}{\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\|} < \frac{\sqrt{10}\sigma/\sqrt{b_{s,t}}}{11\sqrt{5}\,\sigma/(6\sqrt{b_{s,t}})} = \frac{6\sqrt{2}}{11}.$$

Since $6\sqrt{2}/11 < 1$, an accurate block average $\bar{u}_{s,i,k}^t$ cannot equal $\tilde{u}_{s,i}^t$. Let $\theta \in [0, \pi]$ be the angle between $\bar{u}_{s,i,k}^t - \tilde{u}_{s,i}^t$ and v. The line through $\tilde{u}_{s,i}^t$ and $\bar{u}_{s,i,k}^t$ has direction $\bar{u}_{s,i,k}^t - \tilde{u}_{s,i}^t$; denote this line by L. Because L passes through $\tilde{u}_{s,i}^t$, elementary right-triangle geometry gives

$$\|\nabla f_{s,i}(x^t) - \tilde{u}_{s,i}^t\| \sin\theta = \operatorname{dist}(\nabla f_{s,i}(x^t), L).$$

Moreover, $\bar{u}_{s,i,k}^t \in L$, whereas the perpendicular distance is the smallest distance to any point on L. Therefore,

$$\sin\theta = \frac{\operatorname{dist}(\nabla f_{s,i}(x^t), L)}{\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\|} \le \frac{\|\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t)\|}{\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\|} < \frac{6\sqrt{2}}{11}.$$

The sine bound alone does not exclude an angle close to π. To rule this out, decompose $\bar{u}_{s,i,k}^t - \tilde{u}_{s,i}^t = (\nabla f_{s,i}(x^t) - \tilde{u}_{s,i}^t) + (\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t))$ and project onto v:

$$\begin{aligned}\langle \bar{u}_{s,i,k}^t - \tilde{u}_{s,i}^t, v\rangle &= \|\nabla f_{s,i}(x^t) - \tilde{u}_{s,i}^t\| + \langle \bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t), v\rangle \\ &\ge \|\nabla f_{s,i}(x^t) - \tilde{u}_{s,i}^t\| - \|\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t)\| > 0.\end{aligned}$$

The inequality uses Cauchy–Schwarz and $\|v\| = 1$; its final strict positivity follows from the preceding distance ratio, which is smaller than 1. Hence $\cos\theta > 0$, so θ is acute. Combining this fact with the sine bound gives $\cos\theta > \sqrt{1-(6\sqrt{2}/11)^2} = 7/11$, or equivalently,

$$\cos\theta = \left\langle \frac{\bar{u}^t_{s,i,k} - \tilde{u}^t_{s,i}}{\|\bar{u}^t_{s,i,k} - \tilde{u}^t_{s,i}\|}, v \right\rangle > \frac{7}{11}.$$

Contradiction with geometric-median optimality. Let $\Phi(u) = \sum_{k=1}^{K_{s,t}} \|u - \bar{u}^t_{s,i,k}\|$. Since $\tilde{u}^t_{s,i}$ minimizes Φ, moving from $\tilde{u}^t_{s,i}$ a small distance along v cannot decrease Φ. Thus

$$\lim_{a\downarrow 0} \frac{\Phi(\tilde{u}^t_{s,i} + av) - \Phi(\tilde{u}^t_{s,i})}{a} \geq 0.$$

We now compute how each distance term changes under this move. For an accurate block, its block average differs from $\tilde{u}^t_{s,i}$, so the norm is differentiable there and

$$\lim_{a\downarrow 0} \frac{\|\tilde{u}^t_{s,i} + av - \bar{u}^t_{s,i,k}\| - \|\tilde{u}^t_{s,i} - \bar{u}^t_{s,i,k}\|}{a} = \left\langle \frac{\tilde{u}^t_{s,i} - \bar{u}^t_{s,i,k}}{\|\tilde{u}^t_{s,i} - \bar{u}^t_{s,i,k}\|}, v \right\rangle < -\frac{7}{11}.$$

Thus moving along v decreases the distance to each accurate block at a rate greater than $7/11$. For every other block, the triangle inequality gives, for $a > 0$,

$$\|\tilde{u}^t_{s,i} + av - \bar{u}^t_{s,i,k}\| - \|\tilde{u}^t_{s,i} - \bar{u}^t_{s,i,k}\| \leq a\|v\| = a.$$

Hence each remaining block can increase the objective at a rate of at most one. This bound also covers a block average equal to $\tilde{u}^t_{s,i}$, for which the one-sided rate is exactly one.

The one-sided derivative of the finite sum Φ is the sum of these per-block derivatives. There are at least $11K_{s,t}/18$ accurate blocks and at most $7K_{s,t}/18$ remaining blocks. Therefore, its derivative along v is strictly smaller than

$$-\frac{11K_{s,t}}{18}\frac{7}{11} + \frac{7K_{s,t}}{18} = 0,$$

because the guaranteed decrease from the accurate blocks outweighs the largest possible increase from all remaining blocks. This contradicts the nonnegative derivative required at a minimizer. Thus the GMOM bound holds whenever at least $11K_{s,t}/18$ blocks are accurate. Taking the contrapositive gives the equivalent statement about inaccurate blocks. □

D.2 Proof of Lemma 3.8

Lemma 3.8 (Stochastic TWGM error). *Suppose Assumptions 3.1, 3.3, and 3.7 hold. Fix s, t, and let $K_{s,t} \leq B/2$ satisfy $|\mathcal{R}_s| \exp(-K_{s,t}/3.5) \leq \delta_{s,t}$. Then*

$$\Pr\Big(\|\tilde{g}^t_s - \nabla f^{\mathcal{H}}_s(x^t)\| \leq r^{\mathrm{hp}}_{s,t} \mid \mathcal{F}_t\Big) \geq 1 - \delta_{s,t}, \quad r^{\mathrm{hp}}_{s,t} := \frac{2\alpha^{\mathcal{H}}_s(\zeta_s + \tau_{s,t}) + \epsilon}{2\alpha^{\mathcal{H}}_s - 1}, \quad \tau_{s,t} := \frac{11\sqrt{5}\,\sigma}{6\sqrt{b_{s,t}}},$$

$$\mathbb{E}[\|\tilde{g}^t_s - \nabla f^{\mathcal{H}}_s(x^t)\|^2 \mid \mathcal{F}_t] \leq (R^{\mathrm{ms}}_{s,t})^2 := \frac{8(\alpha^{\mathcal{H}}_s)^2(\zeta_s + 2\sigma/\sqrt{b_{s,t}})^2 + 2\epsilon^2}{(2\alpha^{\mathcal{H}}_s - 1)^2}.$$

Proof. High-probability bound. Fix s, t and one honest client $i \in \mathcal{R}_s \cap \mathcal{H}$, and condition on $\mathcal{F}_t$. Each block contains $b_{s,t}$ conditionally independent samples. Conditional unbiasedness makes the cross terms between distinct samples vanish, and the variance bound in Assumption 3.7 therefore gives

$$\begin{aligned} &\mathbb{E}\big[\|\bar{u}^t_{s,i,k} - \nabla f_{s,i}(x^t)\|^2 \mid \mathcal{F}_t\big] \\ &\quad = \frac{1}{b^2_{s,t}} \sum\nolimits_{\xi \in \Xi^t_{s,i,k}} \mathbb{E}\big[\|\nabla f_{s,i}(x^t, \xi) - \nabla f_{s,i}(x^t)\|^2 \mid \mathcal{F}_t\big] \leq \frac{\sigma^2}{b_{s,t}}. \end{aligned} \tag{12}$$

If $\sigma = 0$, every block average, and hence its geometric median, equals the client's full gradient almost surely, so the local high-probability bound holds directly. For $\sigma > 0$, applying Markov's inequality to the squared norm yields

$$\Pr\left(\|\bar{u}^t_{s,i,k} - \nabla f_{s,i}(x^t)\| > \frac{\sqrt{10}\sigma}{\sqrt{b_{s,t}}} \,\middle|\, \mathcal{F}_t \right) \le \frac{\sigma^2/b_{s,t}}{10\sigma^2/b_{s,t}} = 0.1.$$

Thus each block average lies within the displayed radius of the client's full gradient with conditional probability at least 0.9. Call a block inaccurate when this bound fails. The blocks of one client are conditionally independent because they use disjoint sets of independent samples.

The remaining argument shows how the geometric median boosts this constant-probability guarantee to high probability. By Lemma D.1, the local report can be farther than $\tau_{s,t}$ from $\nabla f_{s,i}(x^t)$ only if more than $7K_{s,t}/18$ blocks are inaccurate.

To bound the probability of this joint failure, define

$$Z_k := \mathbf{1}\left\{ \|\bar{u}^t_{s,i,k} - \nabla f_{s,i}(x^t)\| > \frac{\sqrt{10}\sigma}{\sqrt{b_{s,t}}} \right\}, \qquad Z := \sum\nolimits_{k=1}^{K_{s,t}} Z_k.$$

Thus Z counts the inaccurate blocks of the fixed client. Conditional on $\mathcal{F}_t$, the indicators are independent and $\Pr(Z_k = 1 \mid \mathcal{F}_t) \le 0.1$. For any $\lambda > 0$, Markov's inequality applied to $e^{\lambda Z}$ gives

$$\begin{aligned}
\Pr\left(Z \ge \frac{7}{18}K_{s,t} \,\middle|\, \mathcal{F}_t \right) &= \Pr\left(e^{\lambda Z} \ge e^{\lambda(7/18)K_{s,t}} \,\middle|\, \mathcal{F}_t \right) \\
&\le e^{-\lambda(7/18)K_{s,t}} \mathbb{E}[e^{\lambda Z} \mid \mathcal{F}_t] \\
&\le \left[e^{-7\lambda/18}(0.9 + 0.1e^{\lambda}) \right]^{K_{s,t}}.
\end{aligned}$$

The final inequality uses conditional independence across the blocks:

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z} \mid \mathcal{F}_t] &= \mathbb{E}\left[\prod\nolimits_{k=1}^{K_{s,t}} e^{\lambda Z_k} \,\middle|\, \mathcal{F}_t \right] \\
&= \prod\nolimits_{k=1}^{K_{s,t}} \mathbb{E}[e^{\lambda Z_k} \mid \mathcal{F}_t] \le (0.9 + 0.1e^{\lambda})^{K_{s,t}}.
\end{aligned}$$

The last bound follows because, for each block,

$$\mathbb{E}[e^{\lambda Z_k} \mid \mathcal{F}_t] = \Pr(Z_k = 0 \mid \mathcal{F}_t)e^0 + \Pr(Z_k = 1 \mid \mathcal{F}_t)e^{\lambda} \le 0.9 + 0.1e^{\lambda}.$$

Minimizing the last bound over $\lambda > 0$ yields the binary Chernoff exponent. Indeed, differentiating the logarithm of its bracketed term gives the optimality condition $-7/18 + 0.1e^{\lambda}/(0.9 + 0.1e^{\lambda}) = 0$, hence $e^{\lambda} = 63/11$. Substitution gives

$$\Pr\left(Z \ge \frac{7}{18}K_{s,t} \,\middle|\, \mathcal{F}_t \right) \le \exp\left[-K_{s,t} \left(\frac{11}{18} \log \frac{11/18}{0.9} + \frac{7}{18} \log \frac{7/18}{0.1} \right) \right].$$

The expression in parentheses is the binary relative entropy between the threshold failure rate $7/18$ and the per-block upper bound 0.1, and it is approximately 0.292, which is greater than $1/3.5 \approx 0.286$. Moreover,

$$\left\{ \|\tilde{u}^t_{s,i} - \nabla f_{s,i}(x^t)\| > \tau_{s,t} \right\} \subseteq \left\{ Z > \frac{7}{18}K_{s,t} \right\} \subseteq \left\{ Z \ge \frac{7}{18}K_{s,t} \right\}.$$

Hence the local-report failure probability is at most $\exp(-K_{s,t}/3.5)$. Therefore, for every honest client,

$$\Pr\left(\|\tilde{u}^t_{s,i} - \nabla f_{s,i}(x^t)\| > \tau_{s,t} \mid \mathcal{F}_t \right) \le \exp(-K_{s,t}/3.5) \le \frac{\delta_{s,t}}{|\mathcal{R}_s|},$$

where the final inequality is the condition on $K_{s,t}$ in the lemma. Independence across clients is unnecessary: a union bound over at most $|\mathcal{R}_s|$ honest clients shows that, with conditional probability at least $1 - \delta_{s,t}$, every $i \in \mathcal{R}_s \cap \mathcal{H}$ satisfies

$$\|\tilde{u}^t_{s,i} - \nabla f^{\mathcal{H}}_s(x^t)\| \le \|\tilde{u}^t_{s,i} - \nabla f_{s,i}(x^t)\| + \|\nabla f_{s,i}(x^t) - \nabla f^{\mathcal{H}}_s(x^t)\| \le \tau_{s,t} + \zeta_s.$$

We next record the deterministic TWGM inequality used in both parts of the proof. It holds for every realization of the submitted vectors, independently of the simultaneous event above. For every honest client, the triangle inequality gives

$$\begin{aligned}&\|\tilde{g}_s^t - \tilde{u}_{s,i}^t\| - \|\nabla f_s^{\mathcal{H}}(x^t) - \tilde{u}_{s,i}^t\| \\ &\quad \geq \|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| - 2\|\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\|.\end{aligned}$$

For every Byzantine client, the reverse triangle inequality gives

$$\|\tilde{g}_s^t - \tilde{u}_{s,i}^t\| - \|\nabla f_s^{\mathcal{H}}(x^t) - \tilde{u}_{s,i}^t\| \geq -\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|.$$

By ϵ-optimality of $\tilde{g}_s^t$, the weighted sum of the left-hand sides is at most ϵ. Splitting this sum into its honest and Byzantine parts, and using that their weights sum to $\alpha_s^{\mathcal{H}}$ and $1 - \alpha_s^{\mathcal{H}}$, respectively, therefore gives

$$(2\alpha_s^{\mathcal{H}} - 1)\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \leq 2\sum\nolimits_{i \in \mathcal{R}_s \cap \mathcal{H}} \alpha_{s,i}\|\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\| + \epsilon. \tag{13}$$

On the simultaneous event established above, the right-hand side is at most $2\alpha_s^{\mathcal{H}}(\zeta_s + \tau_{s,t}) + \epsilon$, because the honest weights sum to $\alpha_s^{\mathcal{H}}$. Division by $2\alpha_s^{\mathcal{H}} - 1 > 0$ proves the stated conditional coverage of $r_{s,t}^{\mathrm{hp}}$.

Mean-square bound. This part does not condition on the preceding high-probability event; all expectations below are conditional only on $\mathcal{F}_t$. Fix an honest client $i \in \mathcal{R}_s \cap \mathcal{H}$. Since $\tilde{u}_{s,i}^t$ minimizes the sum of distances to the block averages, its objective value cannot exceed the value at the client's full gradient:

$$\sum\nolimits_{k=1}^{K_{s,t}} \|\tilde{u}_{s,i}^t - \bar{u}_{s,i,k}^t\| \leq \sum\nolimits_{k=1}^{K_{s,t}} \|\nabla f_{s,i}(x^t) - \bar{u}_{s,i,k}^t\|.$$

For every block k, the triangle inequality gives

$$\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\| \leq \|\tilde{u}_{s,i}^t - \bar{u}_{s,i,k}^t\| + \|\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t)\|.$$

Summing over k and applying the preceding comparison inequality yields

$$K_{s,t}\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\| \leq 2\sum\nolimits_{k=1}^{K_{s,t}} \|\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t)\|.$$

Squaring this inequality and applying Cauchy–Schwarz to the sum over the blocks gives

$$\begin{aligned}K_{s,t}^2\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\|^2 &\leq 4\left(\sum\nolimits_{k=1}^{K_{s,t}} \|\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t)\|\right)^2 \\ &\leq 4K_{s,t}\sum\nolimits_{k=1}^{K_{s,t}} \|\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t)\|^2.\end{aligned}$$

After division by $K_{s,t}^2$, taking conditional expectations and applying equation 12 to each block gives

$$\begin{aligned}&\mathbb{E}[\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\|^2 \mid \mathcal{F}_t] \\ &\quad \leq \frac{4}{K_{s,t}}\sum\nolimits_{k=1}^{K_{s,t}} \mathbb{E}[\|\bar{u}_{s,i,k}^t - \nabla f_{s,i}(x^t)\|^2 \mid \mathcal{F}_t] \leq \frac{4\sigma^2}{b_{s,t}}.\end{aligned} \tag{14}$$

Next, decompose the local-to-task error as

$$\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t) = \big(\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\big) + \big(\nabla f_{s,i}(x^t) - \nabla f_s^{\mathcal{H}}(x^t)\big).$$

Conditional on $\mathcal{F}_t$, the second term is fixed. Conditional Minkowski inequality, followed by equation 14 and Assumption 3.3, gives

$$\begin{aligned}&\big(\mathbb{E}[\|\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \mid \mathcal{F}_t]\big)^{1/2} \\ &\quad \leq \big(\mathbb{E}[\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\|^2 \mid \mathcal{F}_t]\big)^{1/2} + \|\nabla f_{s,i}(x^t) - \nabla f_s^{\mathcal{H}}(x^t)\| \\ &\quad \leq \zeta_s + \frac{2\sigma}{\sqrt{b_{s,t}}}.\end{aligned} \tag{15}$$

It remains to propagate the client-level errors through TWGM. As noted above, inequality equation 13 holds pointwise for every realization of the submitted reports. Its use here therefore neither

conditions on the preceding high-probability event nor requires independence across clients. Squaring that inequality and using $(a+b)^2 \le 2a^2 + 2b^2$ gives

$$\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \le \frac{8}{(2\alpha_s^{\mathcal{H}}-1)^2}\left(\sum\nolimits_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i}\|\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\|\right)^2 + \frac{2\epsilon^2}{(2\alpha_s^{\mathcal{H}}-1)^2}.$$

For the first term, weighted Cauchy–Schwarz gives

$$\left(\sum\nolimits_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i}\|\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\|\right)^2 \le \alpha_s^{\mathcal{H}} \sum\nolimits_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i}\|\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2.$$

This gives the pointwise squared-error bound

$$\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \le \frac{8\alpha_s^{\mathcal{H}}}{(2\alpha_s^{\mathcal{H}}-1)^2} \sum\nolimits_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i}\|\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 + \frac{2\epsilon^2}{(2\alpha_s^{\mathcal{H}}-1)^2}. \tag{16}$$

Taking conditional expectations in equation 16 and applying the squared form of equation 15 to every honest client yields

$$\begin{aligned}
&\mathbb{E}[\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \mid \mathcal{F}_t] \\
&\le \frac{8\alpha_s^{\mathcal{H}}}{(2\alpha_s^{\mathcal{H}}-1)^2} \sum\nolimits_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i}\mathbb{E}[\|\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \mid \mathcal{F}_t] + \frac{2\epsilon^2}{(2\alpha_s^{\mathcal{H}}-1)^2} \\
&\le \frac{8\alpha_s^{\mathcal{H}}}{(2\alpha_s^{\mathcal{H}}-1)^2} \sum\nolimits_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i}\left(\zeta_s + \frac{2\sigma}{\sqrt{b_{s,t}}}\right)^2 + \frac{2\epsilon^2}{(2\alpha_s^{\mathcal{H}}-1)^2} \\
&= \frac{8(\alpha_s^{\mathcal{H}})^2(\zeta_s + 2\sigma/\sqrt{b_{s,t}})^2 + 2\epsilon^2}{(2\alpha_s^{\mathcal{H}}-1)^2} = (R_{s,t}^{\mathrm{ms}})^2,
\end{aligned}$$

where the equality uses $\sum_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i} = \alpha_s^{\mathcal{H}}$. □

D.3 Sample-Mean Reports and Confidence Dependence

We derive the sample-mean certification radius underlying the comparison in Remark 3.9. In this subsection, each honest client reports the sample mean

$$\tilde{u}_{s,i}^t = \frac{1}{B}\sum\nolimits_{\xi\in\Xi_{s,i}^t} \nabla f_{s,i}(x^t, \xi),$$

and the server aggregates these reports into $\tilde{g}_s^t$ using task-wise TWGM. Conditional independence and unbiasedness in Assumption 3.7 give

$$\mathbb{E}[\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t) \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\|^2 \mid \mathcal{F}_t] \le \frac{\sigma^2}{B}.$$

The difference between the local full gradient and the honest task gradient is fixed conditional on $\mathcal{F}_t$. Expanding the squared local-to-task error, its cross term vanishes by conditional unbiasedness. Thus

$$\begin{aligned}
&\mathbb{E}[\|\tilde{u}_{s,i}^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \mid \mathcal{F}_t] \\
&= \mathbb{E}[\|\tilde{u}_{s,i}^t - \nabla f_{s,i}(x^t)\|^2 \mid \mathcal{F}_t] + \|\nabla f_{s,i}(x^t) - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \le \frac{\sigma^2}{B} + \zeta_s^2.
\end{aligned}$$

The pointwise TWGM bound in equation 16 applies to these reports as well. Taking conditional expectations gives

$$\begin{aligned}
&\mathbb{E}[\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \mid \mathcal{F}_t] \\
&\le \frac{8\alpha_s^{\mathcal{H}}}{(2\alpha_s^{\mathcal{H}}-1)^2} \sum\nolimits_{i\in\mathcal{R}_s\cap\mathcal{H}} \alpha_{s,i}\left(\zeta_s^2 + \frac{\sigma^2}{B}\right) + \frac{2\epsilon^2}{(2\alpha_s^{\mathcal{H}}-1)^2} \\
&= \frac{8(\alpha_s^{\mathcal{H}})^2(\zeta_s^2 + \sigma^2/B) + 2\epsilon^2}{(2\alpha_s^{\mathcal{H}}-1)^2}.
\end{aligned}$$

For sample-mean reports, define the certification radius

$$r_{s,t}^{\mathrm{hp}} := \sqrt{\frac{8(\alpha_s^{\mathcal{H}})^2(\zeta_s^2 + \sigma^2/B) + 2\epsilon^2}{\delta_{s,t}(2\alpha_s^{\mathcal{H}} - 1)^2}}.$$

For $\delta_{s,t} \in (0,1)$ and $r_{s,t}^{\mathrm{hp}} > 0$, Markov's inequality conditional on $\mathcal{F}_t$ gives

$$\begin{aligned}
&\Pr\Big(\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| > r_{s,t}^{\mathrm{hp}} \mid \mathcal{F}_t\Big) \\
&\le \frac{\mathbb{E}[\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \mid \mathcal{F}_t]}{(r_{s,t}^{\mathrm{hp}})^2} \\
&\le \frac{8(\alpha_s^{\mathcal{H}})^2(\zeta_s^2 + \sigma^2/B) + 2\epsilon^2}{(2\alpha_s^{\mathcal{H}} - 1)^2 (r_{s,t}^{\mathrm{hp}})^2} = \delta_{s,t},
\end{aligned}$$

where the equality substitutes the radius defined above. If that radius is zero, the mean-square bound is also zero and the coverage holds almost surely.

For comparison, consider the GMOM choice $K_{s,t} = 1 + \lfloor 3.5 \log(|\mathcal{R}_s|/\delta_{s,t}) \rfloor$ from Section 3.3, for budgets such that $K_{s,t} \le B/2$. This choice satisfies

$$K_{s,t} \le 1 + 3.5 \log \frac{|\mathcal{R}_s|}{\delta_{s,t}}.$$

Since $K_{s,t} \le B/2$, the block size obeys $b_{s,t} = \lfloor B/K_{s,t} \rfloor \ge B/(2K_{s,t})$. Hence

$$\tau_{s,t} = \frac{11\sqrt{5}\,\sigma}{6\sqrt{b_{s,t}}} \le \frac{11\sqrt{10}\,\sigma}{6}\sqrt{\frac{K_{s,t}}{B}} \le \frac{11\sqrt{10}\,\sigma}{6}\sqrt{\frac{1 + 3.5\log(|\mathcal{R}_s|/\delta_{s,t})}{B}}.$$

In the server-side radius, this sampling term is multiplied by $2\alpha_s^{\mathcal{H}}/(2\alpha_s^{\mathcal{H}} - 1)$. With B, $|\mathcal{R}_s|$, and the other problem parameters fixed, the displayed bound grows as $\sqrt{1 + \log(1/\delta_{s,t})}$ over this feasible range of failure budgets, whereas the sample-mean radius derived above scales as $1/\sqrt{\delta_{s,t}}$.

D.4 Probabilistic Descent and Stationarity Certificates

Proposition D.2 (Probabilistic descent and stationarity-gap certificates)**.** *For each task, let $(r_{s,t}^{\mathrm{hp}}, \delta_{s,t})$ be $\mathcal{F}_t$-measurable, with $r_{s,t}^{\mathrm{hp}} \ge 0$ and $\delta_{s,t} \in [0,1]$, and suppose that*

$$\Pr\Big(\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_{s,t}^{\mathrm{hp}} \mid \mathcal{F}_t\Big) \ge 1 - \delta_{s,t}.$$

For any nonzero candidate v^t computed from the current reports, define

$$\mathsf{A}_t(v^t) := \{\langle \tilde{g}_s^t, v^t \rangle > r_{s,t}^{\mathrm{hp}} \|v^t\|,\ \forall s \in [S]\}.$$

Then

$$\Pr\big(\mathsf{A}_t(v^t) \cap \{\exists s : \langle \nabla f_s^{\mathcal{H}}(x^t), v^t \rangle \le 0\} \mid \mathcal{F}_t\big) \le \sum\nolimits_{s\in[S]} \delta_{s,t}. \tag{17}$$

The stationarity-gap upper bound

$$\Psi_{\mathrm{ub,hp}}^t := \min\nolimits_{\lambda \in \Delta_S} \left[\|\widetilde{G}^t \lambda\| + \sum\nolimits_{s\in[S]} \lambda_s r_{s,t}^{\mathrm{hp}}\right]$$

satisfies

$$\Pr\big(\Psi(x^t) > \Psi_{\mathrm{ub,hp}}^t \mid \mathcal{F}_t\big) \le \sum\nolimits_{s\in[S]} \delta_{s,t}. \tag{18}$$

Proof. Fix round t and condition on $\mathcal{F}_t$. Let

$$\mathcal{C}_t := \bigcap_{s\in[S]} \left\{\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_{s,t}^{\mathrm{hp}}\right\}.$$

By the union bound and the task-wise coverage guarantee, the probability that at least one task violates its radius satisfies

$$\Pr\left(\bigcup_{s\in[S]}\left\{\|\tilde{g}_s^t-\nabla f_s^{\mathcal{H}}(x^t)\|>r_{s,t}^{\mathrm{hp}}\right\}\,\middle|\,\mathcal{F}_t\right)$$
$$\le\sum\nolimits_{s\in[S]}\Pr\Big(\|\tilde{g}_s^t-\nabla f_s^{\mathcal{H}}(x^t)\|>r_{s,t}^{\mathrm{hp}}\,\Big|\,\mathcal{F}_t\Big)\le\sum\nolimits_{s\in[S]}\delta_{s,t}.$$

False-acceptance probability. Fix any realized nonzero candidate v^t computed from the current reports. On $\mathcal{C}_t$, for every task $s\in[S]$, Cauchy–Schwarz gives

$$\begin{aligned}\langle\nabla f_s^{\mathcal{H}}(x^t),v^t\rangle&=\langle\tilde{g}_s^t,v^t\rangle+\langle\nabla f_s^{\mathcal{H}}(x^t)-\tilde{g}_s^t,v^t\rangle\\&\ge\langle\tilde{g}_s^t,v^t\rangle-\|\nabla f_s^{\mathcal{H}}(x^t)-\tilde{g}_s^t\|\|v^t\|\\&\ge\langle\tilde{g}_s^t,v^t\rangle-r_{s,t}^{\mathrm{hp}}\|v^t\|.\end{aligned}$$

If $\mathsf{A}_t(v^t)$ occurs, its strict margin condition makes the last quantity positive for every task. Hence

$$\mathsf{A}_t(v^t)\cap\{\exists s:\langle\nabla f_s^{\mathcal{H}}(x^t),v^t\rangle\le 0\}\cap\mathcal{C}_t=\varnothing,$$

so a false acceptance can occur only if at least one task violates its radius. The preceding probability bound therefore gives

$$\Pr\big(\mathsf{A}_t(v^t)\cap\{\exists s:\langle\nabla f_s^{\mathcal{H}}(x^t),v^t\rangle\le 0\}\,\big|\,\mathcal{F}_t\big)\le\sum\nolimits_{s\in[S]}\delta_{s,t}.$$

This implication is pointwise for the realized reports and candidate, so it does not require independence between v^t and the current gradient errors. This proves equation 17.

Stationarity-bound coverage. On $\mathcal{C}_t$, for every $\lambda\in\Delta_S$,

$$G_{\mathcal{H}}^t\lambda=\widetilde{G}^t\lambda+\sum\nolimits_{s\in[S]}\lambda_s\big(\nabla f_s^{\mathcal{H}}(x^t)-\tilde{g}_s^t\big).$$

Because $\lambda_s\ge 0$, the triangle inequality and the definition of $\mathcal{C}_t$ give

$$\|G_{\mathcal{H}}^t\lambda\|\le\|\widetilde{G}^t\lambda\|+\sum\nolimits_{s\in[S]}\lambda_s r_{s,t}^{\mathrm{hp}}.$$

Taking the minimum over $\lambda\in\Delta_S$ on both sides yields

$$\begin{aligned}\Psi(x^t)&=\min\nolimits_{\lambda\in\Delta_S}\|G_{\mathcal{H}}^t\lambda\|\\&\le\min\nolimits_{\lambda\in\Delta_S}\left[\|\widetilde{G}^t\lambda\|+\sum\nolimits_{s\in[S]}\lambda_s r_{s,t}^{\mathrm{hp}}\right]=\Psi_{\mathrm{ub,hp}}^t.\end{aligned}$$

Consequently,

$$\{\Psi(x^t)>\Psi_{\mathrm{ub,hp}}^t\}\cap\mathcal{C}_t=\varnothing,$$

so the stationarity bound can fail only if at least one task violates its radius. The same probability bound gives

$$\Pr\big(\Psi(x^t)>\Psi_{\mathrm{ub,hp}}^t\,\big|\,\mathcal{F}_t\big)\le\sum\nolimits_{s\in[S]}\delta_{s,t}.$$

This proves equation 18. □

CR-MGDA returns the stationarity-gap bound when neither candidate direction is accepted. Its coverage follows from $\mathcal{C}_t$ and holds whether the server then terminates or pauses the update and collects further reports.

D.5 Trajectory-wide Certification

Let $\mathcal{T}_T\subseteq\{0,\dots,T-1\}$ denote the possibly random set of rounds at which the algorithm collects reports. Whether $t\in\mathcal{T}_T$ is determined by $\mathcal{F}_t$, because it depends only on whether the algorithm terminated in an earlier round. Let $\mathcal{C}_t$ be the simultaneous coverage event from the preceding proof on reached rounds, and the entire sample space otherwise. Suppose that the cumulative budget over

the reached rounds is at most δ_{tot} almost surely. A union bound over the reached rounds and the tower property give

$$\begin{aligned}\Pr(\exists t \in \mathcal{T}_T : \mathcal{C}_t^c) &\le \sum_{t=0}^{T-1} \mathbb{E}\big[\mathbf{1}_{\{t\in\mathcal{T}_T\}}\mathbf{1}_{\mathcal{C}_t^c}\big] \\ &= \sum_{t=0}^{T-1} \mathbb{E}\big[\mathbf{1}_{\{t\in\mathcal{T}_T\}}\Pr(\mathcal{C}_t^c \mid \mathcal{F}_t)\big] \\ &\le \mathbb{E}\Big[\sum\nolimits_{t\in\mathcal{T}_T}\sum\nolimits_{s\in[S]}\delta_{s,t}\Big] \le \delta_{\mathrm{tot}}.\end{aligned}$$

The equality uses the $\mathcal{F}_t$-measurability of $\mathbf{1}_{\{t\in\mathcal{T}_T\}}$; the next inequality uses the conditional task-wise coverage guarantees. The predetermined budgets in the main text satisfy the cumulative budget condition since $\mathcal{T}_T \subseteq \{0,\dots,T-1\}$. Consequently, with probability at least $1-\delta_{\mathrm{tot}}$, every accepted direction and every returned stationarity-gap bound is valid. No independence across tasks or rounds is required.

E Proofs and Derivations for Section 4

This section proves the non-convex and strongly convex convergence results in Section 4, with full-gradient and stochastic reports treated separately.

E.1 Proof of Theorem 4.1: Full-Gradient Reports

Theorem 4.1 (Non-convex convergence). *Suppose Assumptions 3.1, 3.2, and 3.3 hold, and let the recovery program in equation 7 use a fixed $\kappa \in (0, 1/2]$. Use a constant step size $0 < \eta \le \kappa/L$. At each round $t < T$, CR-MGDA uses full-gradient reports and executes an update using an accepted proposal or recovery direction, or sets $x^{t+1} = x^t$ if recovery fails. Set $\Delta_0 := \max_{s\in[S]}(f_s^{\mathcal{H}}(x^0) - f_s^{\mathcal{H},\inf})$. Then*

$$\frac{1}{T}\sum\nolimits_{t=0}^{T-1}\Psi(x^t)^2 \le \frac{4\Delta_0}{\eta\kappa T} + \left(\frac{4}{\kappa}+2\right)r^2 + \frac{\gamma}{2\kappa}.$$

Proof. On a round that executes an update, let (λ^t, d^t) denote the accepted proposal or recovery pair.

Proposal updates. Testing the first-order variational inequality of equation 5 against the s-th simplex vertex gives

$$2\langle \hat{g}_s^t, d^t\rangle + \gamma\lambda_s^t \ge 2\|d^t\|^2 + \gamma\|\lambda^t\|^2.$$

Since $\lambda_s^t - \|\lambda^t\|^2 \le \lambda_s^t - (\lambda_s^t)^2 \le 1/4$, it follows that $\langle \hat{g}_s^t, d^t\rangle \ge \|d^t\|^2 - \gamma/8$. Combining this inequality with the full-gradient bound $\|\hat{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\| \le r_s \le r$ gives

$$\begin{aligned}\langle \nabla f_s^{\mathcal{H}}(x^t), d^t\rangle &= \langle \hat{g}_s^t, d^t\rangle + \langle \nabla f_s^{\mathcal{H}}(x^t) - \hat{g}_s^t, d^t\rangle \\ &\ge \|d^t\|^2 - r\|d^t\| - \gamma/8.\end{aligned}$$

By L_s-smoothness and $x^{t+1} = x^t - \eta d^t$,

$$\begin{aligned}f_s^{\mathcal{H}}(x^{t+1}) &\le f_s^{\mathcal{H}}(x^t) + \langle \nabla f_s^{\mathcal{H}}(x^t), x^{t+1} - x^t\rangle + \frac{L_s}{2}\|x^{t+1} - x^t\|^2 \\ &= f_s^{\mathcal{H}}(x^t) - \eta\langle \nabla f_s^{\mathcal{H}}(x^t), d^t\rangle + \frac{L_s\eta^2}{2}\|d^t\|^2 \\ &\le f_s^{\mathcal{H}}(x^t) - \frac{\eta}{2}\|d^t\|^2 + \eta r\|d^t\| + \frac{\eta\gamma}{8} \\ &\le f_s^{\mathcal{H}}(x^t) - \frac{\eta}{4}\|d^t\|^2 + \eta r^2 + \frac{\eta\gamma}{8}.\end{aligned}$$

The third line uses $\eta L_s \le 1$; the last uses $ab \le a^2 + b^2/4$ with $a = r$ and $b = \|d^t\|$. Since $\kappa \le 1/2$, an update using the proposal therefore satisfies

$$\frac{\eta\kappa}{2}\|d^t\|^2 \le f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^{t+1}) + \eta r^2 + \eta\gamma/8. \tag{19}$$

Recovery updates. For a recovery update, $d^t = d_{\mathrm{r}}^t = \widehat{G}^t \lambda_{\mathrm{r}}^t$. Feasibility of $(\lambda_{\mathrm{r}}^t, q_{\mathrm{r}}^t, m_{\mathrm{r}}^t)$ in the recovery program equation 7 gives, for every task s,

$$\|d^t\| \le q_{\mathrm{r}}^t, \qquad \langle \hat{g}_s^t, d^t \rangle - r_s q_{\mathrm{r}}^t \ge m_{\mathrm{r}}^t.$$

The full-gradient aggregation guarantee $\|\nabla f_s^{\mathcal{H}}(x^t) - \hat{g}_s^t\| \le r_s$ and the Cauchy–Schwarz inequality therefore yield

$$\begin{aligned}
\langle \nabla f_s^{\mathcal{H}}(x^t), d^t \rangle &= \langle \hat{g}_s^t, d^t \rangle + \langle \nabla f_s^{\mathcal{H}}(x^t) - \hat{g}_s^t, d^t \rangle \\
&\ge \langle \hat{g}_s^t, d^t \rangle - r_s \|d^t\| \\
&\ge \langle \hat{g}_s^t, d^t \rangle - r_s q_{\mathrm{r}}^t \ge m_{\mathrm{r}}^t > \kappa (q_{\mathrm{r}}^t)^2 \ge \kappa \|d^t\|^2.
\end{aligned}$$

The strict inequality follows directly from the recovery acceptance rule. Since $f_s^{\mathcal{H}}$ is L_s-smooth and $x^{t+1} - x^t = -\eta d^t$, the smoothness inequality gives

$$\begin{aligned}
f_s^{\mathcal{H}}(x^{t+1}) &\le f_s^{\mathcal{H}}(x^t) + \langle \nabla f_s^{\mathcal{H}}(x^t), x^{t+1} - x^t \rangle + \frac{L_s}{2} \|x^{t+1} - x^t\|^2 \\
&= f_s^{\mathcal{H}}(x^t) - \eta \langle \nabla f_s^{\mathcal{H}}(x^t), d^t \rangle + \frac{L_s \eta^2}{2} \|d^t\|^2 \\
&\le f_s^{\mathcal{H}}(x^t) - \eta \left(\kappa - \frac{L_s \eta}{2} \right) \|d^t\|^2 \\
&\le f_s^{\mathcal{H}}(x^t) - \frac{\eta \kappa}{2} \|d^t\|^2.
\end{aligned}$$

The step size condition gives $L_s \eta \le L \eta \le \kappa$, which implies $\kappa - L_s \eta / 2 \ge \kappa / 2$ and proves the last line. Rearranging the last inequality and adding the nonnegative terms $\eta r^2 + \eta \gamma / 8$ to its right-hand side shows that a recovery update also satisfies equation 19.

Stationarity-gap bound and summation. On either type of update round, $d^t = \widehat{G}^t \lambda^t$, so

$$\begin{aligned}
\Psi(x^t) &\le \|G_{\mathcal{H}}^t \lambda^t\| \\
&\le \|d^t\| + \left\| \sum\nolimits_{s \in [S]} \lambda_s^t (\nabla f_s^{\mathcal{H}}(x^t) - \hat{g}_s^t) \right\| \\
&\le \|d^t\| + \sum\nolimits_{s \in [S]} \lambda_s^t r_s \le \|d^t\| + r.
\end{aligned}$$

Hence $\Psi(x^t)^2 \le 2\|d^t\|^2 + 2r^2$. Combining this with equation 19 gives, for every task s,

$$\Psi(x^t)^2 \le \frac{4}{\eta \kappa} \big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^{t+1}) \big) + \left(\frac{4}{\kappa} + 2 \right) r^2 + \frac{\gamma}{2\kappa}. \tag{20}$$

If recovery fails, $x^{t+1} = x^t$ and equation 8 gives $\Psi(x^t)^2 \le [(2-\kappa)/(1-\kappa)]^2 r^2$. For $\kappa \in (0, 1/2]$, direct algebra gives $[(2-\kappa)/(1-\kappa)]^2 \le 4/\kappa + 2$; together with $\gamma \ge 0$, this bounds a failed-recovery round by the residual term in equation 20. Thus equation 20 also holds on these rounds. Summing this bound over all T rounds telescopes the objective differences. Using $f_s^{\mathcal{H}}(x^T) \ge f_s^{\mathcal{H},\inf}$ and the definition of Δ_0 yields

$$\frac{1}{T} \sum\nolimits_{t=0}^{T-1} \Psi(x^t)^2 \le \frac{4\Delta_0}{\eta \kappa T} + \left(\frac{4}{\kappa} + 2 \right) r^2 + \frac{\gamma}{2\kappa}.$$

□

Why the fixed-step bound requires $\kappa > 0$. Consider two exact one-dimensional gradients $g_1 = 1$ and $g_2 = M$, where $M > 1$, with zero error radii. Any convex combination gives a direction $d \in [1, M]$. When $\kappa = 0$, its certified margin is

$$\min\{g_1 d, g_2 d\} = d,$$

so maximum-margin recovery selects $d = M$ and obtains margin $m = M$. However, $m/d^2 = 1/M$, which has no positive lower bound as M grows. A fixed step size can then overshoot. For example, let

$$f_1(x) = \tfrac{1}{2}(x+1)^2, \qquad f_2(x) = \tfrac{1}{2}(x+M)^2.$$

Both objectives are 1-smooth. At $x = 0$, their gradients are 1 and M, and $-d = -M$ is a common-descent direction. Yet the update $x^+ = -\eta M$ increases f_1 whenever $\eta M > 2$. Thus, even with fixed smoothness $L = 1$, any fixed $\eta > 0$ fails for sufficiently large M. A positive margin alone therefore does not provide a single step size that guarantees descent for every accepted recovery direction.

Proposal-only specialization. If every round executes the regularized MGDA proposal, the recovery argument is unnecessary. Under the weaker step size condition $0 < \eta \le 1/L$, the proposal descent estimate above directly gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\Psi(x^t)^2 \le \frac{8\Delta_0}{\eta T} + 10r^2 + \gamma.$$

High-probability bound with stochastic reports. Under the failure-budget condition in Section 3.3, Appendix D.5 ensures simultaneous coverage of all task-gradient errors over the T rounds with probability at least $1-\delta_{\mathrm{tot}}$. On this event, each error is at most $r^{\mathrm{hp}}_{s,t} \le \bar{r}_{\mathrm{hp}}$, where $\bar{r}_{\mathrm{hp}} = \max_{t<T, s\in[S]} r^{\mathrm{hp}}_{s,t}$. Under the step-size condition of Theorem 4.1, the proposal estimate therefore holds with r replaced by $\bar{r}_{\mathrm{hp}}$, while the recovery estimate uses the task-wise radii $r^{\mathrm{hp}}_{s,t}$ in the same way as above. On a round with failed recovery, the argument in Appendix C.3 gives

$$\Psi(x^t) \le \frac{2-\kappa}{1-\kappa}\max_{s\in[S]} r^{\mathrm{hp}}_{s,t} \le \frac{2-\kappa}{1-\kappa}\bar{r}_{\mathrm{hp}}.$$

Thus equation 20 holds on every round with r replaced by $\bar{r}_{\mathrm{hp}}$; summing it proves the high-probability convergence bound stated in Section 4.1.

E.2 Proof of Theorem 4.2: Stochastic Reports

Theorem 4.2 (Expected non-convex convergence). *Suppose Assumptions 3.1, 3.2, 3.3, and 3.7 hold, and the deterministic local-GMOM block counts $K_{s,t}$ satisfy Lemma 3.8. Let CR-MGDA use the same update rule as in Theorem 4.1, with a fixed $\kappa \in (0, 1/2]$, and let $R^2_{\mathrm{ms}} := \max_{0\le t<T}\sum_{s\in[S]}(R^{\mathrm{ms}}_{s,t})^2$. For $0 < \eta \le \kappa/L$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\Psi(x^t)^2] \le \frac{8\Delta_0}{\eta\kappa T} + \left(\frac{8}{\kappa^2} + 2 + \frac{6}{(1-\kappa)^2}\right)R^2_{\mathrm{ms}} + \frac{\gamma}{\kappa}.$$

Proof. The step-size condition $0 < \eta \le \kappa/L$ gives $\eta L_s \le \kappa$ for every task. We first bound each realized update or pause, allowing the task weights and algorithmic decisions to depend on the current reports, and then take expectations.

Executed updates. For any simplex weights λ, convexity of the squared norm gives the pointwise bound

$$\begin{aligned}\left\|\sum_s \lambda_s\big(\tilde{g}^t_s - \nabla f^{\mathcal{H}}_s(x^t)\big)\right\|^2 &\le \sum_s \lambda_s\|\tilde{g}^t_s - \nabla f^{\mathcal{H}}_s(x^t)\|^2 \\ &\le \sum_s \|\tilde{g}^t_s - \nabla f^{\mathcal{H}}_s(x^t)\|^2.\end{aligned}$$

On a round that executes an update, let (λ^t, d^t) be the accepted proposal or recovery pair. For a proposal, the same QP optimality argument as in Appendix E.1 gives $\langle \tilde{g}^t_s, d^t\rangle \ge \|d^t\|^2 - \gamma/8$. For recovery, feasibility of the stochastic recovery program, its acceptance rule, and nonnegativity of $r^{\mathrm{hp}}_{s,t}$ give

$$\langle \tilde{g}^t_s, d^t\rangle \ge m^t_{\mathrm{r}} + r^{\mathrm{hp}}_{s,t}q^t_{\mathrm{r}} > \kappa(q^t_{\mathrm{r}})^2 \ge \kappa\|d^t\|^2.$$

Since $\kappa \le 1$, either executed direction satisfies

$$\langle \nabla f^{\mathcal{H}}_s(x^t), d^t\rangle \ge \kappa\|d^t\|^2 - \|\tilde{g}^t_s - \nabla f^{\mathcal{H}}_s(x^t)\|\|d^t\| - \frac{\gamma}{8}.$$

Smoothness and $\eta L_s \le \kappa$ therefore imply

$$\begin{aligned}f^{\mathcal{H}}_s(x^{t+1}) &\le f^{\mathcal{H}}_s(x^t) - \frac{\eta\kappa}{2}\|d^t\|^2 + \eta\|\tilde{g}^t_s - \nabla f^{\mathcal{H}}_s(x^t)\|\|d^t\| + \frac{\eta\gamma}{8} \\ &\le f^{\mathcal{H}}_s(x^t) - \frac{\eta\kappa}{4}\|d^t\|^2 + \frac{\eta}{\kappa}\|\tilde{g}^t_s - \nabla f^{\mathcal{H}}_s(x^t)\|^2 + \frac{\eta\gamma}{8},\end{aligned}$$

where the last inequality uses $ab \le \kappa a^2/4 + b^2/\kappa$.

On an update round, the comparison with the honest gradient combination gives

$$\Psi(x^t) \le \|G^t_{\mathcal{H}}\lambda^t\| \le \|d^t\| + \left\|\sum_s \lambda^t_s\big(\tilde{g}^t_s - \nabla f^{\mathcal{H}}_s(x^t)\big)\right\|.$$

Consequently, $\Psi(x^t)^2 \le 2\|d^t\|^2 + 2\sum_s \|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2$. Combining this with the preceding descent inequality yields, for every task s,

$$\Psi(x^t)^2 \le \frac{8}{\eta\kappa}\big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^{t+1})\big) + \frac{8}{\kappa^2}\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 + 2\sum\nolimits_j \|\tilde{g}_j^t - \nabla f_j^{\mathcal{H}}(x^t)\|^2 + \frac{\gamma}{\kappa}.$$

Failed recovery. As in Appendix C.3, let $\lambda_0^t \in \arg\min_{\lambda\in\Delta_S} \|\widetilde{G}^t\lambda\|$ and $d_0^t := \widetilde{G}^t\lambda_0^t$. Projection optimality gives $\langle \tilde{g}_s^t, d_0^t\rangle \ge \|d_0^t\|^2$ for every task. If $\|d_0^t\| > \max_{j\in[S]} r_{j,t}^{\mathrm{hp}}/(1-\kappa)$, then

$$\min_s\big[\langle \tilde{g}_s^t, d_0^t\rangle - r_{s,t}^{\mathrm{hp}}\|d_0^t\|\big] - \kappa\|d_0^t\|^2 \ge (1-\kappa)\|d_0^t\|^2 - \max_{j\in[S]} r_{j,t}^{\mathrm{hp}}\|d_0^t\| > 0,$$

contradicting recovery failure. Thus $\|d_0^t\| \le \max_{j\in[S]} r_{j,t}^{\mathrm{hp}}/(1-\kappa)$ when recovery fails. This argument concerns only the estimated gradients and certification radii; it does not require the error balls to contain the honest gradients. Using the actual estimation errors gives

$$\Psi(x^t) \le \|G_{\mathcal{H}}^t\lambda_0^t\| \le \frac{\max_{j\in[S]} r_{j,t}^{\mathrm{hp}}}{1-\kappa} + \Big\|\sum\nolimits_s (\lambda_0^t)_s\big(\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\big)\Big\|,$$

and hence

$$\Psi(x^t)^2 \le \frac{2}{(1-\kappa)^2}\max_{j\in[S]}(r_{j,t}^{\mathrm{hp}})^2 + 2\sum\nolimits_s \|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2.$$

Since $x^{t+1} = x^t$ when recovery fails, both types of rounds satisfy

$$\Psi(x^t)^2 \le \frac{8}{\eta\kappa}\big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^{t+1})\big) + \frac{8}{\kappa^2}\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 + 2\sum\nolimits_j \|\tilde{g}_j^t - \nabla f_j^{\mathcal{H}}(x^t)\|^2 + \frac{\gamma}{\kappa} + \frac{2}{(1-\kappa)^2}\mathbf{1}_{\{\text{recovery fails at } t\}}\max_{j\in[S]}(r_{j,t}^{\mathrm{hp}})^2. \tag{21}$$

Expectation and residual bound. By Lemma 3.8 and the tower property,

$$\mathbb{E}\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 = \mathbb{E}\big[\mathbb{E}[\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \mid \mathcal{F}_t]\big] \le (R_{s,t}^{\mathrm{ms}})^2,$$

so

$$\sum\nolimits_s \mathbb{E}\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \le R_{\mathrm{ms}}^2.$$

Taking expectations in equation 21, summing over all rounds, and using the objective lower bound yields the intermediate estimate

$$\frac{1}{T}\sum\nolimits_{t=0}^{T-1}\mathbb{E}[\Psi(x^t)^2] \le \frac{8\Delta_0}{\eta\kappa T} + \left(\frac{8}{\kappa^2}+2\right)R_{\mathrm{ms}}^2 + \frac{\gamma}{\kappa} + \frac{2}{(1-\kappa)^2 T}\sum\nolimits_{t=0}^{T-1}\mathbb{E}[\mathbf{1}_{\{\text{recovery fails at } t\}}\max_{j\in[S]}(r_{j,t}^{\mathrm{hp}})^2]. \tag{22}$$

The deterministic block counts make the radii deterministic, so the last term equals

$$\frac{2}{(1-\kappa)^2 T}\sum_{t=0}^{T-1}\left[\max_{j\in[S]}(r_{j,t}^{\mathrm{hp}})^2\right]\Pr(\text{recovery fails at } t).$$

It remains to bound this term using the explicit GMOM radii. For fixed s,t, the definition of $\tau_{s,t}$, together with $\zeta_s \ge 0$ and $11\sqrt{5}/12 > 1$, gives

$$2\alpha_s^{\mathcal{H}}(\zeta_s + \tau_{s,t}) \le \frac{11\sqrt{5}}{12}\,2\alpha_s^{\mathcal{H}}\left(\zeta_s + \frac{2\sigma}{\sqrt{b_{s,t}}}\right).$$

Substituting into the GMOM radius and applying Cauchy–Schwarz yields

$$
\begin{aligned}
(r_{s,t}^{\mathrm{hp}})^2 &\le \frac{\left[\frac{11\sqrt{5}}{12}\, 2\alpha_s^{\mathcal{H}}(\zeta_s + 2\sigma/\sqrt{b_{s,t}}) + \epsilon\right]^2}{(2\alpha_s^{\mathcal{H}} - 1)^2} \\
&\le \left[\left(\frac{11\sqrt{5}}{12}\right)^2 + 1\right] \frac{4(\alpha_s^{\mathcal{H}})^2(\zeta_s + 2\sigma/\sqrt{b_{s,t}})^2 + \epsilon^2}{(2\alpha_s^{\mathcal{H}} - 1)^2} \\
&= \frac{749}{288}(R_{s,t}^{\mathrm{ms}})^2 \le 3(R_{s,t}^{\mathrm{ms}})^2.
\end{aligned}
$$

Hence $\max_{j\in[S]}(r_{j,t}^{\mathrm{hp}})^2 \le 3\max_s(R_{s,t}^{\mathrm{ms}})^2 \le 3R_{\mathrm{ms}}^2$. Substituting this into equation 22 proves

$$
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\Psi(x^t)^2] \le \frac{8\Delta_0}{\eta\kappa T} + \left(\frac{8}{\kappa^2} + 2 + \frac{6}{(1-\kappa)^2}\right) R_{\mathrm{ms}}^2 + \frac{\gamma}{\kappa}.
$$

If an update is executed at every round almost surely, then $\Pr(\text{recovery fails at } t) = 0$ for every $t < T$. The last term in equation 22 vanishes, giving the tighter coefficient $8/\kappa^2 + 2$. □

E.3 Proof of Theorem 4.4: Full-Gradient Reports

Theorem 4.4 (Strongly convex convergence). *Suppose Assumptions 3.1, 3.3, and 4.3 hold. Let $0 < \eta < 2/\mu$ and $w_t := (1-\eta\mu/2)^{-(t+1)}$. With full-gradient reports, suppose CR-MGDA executes an update using the regularized MGDA proposal or a recovery direction at every round $t < T$. Let (λ^t, d^t) denote the executed pair, so $d^t = \widehat{G}^t\lambda^t$. Then*

$$
\frac{\sum_{t=0}^{T-1} w_t \sum_{s\in[S]} \lambda_s^t\big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^\star)\big)}{\sum_{t=0}^{T-1} w_t} \le \frac{\mu\|x^0 - x^\star\|^2}{4((1-\eta\mu/2)^{-T} - 1)} + \frac{r^2}{\mu} + \eta(G_{\max}^2 + r^2).
$$

Proof. For every task s, the normalized honest weights in equation 2 and the local-gradient bound in Assumption 4.3 give

$$
\|\nabla f_s^{\mathcal{H}}(x^t)\| \le \sum\nolimits_{i\in\mathcal{R}_s\cap\mathcal{H}} \frac{\alpha_{s,i}}{\alpha_s^{\mathcal{H}}}\|\nabla f_{s,i}(x^t)\| \le G_{\max}.
$$

Since $\lambda^t \in \Delta_S$, convexity of the norm then gives $\|G_{\mathcal{H}}^t\lambda^t\| \le G_{\max}$. Both the proposal and recovery directions satisfy $d^t = \widehat{G}^t\lambda^t$; hence

$$
\begin{aligned}
\|G_{\mathcal{H}}^t\lambda^t - d^t\| &\le \sum\nolimits_{s\in[S]} \lambda_s^t\|\nabla f_s^{\mathcal{H}}(x^t) - \hat{g}_s^t\| \le r, \\
\|d^t\|^2 &\le 2\|G_{\mathcal{H}}^t\lambda^t\|^2 + 2\|d^t - G_{\mathcal{H}}^t\lambda^t\|^2 \le 2G_{\max}^2 + 2r^2.
\end{aligned}
$$

Applying strong convexity to each task with $y = x^\star$, multiplying by λ_s^t, and summing over s gives

$$
\langle x^t - x^\star, G_{\mathcal{H}}^t\lambda^t\rangle \ge \sum\nolimits_{s\in[S]} \lambda_s^t\big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^\star)\big) + \frac{\mu}{2}\|x^t - x^\star\|^2.
$$

Now expand the update $x^{t+1} = x^t - \eta d^t$ and separate the honest direction from its estimation error:

$$
\begin{aligned}
\|x^{t+1} - x^\star\|^2 = \|x^t - x^\star\|^2 &- 2\eta\langle x^t - x^\star, G_{\mathcal{H}}^t\lambda^t\rangle \\
&- 2\eta\langle x^t - x^\star, d^t - G_{\mathcal{H}}^t\lambda^t\rangle + \eta^2\|d^t\|^2.
\end{aligned}
$$

Cauchy–Schwarz, the error bound $\|d^t - G_{\mathcal{H}}^t\lambda^t\| \le r$, and Young's inequality

$$
2\|x^t - x^\star\|r \le \frac{\mu}{2}\|x^t - x^\star\|^2 + \frac{2r^2}{\mu}
$$

give

$$
-2\eta\langle x^t - x^\star, d^t - G_{\mathcal{H}}^t\lambda^t\rangle \le \frac{\eta\mu}{2}\|x^t - x^\star\|^2 + \frac{2\eta r^2}{\mu}.
$$

Combining the preceding three displays and rearranging yields

$$\sum\nolimits_{s\in[S]} \lambda_s^t\big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^\star)\big) \le \frac{1-\eta\mu/2}{2\eta}\|x^t - x^\star\|^2 - \frac{1}{2\eta}\|x^{t+1} - x^\star\|^2 + \frac{r^2}{\mu} + \eta(G_{\max}^2 + r^2).$$

Multiplying by $w_t = (1-\eta\mu/2)^{-(t+1)}$ and summing over t telescopes the distance terms because $w_t(1-\eta\mu/2) = (1-\eta\mu/2)^{-t}$:

$$\sum\nolimits_{t=0}^{T-1} w_t \sum\nolimits_{s\in[S]} \lambda_s^t\big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^\star)\big) \le \frac{\|x^0 - x^\star\|^2 - (1-\eta\mu/2)^{-T}\|x^T - x^\star\|^2}{2\eta} + \left(\frac{r^2}{\mu} + \eta(G_{\max}^2 + r^2)\right)\sum\nolimits_{t=0}^{T-1} w_t.$$

The final squared distance has a nonpositive coefficient, so omitting its contribution gives an upper bound. Since $0 < 1-\eta\mu/2 < 1$, the geometric sum is

$$\sum\nolimits_{t=0}^{T-1} w_t = \sum\nolimits_{t=0}^{T-1}(1-\eta\mu/2)^{-(t+1)} = \frac{2((1-\eta\mu/2)^{-T} - 1)}{\eta\mu}.$$

Dividing by this sum produces exactly the full-gradient inequality in Theorem 4.4. □

E.4 Stochastic Strongly Convex Counterpart

Theorem E.1 (Stochastic strongly convex convergence)**.** *Suppose Assumptions 3.1, 3.3, 3.7, and 4.3 hold. Suppose the deterministic block counts $K_{s,t}$ satisfy the local-GMOM conditions of Lemma 3.8. Let $0 < \eta < 2/\mu$ and $w_t := (1-\eta\mu/2)^{-(t+1)}$. Suppose the algorithm updates using the stochastic proposal or a recovery direction at every round $t < T$, almost surely. Let (λ^t, d^t) denote the executed pair, so $d^t = \widetilde{G}^t\lambda^t$. Then*

$$\frac{\sum_{t=0}^{T-1} w_t \mathbb{E}\Big[\sum_{s\in[S]} \lambda_s^t\big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^\star)\big)\Big]}{\sum_{t=0}^{T-1} w_t} \le \frac{\mu\|x^0 - x^\star\|^2}{4((1-\eta\mu/2)^{-T} - 1)} + \frac{R_{\mathrm{ms}}^2}{\mu} + \eta(G_{\max}^2 + R_{\mathrm{ms}}^2).$$

Proof. As in the full-gradient proof, $\|G_{\mathcal{H}}^t\lambda^t\| \le G_{\max}$. Since $d^t = \widetilde{G}^t\lambda^t$, pointwise convexity of the squared norm gives

$$\|d^t - G_{\mathcal{H}}^t\lambda^t\|^2 \le \sum\nolimits_{s\in[S]} \lambda_s^t\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2.$$

Taking expectations, using $0 \le \lambda_s^t \le 1$ pointwise, and then applying the conditional mean-square bound in Lemma 3.8 yields

$$\begin{aligned}\mathbb{E}\|d^t - G_{\mathcal{H}}^t\lambda^t\|^2 &\le \sum\nolimits_{s\in[S]} \mathbb{E}\big[\lambda_s^t\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2\big] \\ &\le \sum\nolimits_{s\in[S]} \mathbb{E}\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \\ &= \sum\nolimits_{s\in[S]} \mathbb{E}\big[\mathbb{E}[\|\tilde{g}_s^t - \nabla f_s^{\mathcal{H}}(x^t)\|^2 \mid \mathcal{F}_t]\big] \\ &\le \sum\nolimits_{s\in[S]} (R_{s,t}^{\mathrm{ms}})^2 \le R_{\mathrm{ms}}^2.\end{aligned}$$

Consequently,

$$\mathbb{E}\|d^t\|^2 \le 2\mathbb{E}\|G_{\mathcal{H}}^t\lambda^t\|^2 + 2\mathbb{E}\|d^t - G_{\mathcal{H}}^t\lambda^t\|^2 \le 2G_{\max}^2 + 2R_{\mathrm{ms}}^2.$$

The strong-convexity argument from the full-gradient proof is unchanged:

$$\langle x^t - x^\star, G_{\mathcal{H}}^t\lambda^t\rangle \ge \sum\nolimits_{s\in[S]} \lambda_s^t\big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^\star)\big) + \frac{\mu}{2}\|x^t - x^\star\|^2.$$

Expanding $x^{t+1} = x^t - \eta d^t$, separating $G_{\mathcal{H}}^t \lambda^t$ from $d^t - G_{\mathcal{H}}^t \lambda^t$, and applying the same Cauchy–Schwarz and Young inequalities as in the full-gradient proof gives

$$\|x^{t+1} - x^\star\|^2 \le (1 - \eta\mu/2)\|x^t - x^\star\|^2 - 2\eta \sum\nolimits_{s\in[S]} \lambda_s^t \big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^\star)\big) + \frac{2\eta}{\mu}\|d^t - G_{\mathcal{H}}^t \lambda^t\|^2 + \eta^2 \|d^t\|^2.$$

Rearranging and taking expectations yields

$$\mathbb{E}\left[\sum\nolimits_{s\in[S]} \lambda_s^t \big(f_s^{\mathcal{H}}(x^t) - f_s^{\mathcal{H}}(x^\star)\big)\right] \le \frac{1 - \eta\mu/2}{2\eta}\mathbb{E}\|x^t - x^\star\|^2 - \frac{1}{2\eta}\mathbb{E}\|x^{t+1} - x^\star\|^2 + \frac{R_{\mathrm{ms}}^2}{\mu} + \eta(G_{\max}^2 + R_{\mathrm{ms}}^2).$$

Multiplying by $w_t = (1 - \eta\mu/2)^{-(t+1)}$ and summing leaves the distance contribution

$$\frac{\|x^0 - x^\star\|^2 - (1 - \eta\mu/2)^{-T}\mathbb{E}\|x^T - x^\star\|^2}{2\eta}.$$

Omitting the nonpositive contribution of the final squared distance and dividing by the geometric sum,

$$\sum\nolimits_{t=0}^{T-1} w_t = \frac{2((1 - \eta\mu/2)^{-T} - 1)}{\eta\mu},$$

yields the stochastic-report inequality stated after Theorem 4.4. □