ASKING THE CROWD THE RIGHT QUESTION: BIAS-CANCELLING WEIGHTS FOR FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

A federated objective is a weighted sum of client risks, and the weights are almost always fixed in advance. We treat them instead as the only instrument of a wisdom-of-crowds mechanism: clients are noisy views of one truth, each seeing it through an independent distortion that is unbiased across the crowd. That the optimal weights are inversely proportional to the clients' error energies is classical; we begin at the question that answer presupposes, which energies belong there and whether a crowd can recover them from itself. Excess risk on the truth is of the exact order of the aggregate bias energy, so no optimizer can repair a bad weight vector; the truth itself is identifiable only up to a linear tilt, so subtracting estimated client biases provably reproduces uniform weighting. The expected per-client second moments, however, are exactly identified from the law of the crowd's disagreement by a well-conditioned linear inversion, a step random-effects meta-analysis cannot take because a source reports once; their realised counterparts are estimable up to an incoherence floor the algorithm can measure. This yields CROWD, which reads the disagreement off the optimization trajectory at no extra cost and matches a Bayesian minimax lower bound in the same constant: per instance as the horizon grows, and unconditionally as the prior becomes diffuse. For arbitrary distortions it stays competitive with the optimal weights, at a ratio governed by a geometric incoherence the algorithm can measure. On real scans split into sites with their own miscalibrated detectors it attains the oracle excess risk; on a companion federation that pulls bias and noise apart, weighting by noise variance is worse than not weighting at all, and CROWD is not.

1 INTRODUCTION

Every federated system optimises a weighted objective $F_w = \sum_{i=1}^{n} w_i f_i$ over $w \in \Delta_n$, and the weights are chosen once and never revisited. FEDAVG sets $w_i \propto n_i$ (McMahan et al., 2017); agnostic and fair variants replace w by a worst-case or tilted simplex point (Mohri et al., 2019; Li et al., 2020b; 2021); client sampling chooses whom to query but leaves the target fixed (Cho et al., 2022; Chen et al., 2020). The weights encode a preference, and F_w is treated as ground truth.

This paper takes the opposite view. Suppose instead that the clients are n noisy views of one truth: n hospitals measuring the same biology through different instruments, n annotators labelling one concept through different habits. Then there is a single correct objective f_0, no client's risk equals it, and every weight vector produces a different wrong objective. This is the wisdom of crowds (Galton, 1907; Surowiecki, 2004): the crowd beats its members when the errors are individual, independent and not systematically aligned. *If the weights are the only instrument, what can the crowd's own disagreement tell us about how to set them, and what survives when the crowd does not behave as assumed?*

We posit $f_i = f_0 + h_i$ with h_i independent across clients and $\mathbb{E}\nabla h_i(x) = 0$. Writing $H_w = \sum_i w_i h_i$, the weighted objective is $F_w = f_0 + H_w$, and everything is governed by the single vector

$$b(w) := \nabla H_w(x_0^\star) = \sum_{i=1}^{n} w_i \beta_i, \qquad \beta_i := \nabla h_i(x_0^\star), \qquad x_0^\star := \arg\min f_0, \tag{1}$$

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.

through its *bias energy* $\|b(w)\|^2$. Section 2 sandwiches the excess risk on the truth between $c\|b(w)\|^2/L$ and $C\|b(w)\|^2/\mu$. Hence no optimiser can repair a bad w, and no w is punished beyond its bias energy. The linear tilt is also the only crowd model whose members are simultaneously convex and mean-zero.

Contributions. Write $\zeta_i^2 = \mathbb{E}\|\beta_i\|^2$ and AM, HM for the arithmetic and harmonic means of $\zeta_1^2, \dots, \zeta_n^2$. Weights inversely proportional to error energies, with the harmonic mean as value, are classical (Hedges & Olkin, 1985; DerSimonian & Laird, 1986; Markowitz, 1952) and have been applied to federated aggregation (Reyes et al., 2021; Tölle et al., 2024). We ask which energies belong there when the target is unobservable, and whether a crowd can recover them from itself.

(1) *The right energy is the bias energy, and it is identifiable* (Sections 2–4.1). The truth is identifiable only up to a linear tilt; the expected per-client bias energies are identified from the law of the disagreement by a well-conditioned inversion, the realised ones up to a measurable incoherence floor.

(2) *An estimator at no extra cost, with a matching lower bound* (Sections 3–4.3). The residual cancels ∇f_0 at any iterate, which gives CROWD. Its excess risk matches the bias part $\Psi_0/(2\mu)$ of a Bayesian minimax lower bound for a Gaussian crowd in the same constant, to a factor $1 + o(1)$. The match holds per instance as the horizon grows, and unconditionally in the diffuse-prior limit (Theorem C.11).

(3) *A guarantee beyond the model* (Section 4.4). For arbitrary realised tilts CROWD is $\frac{1+\kappa}{1-\kappa}$-competitive with the optimal weights up to estimation error, κ being a geometric incoherence.

1.1 RELATED WORK

Heterogeneity as a rate obstacle. Federated optimisation mostly treats client heterogeneity as an obstacle to *optimising* a given weighted objective (Kairouz et al., 2021): local-update drift (McMahan et al., 2017; Stich, 2019; Koloskova et al., 2020; Woodworth et al., 2020), proximal terms (Li et al., 2020a), control variates (Karimireddy et al., 2020), objective-inconsistency corrections (Wang et al., 2020), server adaptivity (Reddi et al., 2021). There the target is fixed and heterogeneity inflates the rate through a gradient-dissimilarity constant; under linear tilts that constant instead shifts the limit, and weighted minibatching converges at the homogeneous rate.

Weights as a preference, and as a quality proxy. The rules named above encode a normative goal and treat the resulting objective as ground truth (Mohri et al., 2019; Li et al., 2020b; 2021; Sagawa et al., 2020). Reading the weights as an estimator is orthogonal to these: it replaces the target rather than the rule. A second family weights by a surrogate for update quality: inverse gradient or parameter variance (Reyes et al., 2021; Tölle et al., 2024), a generalisation-bound surrogate (Xu et al., 2023), validation loss, or consistency with an estimated consensus (Chen et al., 2026). These reach the closed form of Proposition D.5 by the meta-analytic route. We add three things: the quantity to invert is a bias energy rather than a noise variance, it has an estimator with a guarantee, and the consensus itself is not identifiable.

Meta-analysis, pooling and fusion. Inverse-variance weighting is the canonical combination of unbiased estimators of one quantity (Hedges & Olkin, 1985), and random-effects meta-analysis (DerSimonian & Laird, 1986) adds a between-study variance estimated by a moment identity parallel to Theorem 4.1. It stops at a single shared variance because a study reports once; a federated round is a repeated draw from the same client, which is what makes the per-client energies estimable. The optimal-weight formula is the minimum-variance portfolio of Markowitz (1952), and opinion pooling names when averaging works (Genest & Zidek, 1986; Page, 2007). Forecasting reports the opposite difficulty, estimated optimal weights losing to the simple average because estimating them adds variance (Stock & Watson, 2004; Smith & Wallis, 2009; Claeskens et al., 2016; DeMiguel et al., 2009); on the simplex the squared weights sum to at most one, so that price stays additive and uniform in n.

Learning whom to learn from. Collaborative learning weights peers' gradients to trade bias against variance (Chayti et al., 2022; Grimberg et al., 2021; Even et al., 2022). There the pairwise discrepancies are supplied to the algorithm, which those authors name as the main limitation; here

they are the object to be estimated. Those methods also personalise, so each agent scores itself on held-out data; Proposition 3.1 arises because nobody observes the common objective. Alamouti & Arjomandi (2025) and Kiselman & Levy (2026) also reach inverse-cost weights: the first with biases predictable from covariates, the second by minimising a bound for one designated client. In both the target is observable, so neither an identification question nor a matching lower bound arises. Appendix B carries the full comparison, with annotator-reliability models (Dawid & Skene, 1979; Karger et al., 2011), two-way fixed-effects differencing (Abowd et al., 1999) and cross-fitting (Chernozhukov et al., 2018).

2 Problem statement

We work on a closed convex $\mathcal{X} \subseteq \mathbb{R}^d$ with Euclidean projection $\Pi_{\mathcal{X}}$, write Δ_n for the simplex on $[n]$ and $\mathbf{1}$ for the all-ones vector. Every numbered statement of the paper is proved in its own subsection of Appendices D–F; the claims quoted informally in this section are stated in full in Appendix D.

Assumption A1 (Crowd)**.** $f_i = f_0 + h_i$, where $h_1, \dots, h_n$ are independent random C^2 functions with $\mathbb{E}\nabla h_i(x) = 0$ for every x, drawn once and then fixed. We write $\beta_i := \nabla h_i(x_0^\star)$ for client i's realised tilt at the truth's optimum, $q_i := \|\beta_i\|^2$ for its realised bias energy, and $\zeta_i^2 := \mathbb{E}q_i$ for the expected one. We also write $H_w = \sum_i w_i h_i$ and $\lambda_w := \sup_{\mathcal{X}} \|\nabla^2 H_w\|_{\mathrm{op}}$, bounded by $\ell := \max_i \ell_i$ when ℓ_i bounds $\|\nabla^2 h_i\|_{\mathrm{op}}$. The *linear-tilt* case is $h_i = \langle \beta_i, \cdot \rangle$, i.e. $\ell = 0$.

Two classical wisdom-of-crowds hypotheses do the work. Independence says clients err for unrelated reasons; it fails under a shared preprocessing pipeline, and Section 4.4 shows what survives. Mean-zero says the crowd is centred on the truth rather than on a shared distortion; it is untestable, and assumed whenever an aggregate is called closer to ground truth than its parts.

Assumption A2 (Truth)**.** f_0 is μ-strongly convex and L-smooth on $\mathcal{X}$, $0 < \mu \le L$, with $f_0^\star = \min_{\mathcal{X}} f_0$ attained at $x_0^\star$.

Assumption A3 (Oracle)**.** At round k the server broadcasts x^{k-1} and each client returns $\widehat{g}_i^k = \nabla f_i(x^{k-1}) + \varepsilon_i^k$ computed from b fresh samples, with $\mathbb{E}[\varepsilon_i^k \mid \mathcal{F}_{k-1}] = 0$, $\mathbb{E}[\|\varepsilon_i^k\|^2 \mid \mathcal{F}_{k-1}] \le \sigma_i^2/b$ and conditional independence across i. Query points may be adaptive. We write $\tau_i^2 = \sigma_i^2 + \frac{1}{n^2}\sum_j \sigma_j^2$, which bounds the noise level of client i's report after the crowd mean has been subtracted.

Assumption A2 applies to the truth only. The perturbations h_i need not be convex, and in a non-linear crowd they cannot all be: a crowd of almost surely convex mean-zero perturbations is almost surely affine (Proposition D.1). The way past the linear model is therefore perturbative. Assumption A3 bounds a conditional rather than a uniform second moment and lets the query point depend on everything seen so far, which matters because our statistic is computed at the algorithm's own iterates.

Since $F_w = f_0 + H_w$, choosing w and optimising F_w are the only two things one can do. The bias-transfer lemma (Lemma D.2, Appendix D) says they combine additively through the aggregate bias $\Gamma_w := \|b(w)\|$. For a linear crowd, every x and every $\epsilon > 0$ satisfy $f_0(x) - f_0^\star \le (1+\epsilon)\,\Gamma_w^2/(2\mu) + (1+\epsilon^{-1})(L/\mu)[F_w(x) - \min F_w]$, while the minimiser of F_w itself pays at least $\Gamma_w^2/(2L)$. For a non-linear crowd the same holds with $\mu - \lambda_w$ in place of μ and cruder constants, and $x_w^\star$ lies within $\Gamma_w/(\mu - \lambda_w)$ of $x_0^\star$. So any method converging to $\arg\min F_w$, even exactly with infinite data, pays $\Theta(\Gamma_w^2)$ on the truth. The $1+\epsilon$ refinement is what lets the two bounds of Section 4 meet in one constant.

2.1 The oracle target: a harmonic law of crowds

Suppose an oracle reveals the bias energies. For weights independent of the tilts, independence and mean-zero kill the cross terms, so $\mathbb{E}\Gamma_w^2 = \sum_i w_i^2 \zeta_i^2$. Minimising this quadratic over the simplex gives closed-form weights $w_i^\star \propto \zeta_i^{-2}$ with value $\Psi_0 := \mathrm{HM}/n$, against AM/n for uniform weighting (Proposition D.5, Appendix D). The ratio $\mathrm{AM}/\mathrm{HM} \ge 1$ is one only for equal energies and is unbounded. A correctly weighted crowd is never worse than its best member and behaves as a crowd of $n \cdot \mathrm{AM}/\mathrm{HM}$: one site with $\zeta_1^2 = 100$ among 20 unit-energy ones makes the uniform crowd behave like fewer than four. The simplex constraint is inactive, so signed aggregation helps only when biases are correlated, which is where the weights cease to be identifiable.

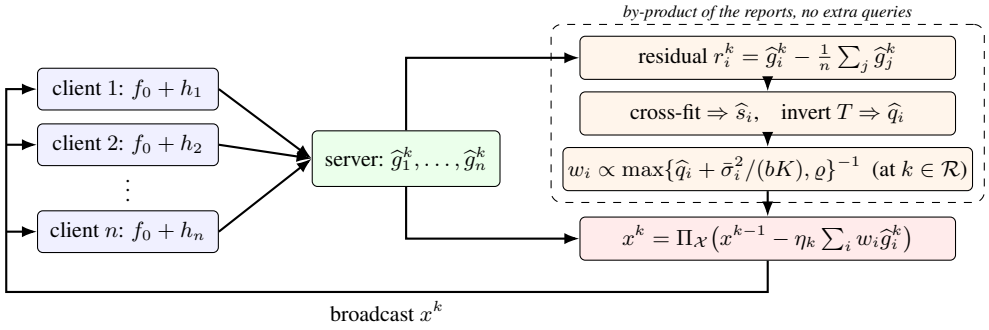


Figure 1: One round of CROWD. The unknown $\nabla f_0(x^{k-1})$ cancels in the residual, so the disagreement is measured at any query point at no extra cost.

Since the aggregate gradient variance is $\sum_i w_i^2 \sigma_i^2 / b$, the same proposition with $c_i = \zeta_i^2 + \sigma_i^2/K$ gives the finite-sample target

$$\Psi(K) := \Big(\sum_i \big(\zeta_i^2 + \sigma_i^2/K\big)^{-1} \Big)^{-1}, \qquad w_i^\star(K) \propto \big(\zeta_i^2 + \sigma_i^2/K\big)^{-1}, \tag{2}$$

where K is the number of samples per client. A client is discounted for bias and for noise on one scale, at one variance unit per K samples. On a Gaussian crowd $\Psi(K)/(2\mu)$ is a worst-case lower bound for every algorithm, and CROWD approaches it per instance as the horizon grows (Section 4.3).

3 ALGORITHM DESIGN

Choosing $w_i \propto \zeta_i^{-2}$ requires knowing the ζ_i^2; this section asks what a crowd can learn about itself and turns the answer into an algorithm.

Proposition 3.1 (Two impossibilities)**.** *Under Assumptions A1–A3: (i) for every $c \in \mathbb{R}^d$, replacing f_0 by $f_0 + \langle c, \cdot \rangle$ and each h_i by $h_i - \langle c, \cdot \rangle$ leaves every client risk, and hence the law of the transcript of any protocol, unchanged, while for $c \neq 0$ and $\mathcal{X} = \mathbb{R}^d$ the target moves; (ii) subtracting from each client's gradients its deviation from the crowd mean makes the aggregate an unbiased gradient of $F_{w_{\mathrm{unif}}}$ for every $w \in \Delta_n$, with expected bias energy AM/n.*

No amount of data identifies the truth, so a claim that an aggregate is closer to ground truth is a claim about the crowd. By (ii) correcting clients does not help – the disagreement reveals deviations from the consensus, not the consensus error – but weighting does, because unequal ζ_i make the dispersion informative.

The residual statistic. Part (i) also tells us which statistic to use. Define the residual

$$r_i^k := \widehat{g}_i^k - \frac{1}{n}\sum_j \widehat{g}_j^k \qquad \text{and write } \ \widetilde{\beta}_i := \beta_i - \tfrac{1}{n}\textstyle\sum_j \beta_j \tag{3}$$

for the corresponding disagreement in the tilts. The residual cancels $\nabla f_0(x^{k-1})$ identically, at any query point and for any f_0. It is a maximal invariant of the per-round shift group (Lemma E.1), so by Proposition 3.1(i) it loses nothing a shift-invariant procedure could have used. For linear tilts $r_i^k = \widetilde{\beta}_i + \widetilde{\varepsilon}_i^k$ exactly.

Two obstacles separate this signal from the weights. First, noise inflates it: the plug-in estimate $\|\widehat{\beta}_i\|^2$ over K_0 rounds carries an offset of at most $\tau_i^2/(bK_0)$, and removing it would need the unknown τ_i. We instead split the probing rounds into two interleaved halves by parity, average within each, and take the inner product of the two averages:

$$\widehat{\beta}_i^{(a)} = \frac{1}{|\mathcal{K}_a|}\sum_{k \in \mathcal{K}_a} r_i^k, \qquad \widehat{s}_i = \big\langle \widehat{\beta}_i^{(1)}, \widehat{\beta}_i^{(2)} \big\rangle. \tag{4}$$

The halves use disjoint rounds with uncorrelated noise, so the inner product is exactly unbiased for $\|\widetilde{\beta}_i\|^2$ with no knowledge of the noise level (Chernozhukov et al., 2018). The factor-two variance

Algorithm 1 CROWD

Require: x^0; stepsizes η_k; batch b; noise bounds $\bar{\sigma}_i \geq \sigma_i$; refresh rounds $\mathcal{R} \subseteq \{2, 3, \dots\}$

$w^1 \leftarrow \mathbf{1}/n$; $S^{(1)}, S^{(2)} \leftarrow 0$; $m_1, m_2 \leftarrow 0$
for $k = 1, 2, \dots$ **do**
  Clients, in parallel
  $\widehat{g}_i^k \leftarrow \nabla f_i(x^{k-1}) + \varepsilon_i^k$ ▷ b fresh samples
  $x^k \leftarrow \Pi_{\mathcal{X}}\big(x^{k-1} - \eta_k \sum_i w_i^k \widehat{g}_i^k\big)$
  Server: statistics from the same reports
  $r_i^k \leftarrow \widehat{g}_i^k - \frac{1}{n}\sum_j \widehat{g}_j^k$ ▷ ∇f_0 cancels
  $a \leftarrow 2 - (k \bmod 2)$ ▷ odd or even half
  $S^{(a)} \leftarrow S^{(a)} + r^k$; $m_a \leftarrow m_a + 1$
  if $k \in \mathcal{R}$ **then**
    $w^{k+1} \leftarrow \textsc{Weights}(S^{(1)}/m_1,\ S^{(2)}/m_2)$
  else $w^{k+1} \leftarrow w^k$

Algorithm 2 $\textsc{Weights}(\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)})$

Require: half-averages $\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}$; rounds K the weights will serve; b and $\bar{\sigma}_i$ as in Algorithm 1; floor $\varrho > 0$

$\widehat{s}_i \leftarrow \big\langle \widehat{\beta}_i^{(1)}, \widehat{\beta}_i^{(2)} \big\rangle$ ▷ by (4)
$\widehat{q}_i \leftarrow \big[\frac{n}{n-2}\big(\widehat{s}_i - \frac{\sum_j \widehat{s}_j}{n(n-1)}\big)\big]_+$ ▷ invert T, clip at zero
$\widehat{c}_i \leftarrow \max\big\{\widehat{q}_i + \bar{\sigma}_i^2/(bK),\ \varrho\big\}$ ▷ by (2); floor
return $w_i = \widehat{c}_i^{-1} / \sum_j \widehat{c}_j^{-1}$

cost is acceptable: an offset would distort most the energies of the low-bias clients, which the weights should favour. Interleaving keeps both halves on the same stretch of trajectory.

Second, subtracting the crowd mean removes part of each client's own tilt and mixes in part of everyone else's, so $\widehat{s}_i$ estimates $(Tq)_i$ rather than q_i, for the map $(Tq)_i = (1 - \frac{2}{n})q_i + \frac{1}{n^2}\sum_j q_j$. The inversion of T is therefore part of the estimator, and Section 4.1 shows it is well conditioned. Algorithm 2 inverts T and weights by $\widehat{q}_i + \bar{\sigma}_i^2/(bK)$ as in (2), K being the remaining horizon if known and the rounds seen so far otherwise. The floor $\varrho > 0$ is not cosmetic. The clipped inversion can return $\widehat{q}_i = 0$ exactly, and on an admissible noiseless instance ($\bar{\sigma}_i = 0$, possible with $\kappa < 1$) the unfloored cost would vanish and the normalisation would divide by zero. Flooring costs at most ϱ in the ℓ_∞ error that Lemma 4.3 charges, so any ϱ far below the energy scale (e.g. 10^{-12} times the largest unfloored cost) is safe.

Algorithm 1 (Figure 1) costs $O(nd)$ per round beyond the optimisation step and calls Algorithm 2 only at refresh rounds; since w^{k+1} depends on rounds $\leq k$, the anytime analysis of Theorem C.5 applies.

4 MAIN THEORETICAL RESULTS

Section 4.1 shows what the disagreement determines; Section 4.2 turns that identity into an estimator and the end-to-end excess risk of CROWD; Section 4.3 matches it from below; Section 4.4 drops the crowd model and keeps a competitive guarantee. Appendix C states the extensions: the guard, correlated crowds, the anytime schedule, the non-linear bound, partial participation.

4.1 WHAT THE DISAGREEMENT DETERMINES

The tilts themselves are out of reach, by Proposition 3.1; their second moments are not. Centring mixes the energies across clients in a way that depends only on the crowd size, and the mixing can be undone.

Theorem 4.1 (Second-moment identifiability)**.** *Let Assumption A1 hold with h_i linear, and $n \geq 3$. With T as in Section 3, the expected disagreement energies $s_i = \mathbb{E}\|\widetilde{\beta}_i\|^2$ satisfy $s = T\zeta^2$, and T is invertible with*

$$(T^{-1}s)_i = \frac{n}{n-2}\Big(s_i - \frac{1}{n(n-1)}\sum_j s_j\Big), \qquad \|T^{-1}\|_{\infty\to\infty} = \frac{n^2-2}{(n-1)(n-2)} =: C_n.$$

Hence ζ^2, and with it the oracle weights, are identified by the law of the disagreement residuals. Moreover $C_3 = 3.5$, $C_{20} \leq 1.17$, $C_n \to 1$, and T is singular for $n = 2$.

The constant C_n is the conditioning of the inversion: an ℓ_∞ error in s is amplified by at most C_n; independence is equally necessary (Proposition C.4). Three claims of different strength must be kept

apart. *Population identifiability* (Theorem 4.1): the law of the disagreement determines the expected energies ζ_i^2 exactly. *Realised estimation* (Section 4.2): the realised $q_i = \|\beta_i\|^2$ are estimable only up to an incoherence floor of order $\kappa\bar{q}$, which no number of rounds removes. *Exact recovery* of the q_i or of the tilts is neither claimed nor possible (Proposition 3.1(i); the floor survives $K_0 \to \infty$). The weights need only the second, crudest claim, through the 2-Lipschitz map of Lemma 4.3.

4.2 Estimation error, and why a crude estimate suffices

Theorem 4.1 concerns expectations, while an algorithm sees one realisation. For it the map from bias to disagreement energies is only approximately T, and one geometric quantity controls the gap.

Assumption A4 (Incoherence)**.** With Q the Gram matrix of the realised tilts and D its diagonal, $\kappa := \|D^{-1/2}(Q-D)D^{-1/2}\|_{\mathrm{op}} < 1$.

Assumption A4 is a property of one realisation, on which the assumption-free guarantee of Section 4.4 also rests. It gives two deterministic facts:

$$\|\widetilde{\beta}_i\|^2 = (Tq)_i + \eta_i \;\text{ with }\; |\eta_i| \le 3\kappa\bar{q}, \qquad (1-\kappa)\,V(w) \;\le\; \Gamma_w^2 \;\le\; (1+\kappa)\,V(w), \tag{5}$$

where $V(w) := \sum_i w_i^2 q_i$ is the diagonal proxy for the bias energy (Lemmas F.1 and D.3). For Gaussian tilts with $d \gtrsim n$, Lemma D.4 gives $\kappa \le C(\sqrt{n/d} + \sqrt{\log(n/\delta)/d})$. The operative condition, however, compares the crowd size with the number of *independent causes* of heterogeneity. That number is at most c^2 for c classes differing through a label channel, and the number of detectors when the instruments differ. For a near-linear crowd κ is estimable from the two half-averages the algorithm already holds (Proposition C.1).

The next lemma carries a *staleness* parameter δ_i: zero for a linear crowd, $\widetilde{\ell}_i \bar{D}$ for a non-linear one, where the residual is measured at a moving point.

Lemma 4.2 (Estimation error)**.** *Let Assumption A3 hold with arbitrary tilts, $n \ge 3$ and K_0 even, and let the disagreement at the query points lie within δ_i of $\widetilde{\beta}_i$ for all $k \le K_0$. Put $\vartheta_i := \tau_i\sqrt{2/(bK_0)}$ and $\widetilde{\zeta}_i := \|\widetilde{\beta}_i\|$.* (i) *if $\delta_i = 0$, $\widehat{s}_i$ is unbiased for $\|\widetilde{\beta}_i\|^2$ for every adaptive query sequence, with a variance of order $\widetilde{\zeta}_i^2\vartheta_i^2 + \vartheta_i^4$ that does not depend on d;* (ii) $\mathbb{E}|\widehat{s}_i - \|\widetilde{\beta}_i\|^2| \le 2\widetilde{\zeta}_i(\delta_i+\vartheta_i) + (\delta_i+\vartheta_i)^2$*;* (iii) *under Assumption A4, $\widehat{q} = (T^{-1}\widehat{s})_+$ satisfies $\Delta := \max_i |\widehat{q}_i - q_i| \le C_n(\max_i |\widehat{s}_i - \|\widetilde{\beta}_i\|^2| + 3\kappa\bar{q})$.*

Lemma F.3 gives the constants and a high-probability form. The three error sources differ in character: ϑ_i decays as $(bK_0)^{-1/2}$, the floor $\kappa\bar{q}$ decays with d and not with K_0, and non-linearity enters through δ_i alone. A crude estimate suffices, the map from cost estimates to attained cost being 2-Lipschitz in ℓ_∞ uniformly in n.

Lemma 4.3 (Weight stability)**.** *Let $c, \widehat{c} \in \mathbb{R}^n_{>0}$ with $\|\widehat{c} - c\|_\infty \le \Delta$ and let $\widehat{w}_i \propto \widehat{c}_i^{-1}$. Then $\sum_i \widehat{w}_i^2 c_i \le \min_{w\in\Delta_n} \sum_i w_i^2 c_i + 2\Delta$. In particular, with $c_i = q_i + \bar{\sigma}_i^2/(bT)$ and the floored costs $\widehat{c}_i = \max\{(\widehat{q}_i)_+ + \bar{\sigma}_i^2/(bT), \varrho\}$ of Algorithm 2, $V(\widehat{w}) \le \widehat{\Psi}(bT) + 2(\Delta+\varrho)$, where $\widehat{\Psi}(K) = (\sum_i (q_i + \bar{\sigma}_i^2/K)^{-1})^{-1}$ is the realised counterpart of (2): clipping at zero never hurts, and the floor adds at most ϱ to the ℓ_∞ error.*

Lemma 4.3 is deterministic and assumes nothing about how $\widehat{c}$ was obtained. The price of estimation enters additively and uniformly in n, unlike in the forecast-combination puzzle (Section 1.1). Composing with Lemma 4.2 gives the end-to-end bound.

Theorem 4.4 (End-to-end excess risk, linear crowd)**.** *Let Assumptions A1–A4 hold with h_i linear, $\mathcal{X} = \mathbb{R}^d$, $n \ge 3$. Run Algorithm 1 with $\mathcal{R} = \{K_0\}$, then T rounds of* SGD *on $F_{\widehat{w}}$ with the tuned stepsize of Lemma F.6, writing $V_\sigma := \sum_i \widehat{w}_i^2\bar{\sigma}_i^2/b$, R_0 an $\mathcal{F}_{K_0}$-measurable bound on $\|x^{K_0} - x^\star_{\widehat{w}}\|$ used in the stepsize (it cannot depend on the unknown minimiser) and $\ell_T := 1 + \ln(e + \mu^2 R_0^2 T/V_\sigma)$. Then, conditionally on the first K_0 rounds and with Δ the realised error of Lemma 4.2(iii) enlarged by the floor ϱ of Algorithm 2, for every $\epsilon > 0$,*

$$\mathbb{E}\big[f_0(x^{K_0+T}) - f_0^\star\big] \;\le\; \frac{(1+\epsilon^{-1})L^2}{2\mu}\Big(R_0^2 e^{-\mu T/(2L)} + \frac{\ell_T V_\sigma}{\mu^2 T}\Big) \;+\; \frac{(1+\epsilon)(1+\kappa)}{2\mu}\big(\widehat{\Psi}(bT) + 2\Delta\big).$$

The first bracket vanishes with the horizon; the second is what the weights cost and survives $T \to \infty$. A small ϵ brings the bias constant arbitrarily close to $1/(2\mu)$ at the price of a larger multiple of the

vanishing terms. Any $\epsilon_T \to 0$ with $\epsilon_T^{-1} = o(T/\log T)$ works. A single refresh at K_0 freezes the sampling part of Δ; refreshing with all rounds seen so far keeps shrinking it. Theorem C.5 covers the geometric grid $\mathcal{R} = \{\lceil \gamma^j \rceil\}$, with $\lceil \log_\gamma T \rceil$ refreshes in T rounds, at a factor $(1-\gamma^{-1})^{-1}$ on the vanishing terms. We use $\gamma = 2$. Refreshing every round behaves the same, and a single early refresh can stall well above the oracle floor (Figure 3).

4.3 A MATCHING LOWER BOUND

CROWD attains the harmonic value; this section bounds what any algorithm can attain. By Proposition 3.1 a weighting rule can only use the disagreement, and the oracle weights exhaust that channel (Proposition F.7). The next theorem drops even the restriction to weighting rules.

Theorem 4.5 (Bayesian minimax lower bound)**.** *Let μ, d, n, the energies ζ_i, the noise levels σ_i and a per-client query budget K all be given, and consider the following instances: the truth is the quadratic $f_0 = \frac{\mu}{2}\|x - x_c\|^2$ with unknown centre x_c, the clients are $f_i = f_0 + \langle \beta_i, \cdot \rangle$ with $\beta_i \sim \mathcal{N}(0, \frac{\zeta_i^2}{d} I_d)$, and each query to client i returns its gradient corrupted by $\mathcal{N}(0, \frac{\sigma_i^2}{d} I_d)$ noise. Then over all algorithms making at most K queries per client – adaptive query points, arbitrary computation on the reports, full knowledge of μ, the ζ_i and the σ_i –*

$$\inf_{A} \sup_{x_c \in \mathbb{R}^d} \mathbb{E}\big[f_0(\widehat{x}) - \min f_0\big] \;\geq\; \frac{1}{2\mu}\Psi(K).$$

An algorithm additionally handed the exact disagreements still pays at least $(\Psi_0 + s^2)/(2\mu)$, where $s^2 = (\sum_i K/\sigma_i^2)^{-1}$; the two floors coincide when the bias and noise profiles are proportional.

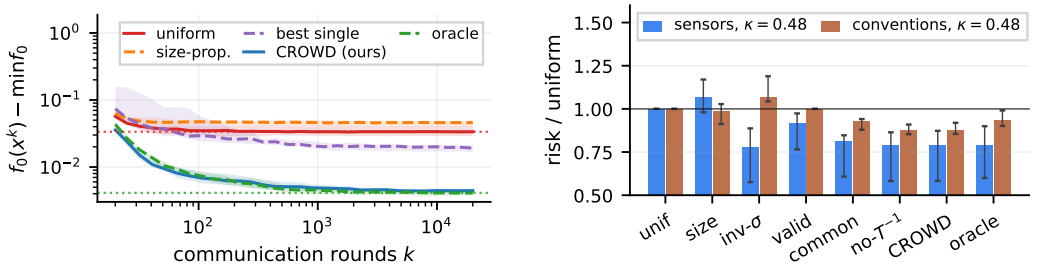
Three features matter. Firstly, the hard instance is perfectly conditioned, so the difficulty is identification rather than optimisation. Secondly, adaptivity does not help, since μ is known and a report carries the same information at every query point. Finally, the upper bound of CROWD has the same form $\Psi(bT)$ up to an estimation term.

Corollary 4.6 (Per-instance sharpness against the minimax floor)**.** *Run Algorithm 1 as in Theorem 4.4 at a fixed centre and let $T \to \infty$. On every realisation with $2\Delta \leq \epsilon_1 \widehat{\Psi}(bT)$ and $\kappa \leq \epsilon_2$, the excess risk is at most $(1 + o(1))(1+\epsilon_1)(1+\epsilon_2)\widehat{\Psi}(bT)/(2\mu)$, the $o(1)$ vanishing more slowly than the horizon term of Theorem 4.4. The lower bound of Theorem 4.5 for the same per-client budget is $\Psi(b(K_0+T))/(2\mu)$. Both values tend to their bias parts, $\widehat{\Psi}_0$ and Ψ_0, where $\widehat{\Psi}_0 := (\sum_i q_i^{-1})^{-1}$: at every fixed centre* CROWD *attains, up to the factor $(1+\epsilon_1)(1+\epsilon_2)\widehat{\Psi}_0/\Psi_0$, the floor below which no algorithm can go uniformly over centres.*

The two bounds share the functional and the constant $1/(2\mu)$, but they are not the same quantity, and the corollary is not a claim of uniform minimax optimality. The upper bound is conditional: it fixes the centre, requires its two hypotheses and uses the realised benchmark $\widehat{\Psi}$. The lower bound is a supremum over centres and an average over the crowd prior. Three bridges narrow the gap. Over centres, the transient of Theorem 4.4 depends on the centre only through R_0, so the upper bound is uniform over any fixed ball once T is large. Over the prior, H is concave (Lemma D.7), so $\mathbb{E}\widehat{\Psi}_0 \leq \Psi_0$ by Jensen. Lemma D.4 puts $\widehat{\Psi}_0/\Psi_0$ within ϵ_0 of one and κ below ϵ_2 once $d \gtrsim \epsilon_2^{-2}(n + \log(n/\delta))$. Over the hypotheses, $2\Delta \leq \epsilon_1 \widehat{\Psi}$ asks for $K_0 \gtrsim \epsilon_1^{-2}$ probing rounds. Theorem C.11 assembles the three into an unconditional statement. The Bayes risk of a prior-aware variant of CROWD matches the Bayes floor (7) as the prior becomes diffuse (Corollary C.12).

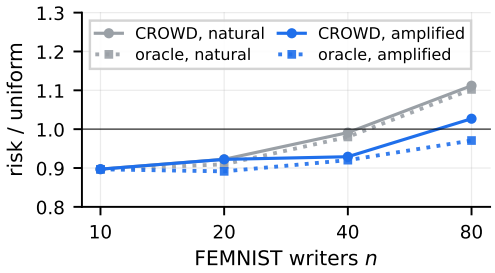
4.4 BEYOND THE LINEAR MODEL

A non-linear crowd necessarily perturbs curvature (Proposition D.1). By Lemma D.2(i) the cost of that perturbation is governed by λ_w/μ. The residual (3) still cancels ∇f_0 for arbitrary h_i, but it now measures the disagreement at the current iterate. Therefore the estimator is stale by $\delta_i \leq \widetilde{\ell}_i \|x^{k-1} - x_0^\star\|$, which does not vanish as probing converges to the uniform minimiser. The analysis goes through with $\mu - \lambda$ in place of μ and with the crude transfer constant; the sharp one is available only for a linear crowd (Theorem C.7). To keep the non-linear part of the estimation error below $\epsilon\Psi_0$, a well-started run needs $\ell/\mu \lesssim \epsilon n^{-1/2}\mathrm{HM}/\mathrm{AM}$ (Remark C.8). The appendix runs show no degradation even where this condition fails. The final result drops the crowd model entirely.

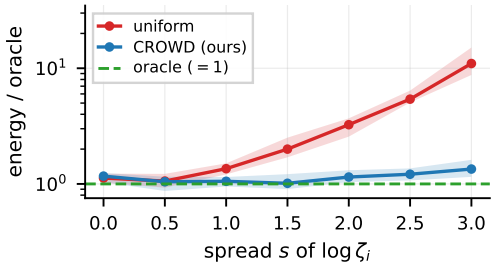


(a) Synthetic: bias floors

(b) Real federations



(c) FEMNIST writers



(d) Gain against the spread of ζ_i

Figure 2: (a) Excess risk on the truth, 12 seeds; dotted lines are the exact floors $\frac{1}{2}b(w)^\top A^{-1}b(w)$. (b, c) Excess risk relative to uniform, 8 and 12 seeds. (d) Bias energy relative to the oracle, 24 seeds. Bands and bars are interquartile.

Proposition 4.7 (Assumption-free operating guarantee)**.** *Let $\bar{x}$ be any point, and let the tilts $\beta_i = \nabla h_i(\bar{x})$ there be arbitrary vectors: no independence, no zero mean, no linearity, no distributional assumption. Let $\kappa < 1$ be their incoherence and let $\widehat{w}_i \propto \widehat{c}_i^{-1}$ be* CROWD*'s weights with $\|\widehat{c} - q\|_\infty \leq \Delta$. Then, deterministically,*

$$\|b(\widehat{w})\|^2 \;\leq\; \frac{1+\kappa}{1-\kappa} \min_{w \in \Delta_n} \|b(w)\|^2 + 2(1+\kappa)\Delta, \tag{6}$$

and the gain over uniform weighting is at least $\frac{1-\kappa}{1+\kappa}\mathrm{AM}(q)/(\mathrm{HM}(q) + 2n\Delta)$, for AM, HM the means of the realised energies.

Inequality (6) makes CROWD $\frac{1+\kappa}{1-\kappa}$-competitive with the exactly optimal weights for the realised instance. This holds even where those weights exploit correlations that CROWD does not model. The crowd hypotheses enter only through the benchmark and the size of Δ. As $\kappa \to 1$ the ratio degrades: nearly aligned tilts admit cancellations that a diagonal rule cannot exploit.

5 Numerical Study

What is compared. We compare nine rules: *uniform*; *size* $w_i \propto n_i$ (FedAvg); *inv-σ*, $w_i \propto \bar{\sigma}_i^{-2}$ (n_i/σ_i^2 on real data); *valid*, a tuned power of each site's held-out loss; *common*, one shared heterogeneity variance; *no-T^{-1}*, CROWD without the inversion; *best single*, the client with the smallest bias; CROWD; and the *oracle*, handed the realised q_i. All rules share optimiser, seeds and gradients, so only w differs.

Bias floors (Figure 2a). The synthetic instance satisfies the model exactly. The truth f_0 is a quadratic on $\mathbb{R}^{150}$ with spectrum geometric in $[\mu, L] = [0.5, 5]$, and each of $n = 20$ clients returns the gradient of $f_i = f_0 + \langle \beta_i, \cdot \rangle$ with Gaussian noise. Bias and noise scales are drawn *independently*, log-uniformly in $[0.2, 3]$ and $[0.5, 3]$. Every rule converges and stalls at the floor set by its own bias. Uniform weighting stops at $3.358 \cdot 10^{-2}$ against $3.361 \cdot 10^{-2}$ predicted, and size weighting and the best single client stop higher still. CROWD stops 8% above the oracle floor, because its weights are built from the identified energies $\widehat{q}_i$. The $7.56\times$ separation is in limit points, not rates.

Model, trained end to end	clients	κ	$\widehat{\kappa}$	inv-σ	CROWD	oracle	CROWD<unif.
two-layer ReLU, raw FEMNIST pixels	10 writers	1.09	1.05	0.815	0.933	0.816	8/8
	20 writers	2.13	2.52	0.971	**0.891**	0.970	7/8
	40 writers, ampl.	3.88	4.39	1.434	**0.950**	1.331	4/6
convolutional net, raw digit scans	8 sensor sites	1.63	1.18	1.141	1.057	1.218	1/6

Table 1: Neural networks trained end to end: excess risk relative to uniform weighting, medians (Appendix H.6). CROWD is unguarded; $\widehat{\kappa}$ is the estimate the guard uses.

Separating bias from noise (Figure 2b). The data are 1797 real digit scans, the model ℓ_2-regularised multinomial logistic regression on a frozen random feature map, so every risk is convex. Eight sites hold disjoint cohorts seen through their own distortion, and f_0 averages the risk over forty distortions from the same law. In the *sensor* federation each site has a detector with its own gain and offset on every pixel. CROWD, the oracle and inverse-noise reach 0.787, 0.789 and 0.778 of the uniform excess risk: a miscalibrated detector inflates noise along with bias. The *convention* federation separates the two costs at the same $\kappa = 0.48$. All detectors share one severity, and sites differ only through their own label-confusion channel. Cohort sizes run *against* quality, so the worst site is the largest one: quiet, but biased. Here precision weighting reaches 1.07, worse than not weighting at all, and validation tuning returns uniform weights. CROWD reaches 0.875, against 0.934 for the bias-energy oracle. The reason is that $\widehat{s}_i$ keeps only the part of a site's disagreement that persists across rounds. Noise averages out, and the quiet but biased site is down-weighted.

Real writers (Figure 2c). In FEMNIST each client is one writer and f_0 the writer-averaged risk of 400 others, so the federation is not constructed. With ten writers $\kappa = 0.52 < 1$, the regime of Theorem 4.4, and CROWD reaches 0.897 of the uniform excess risk, as does the oracle. For natural writers the gain fades and is gone at $n = 40$. The reason is that the bias energy AM/n of the uniform crowd decays with n, so the weights act on a shrinking part of the risk. Amplifying each style by $\sqrt{n/10}$ slows that decay, and the gain holds: 0.929 at $n = 40$, although $\kappa \approx 2$. At $n = 80$ ($\kappa \approx 3.4$) the estimate stops working: CROWD reaches 1.03 and the oracle 0.97.

Unbounded gain (Figure 2d). We take a linear crowd with $n = 20$, $d = 60$, unit noise and $\log \zeta_i$ uniform on an interval of width s. At $s = 3$ uniform weighting falls $11.0\times$ behind the oracle, while CROWD stays within $1.34\times$. The gain of the oracle over uniform weighting is AM/HM of the bias energies, which grows without bound with the spread (Proposition D.5). CROWD captures it because it estimates every q_i separately rather than one shared variance.

Neural networks (Table 1). Outside the convex theory we train a two-layer ReLU network on the writers of Figure 2c and a small convolutional network on the sensor sites. The oracle uses the energies at a reference network. On the two-layer network CROWD lowers the pool risk by 5–11% in the median, with accuracy changing by less than one point. It is the only rule that improves on uniform weighting in all three settings. CROWD measures the disagreement along its own trajectory, whereas inv-σ and the oracle use one fixed point, and at $n = 20, 40$ both lose to it. These gains are unguarded: $\widehat{\kappa} \geq 1$ in 20 of 22 seeds, so the guard below would have kept uniform weights and is conservative here. On the convolutional network no rule improves on uniform weighting, and the guard declines in five of six seeds.

Using CROWD in practice. Probe with uniform weights, refresh on the doubling grid, and at each refresh estimate κ and the drift between the two half-averages (Proposition C.1, Remark C.3). If $\widehat{\kappa} < 1$ and the drift is small, weight, otherwise keep uniform weights. This is a screening diagnostic, not a certificate. The estimate $\widehat{\kappa}$ is not consistent for κ: its false-accept rate is 4.2% in the linear regime, and on a strongly curved crowd it under-reports κ three- to five-fold (Appendix H.2). Appendix H adds further ablations. Among them, local-step FEDAVG, FEDPROX and SCAFFOLD reach the same floor under the same weights (Figure 8).

Limitations. Centring is untestable, so calling an aggregate closer to ground truth is a claim about the crowd. The gain also needs a slowly vanishing summed tilt and not many more clients than independent causes of heterogeneity (Figures 2 and 10).

Reproducibility statement

Each section of the main text has one appendix section of proofs, and every numbered statement has its own proof subsection there; notation is collected in Appendix A. Every experiment is a seeded script. The supplementary materials contain the code, the recorded outputs of every run, a README with per-experiment instructions and runtimes, and a document tracing each quoted number to its recorded output and to the theoretical prediction it is compared against; protocols are in Appendix H.

Ethics statement

The work is theoretical and its experiments use public benchmark data, in part with synthetic distortions; no human subjects, personal data or new data collection are involved. The motivating application is cross-institution federation, and two of the paper's results bear directly on its risks. Proposition 3.1(i) shows that no federated protocol identifies the common truth, so a deployment claiming recovery of ground truth is claiming something about its participants, not about its data; we state it as a theorem because the opposite belief is common. Proposition C.10 shows that if which sites report depends on their own miscalibration – plausible when under-resourced sites are the ones that drop out – then the crowd is centred on a shifted objective, and CROWD returns the weights that are optimal for that shifted objective while appearing well calibrated. A weighting scheme that reduces a site's influence is a governance decision as well as a statistical one, and Proposition 4.7 is stated in competitive form partly so that the worst case against uniform weighting is boundable in advance.

AI use statement

In this work, we used generative AI tools as a general-purpose writing and editing aid – for prose tightening, LaTeX formatting, and for generating literature-search queries whose results were then read and verified by the authors – and, during revision, as an assistant whose output we reviewed: parts of the revised proofs (notably Appendix G.7) and the experimental code of the convolutional stress test (Appendix H.6) were drafted with AI assistance. We have not used generative AI tools to generate the research ideas or the main results. We have reviewed all AI-assisted work: every proof was checked against its statement by hand, and every quoted number is produced by a seeded script and traced to its recorded output in the supplementary materials. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

References

John M. Abowd, Francis Kramarz, and David N. Margolis. High wage workers and high wage firms. *Econometrica*, 67(2):251–333, 1999.

Siavash M. Alamouti and Fay Arjomandi. Optimal information combining for multi-agent systems using adaptive bias learning. *arXiv preprint arXiv:2510.25793*, 2025.

Martyn J. Andrews, Leonard Gill, Thorsten Schank, and Richard Upward. High wage workers and low wage firms: negative assortative matching or limited mobility bias? *Journal of the Royal Statistical Society: Series A*, 171(3):673–697, 2008.

El Mahdi Chayti, Sai Praneeth Karimireddy, Sebastian U. Stich, Nicolas Flammarion, and Martin Jaggi. Linear speedup in personalized collaborative learning. *arXiv preprint arXiv:2111.05968*, 2022.

Rui Chen, Dongyang Bao, and Ning Lu. Robust federated learning for dynamic and heterogeneous environments: an adaptive client selection approach. *Expert Systems with Applications*, 2026.

Wenlin Chen, Samuel Horváth, and Peter Richtárik. Optimal client sampling for federated learning. In *Workshop on Privacy Preserving Machine Learning, NeurIPS*, 2020.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Towards understanding biased client selection in federated learning. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Gerda Claeskens, Jan R. Magnus, Andrey L. Vasnev, and Wendun Wang. The forecast combination puzzle: a simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762, 2016.

William G. Cochran. *Sampling Techniques*. Wiley, 3 edition, 1977.

A. Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C*, 28(1):20–28, 1979.

Ittai Dayan, Holger R. Roth, Aoxiao Zhong, et al. Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27:1735–1743, 2021.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13:165–202, 2012.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: how inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009.

Rebecca DerSimonian and Nan Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7 (3):177–188, 1986.

Mathieu Even, Laurent Massoulié, and Kevin Scaman. On sample optimality in personalized collaborative and federated learning. In *Advances in Neural Information Processing Systems*, 2022.

Francis Galton. Vox populi. *Nature*, 75:450–451, 1907.

Christian Genest and James V. Zidek. Combining probability distributions: a critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

Felix Grimberg, Mary-Anne Hartley, Sai Praneeth Karimireddy, and Martin Jaggi. Optimal model averaging: towards personalized collaborative learning. *arXiv preprint arXiv:2110.12946*, 2021.

Larry V. Hedges and Ingram Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, 1985.

Julian P. T. Higgins and Simon G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

Chi Jin, Praneeth Netrapalli, Rong Ge, Sham M. Kakade, and Michael I. Jordan. A short note on concentration inequalities for random vectors with subGaussian norm. *arXiv preprint arXiv:1902.03736*, 2019.

Peter Kairouz, H. Brendan McMahan, et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2021.

David R. Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 2020.

Yaron Kiselman and Kfir Y. Levy. SP-CACW: Convergence-aware client weighting for selfish personalized learning. *arXiv preprint arXiv:2606.29322*, 2026.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International Conference on Machine Learning*, 2020.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1–2):365–397, 2012.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28(5):1302–1338, 2000.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2020a.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020b.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.

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

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2 edition, 2018.

Scott E. Page. Making the difference: applying a logic of diversity. *Academy of Management Perspectives*, 21(4):6–20, 2007.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Jonatan Reyes, Lisa Di Jorio, Cecile Low-Kam, and Marta Kersten-Oertel. Precision-weighted federated learning. *arXiv preprint arXiv:2107.09627*, 2021.

Nicola Rieke, Jonny Hancox, Wenqi Li, et al. The future of digital health with federated learning. *npj Digital Medicine*, 3(119), 2020.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Micah J. Sheller, Brandon Edwards, G. Anthony Reina, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10 (12598), 2020.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Jeremy Smith and Kenneth F. Wallis. A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71(3):331–355, 2009.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

James H. Stock and Mark W. Watson. Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting*, 23(6):405–430, 2004.

James Surowiecki. *The Wisdom of Crowds*. Doubleday, 2004.

Malte Tölle, Fernando Navarro, Sebastian Eble, Ivo Wolf, Bjoern Menze, and Sandy Engelhardt. FUNAvg: Federated uncertainty weighted averaging for datasets with diverse labels. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Areti Angeliki Veroniki, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian P. T. Higgins, Dean Langan, and Georgia Salanti. Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, 7(1): 55–79, 2016.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Blake Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local SGD better than minibatch SGD? In *International Conference on Machine Learning*, 2020.

Mingwei Xu, Xiaofeng Cao, Ivor W. Tsang, and James T. Kwok. Aggregation weighting of federated learning via generalization bound estimation. *arXiv preprint arXiv:2311.05936*, 2023.

Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I. Jordan. Spectral methods meet EM: a provably optimal algorithm for crowdsourcing. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

A Notation

Symbols used in more than one section, with the place they are defined. Two symbols collide: ℓ and ℓ_i are curvature bounds while ℓ_T is a logarithmic factor in the stepsize, and K counts samples per client in $\Psi(K)$ but rounds in Algorithm 2, where K rounds give bK samples. The family Ψ carries several decorations, listed together below.

Hats denote estimates computed from the transcript and tildes denote quantities centred within the crowd, so $\widetilde{x}_i = x_i - \frac{1}{n}\sum_j x_j$ throughout. Statements labelled *realised* or *deterministic* hold for one fixed draw of the tilts with no expectation.

B Extended related work

Section 1.1 groups the literature by the role the weights play in it. This appendix adds the detail that did not fit, and is confined to the places where the comparison needs more than a clause.

Why meta-analysis stops at a common heterogeneity variance. The normal–normal hierarchical model of random-effects meta-analysis assigns each study a within-study variance and a single shared between-study variance. The latter is estimated by the method of moments, by REML, or by one of the many alternatives surveyed in Veroniki et al. (2016). The standard heterogeneity summaries (Higgins & Thompson, 2002) are likewise scalar. This is structural: a study reports once, so the observed effects cannot support one variance component per study alongside the common mean. A federated round is a repeated draw from the same client, which is what makes the per-client energies estimable here. Cross-fitting over rounds removes the noise offset without knowing the noise level, and Theorem 4.1 converts the resulting disagreement energies into per-client heterogeneity. With a common heterogeneity variance the harmonic law reduces to inverse-noise weighting, and the

Symbol	Defined in	Meaning
n, d, b, K	Sec. 2	clients, dimension, batch, samples per client
$f_0,\ x_0^\star,\ \mu, L$	Assum. A2	the truth, its optimum, its curvature bounds
$h_i,\ f_i = f_0 + h_i$	Assum. A1	client i's perturbation and risk
$\ell_i,\ \ell,\ \lambda_w,\ L'$	Assum. A1	curvature bound of h_i; $\ell = \max_i \ell_i$; $\lambda_w = \sup\|\nabla^2 H_w\|$; $L' = L + \lambda$
ℓ_T	Thm. 4.4	logarithmic factor in the tuned stepsize, unrelated to ℓ
$\beta_i,\ q_i,\ \zeta_i^2$	Assum. A1	realised tilt, realised energy $\|\beta_i\|^2$, expected energy
$\bar\beta,\ \widetilde\beta_i$	Eq. (3)	crowd mean and disagreement of the tilts
$b(w),\ \Gamma_w$	Eq. (1)	aggregate bias vector and its norm
$V(w)$	Sec. 4.2	diagonal proxy $\sum_i w_i^2 q_i$ for Γ_w^2
AM, HM	Prop. D.5	arithmetic and harmonic means of the energies
Ψ_0	Prop. D.5	HM/n, the attainable expected bias energy
$\Psi(K),\ \widehat\Psi(K)$	Eq. (2)	finite-sample target and its realised counterpart
$\kappa,\ \widehat\kappa$	Assum. A4, Prop. C.1	incoherence of the realised tilts and its estimate
$\sigma_i^2,\ \tau_i^2$	Assum. A3	oracle noise and its centred counterpart
$\widehat g_i^k,\ r_i^k$	Eq. (3)	report and residual at round k
$T,\ C_n$	Thm. 4.1	moment map and its ℓ_∞ condition number
$\widehat s_i,\ \widehat q_i,\ \widehat w$	Eq. (4)	cross-fitted statistic, inverted energy, weights
$\Delta,\ \delta_i,\ \vartheta_i$	Lem. 4.2	weight error; its staleness and statistical parts
$\epsilon,\ \epsilon_T$	Lem. D.2(ii)	Young trade-off parameter of the analysis
$P_k,\ m,\ p,\ \rho$	Sec. C.5	participating set, its size and rate m/n, the shrinkage of the disagreement

common-variance curve in Figure 7 recovers none of the gain. The parallel with the moment identity of DerSimonian & Laird (1986) is exact in form: the heterogeneity statistic of meta-analysis and our cross-fitted disagreement statistic are both moment inversions of an ANOVA-type decomposition. Three differences matter. Our sources are optimisation clients queried repeatedly at moving, algorithm-chosen points, rather than fixed summary statistics. The effect is high-dimensional and the relevant functional is a quadratic form. And we need per-client heterogeneity rather than a pooled variance, which the harmonic law of Proposition D.5 requires.

Learning whom to learn from. Even et al. (2022) prove matching upper and lower bounds on sample complexity for collaborative learning, with the gain governed by the number of agents close to the target agent. Chayti et al. (2022) and Grimberg et al. (2021) give related bias–variance trade-offs. In all of these the weights are a design variable chosen from pairwise discrepancy bounds that the algorithm is given, and the authors name the knowledge of those bounds as the main limitation of the approach. In the one-truth case the present paper removes this requirement. The per-client energies play the role of the pairwise discrepancies. By Theorem 4.1 their expectations are identified by the law of the transcript, and the realised ones are estimable up to the incoherence floor of Section 4.2. The settings also differ in target. That line personalises, so each agent has its own risk and its own weight vector, and each agent can evaluate its own risk on held-out data. Here there is one objective that nobody observes, and Proposition 3.1 follows from that.

Bias-aware combination with an observable target. Alamouti & Arjomandi (2025) study multi-agent combination with agent-specific biases and reach the same harmonic value at inverse-cost weights. Two differences matter. First, their biases depend on observable covariates, and their learnability ratio measures the fraction that is predictable from them. Correcting bias is therefore legitimate in their model, whereas Proposition 3.1(ii) shows it is futile in ours, where no such side information exists. Their zero-learnability regime is our setting, and there their main theorem reduces to oracle inverse-variance weighting without addressing how the energies are obtained. Second, their plug-in estimates each agent's energy by the average squared distance to the current consensus, computed from the same data that built the consensus. This is the statistic that Lemma F.1 corrects.

Kiselman & Levy (2026) select aggregation weights by minimising an upper bound on a designated client's convergence error. The rule trades peer bias against stochastic variance and assigns zero weight to harmful peers; it is evaluated on MNIST, CIFAR-100 and LEAF Shakespeare. Read as an algebraic rule, the resulting closed form is again of inverse-cost type and agrees with Proposition D.5. The settings differ in what is observable, and the substantive differences follow from that. Their target is the designated client's own risk, which that client can evaluate on held-out data. Peer bias

relative to the target is then a measurable quantity, and estimating it is a regression problem rather than an identification problem. No analogue of Proposition 3.1(i) can arise when the target is one client's own risk. Consequently they estimate peer bias by comparison with the designated client. Our residual is instead a maximal invariant of the shift group, and the inversion of Theorem 4.1 is needed because no reference client exists. Their guarantee is the minimisation of an upper bound; ours is a matching pair. The lower bound of Theorem 4.5 has no counterpart there, since with an observable target the problem is not statistically obstructed. Finally they produce one weight vector per client and do not aim at a common objective, while we produce one. The two are complementary: a federation could use their rule for personalisation and CROWD for its shared objective. The residual statistic is available to both at the same cost.

Crowdsourcing, differencing, and the optimisation toolkit. Estimating annotator reliabilities from agreement patterns is classical (Dawid & Skene, 1979; Karger et al., 2011; Zhang et al., 2014). Those models are discrete-label and typically identify reliabilities from the spectrum of an agreement matrix. Our setting is continuous, high-dimensional and coupled to an optimisation loop. The items are the algorithm's own iterates, chosen adaptively, and the answers are stochastic gradients whose common part is an unknown nuisance that changes every round. The centring that removes it is the analogue of fixed-effects differencing in the two-way models of econometrics (Abowd et al., 1999). There the same limited-mobility bias in variance components is well documented (Andrews et al., 2008). Our cross-fitting over disjoint halves of the rounds is the sample-splitting device of Chernozhukov et al. (2018), used here to remove the noise-induced inflation exactly. The optimisation components are standard: projected SGD for smooth strongly convex objectives (Robbins & Monro, 1951; Nemirovski et al., 2009; Polyak & Juditsky, 1992) and minibatch stochastic approximation (Lan, 2012; Dekel et al., 2012). Our lower bound is not an oracle-complexity bound: it restricts only the number of gradient queries per client, not computation or adaptivity. The obstruction is purely statistical, a Gaussian location problem with a diffuse nuisance in the spirit of Lehmann & Romano (2005); Tsybakov (2009). The concentration tools are standard (Vershynin, 2018; Laurent & Massart, 2000; Jin et al., 2019).

Applications. Cross-institution medical federations are the motivating example (Rieke et al., 2020; Sheller et al., 2020; Dayan et al., 2021). Site-specific measurement offsets, coding conventions and label-prevalence shifts are the mechanism that the linear-tilt model abstracts, and the size-proportional weights used in practice are, by Proposition D.5, arbitrarily suboptimal when site quality and site size are unrelated. Linearised covariate-shift corrections (Shimodaira, 2000) give another route to the same tilt structure.

C FURTHER RESULTS

This section states the results that the main text relies on but does not state. Each backs a specific claim there; the proofs are in Appendix G.

Result	Claim in the main text it supports
Prop. C.1, Rem. C.2–C.3	κ can be estimated from the two half-averages, which gives the guard used in Section 5 (Section 4.2)
Prop. C.4	independence is necessary for identifying the energies (Section 4.1)
Thm. C.5	the refreshing implementation used in the synthetic experiments keeps the bias constant of Theorem 4.4 (Sections 3, 4.2)
Thm. C.7, Rem. C.8	the end-to-end bound for a non-linear crowd and the condition on ℓ/μ (Section 4.4)
Props. C.9, C.10	partial participation, the usual federated setting, keeps the energies identified unless participation depends on the tilts (ethics statement)
Thm. C.11, Cor. C.12	the per-instance comparison of Corollary 4.6 closes to an unconditional Bayes-optimality statement (Section 4.3)

C.1 What the transcript can and cannot be used to check

Centring (Assumption A1) and incoherence (Assumption A4) differ in whether a practitioner can check them from the transcript. By Proposition 3.1(i), centring cannot be tested at all. Incoherence is a property of the realised tilts, and the residuals carry enough information to estimate it.

Proposition C.1 (Checkability). *Under Assumptions A1–A3:*

(i) (Centring is not checkable.) *For every $c \in \mathbb{R}^d$ the instance $(f_0, h_1, \dots, h_n)$ and the instance obtained by replacing f_0 with $f_0 + \langle c, \cdot\rangle$ and each h_i with $h_i - \langle c, \cdot\rangle$ generate the same joint law of the transcript under every protocol, while their targets and their optimal weights differ. Hence no test based on the transcript has power against "the crowd is centred on a distorted objective".*

(ii) (Incoherence is estimable up to a rank-two correction.) *Assume in addition that the tilts are linear, and let $J := I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. The shift of (i) maps β_i to $\beta_i - c$ and Q to $Q - \mathbf{1}a^\top - a\mathbf{1}^\top + \|c\|^2\mathbf{1}\mathbf{1}^\top$ with $a_i = \langle \beta_i, c\rangle$, which changes κ in general, so κ is not a function of the transcript law; only JQJ is. Let $\widehat{G}$ be the matrix with entries $\widehat{G}_{ij} = \frac{1}{2}(\langle \widehat{\beta}_i^{(1)}, \widehat{\beta}_j^{(2)}\rangle + \langle \widehat{\beta}_i^{(2)}, \widehat{\beta}_j^{(1)}\rangle)$ formed from the two half-averages of (4), let $\widehat{q} = (T^{-1}\operatorname{diag}\widehat{G})_+$, and define*

$$\widehat{Q}_{ij} := \widehat{G}_{ij} + \frac{\widehat{q}_i + \widehat{q}_j}{n} - \frac{1}{n^2}\sum_k \widehat{q}_k, \qquad \widehat{\kappa} := \big\|\widehat{D}^{-1/2}(\widehat{Q} - \widehat{D})\widehat{D}^{-1/2}\big\|_{\mathrm{op}}, \quad \widehat{D} = \operatorname{diag}\widehat{q}.$$

Then $\mathbb{E}[\widehat{G} \mid B] = JQJ$ exactly, for every adaptive query sequence. The correction makes every row sum of $\widehat{Q} - \widehat{D}$ vanish identically, and zero is also the mean, under Assumption A1, of the off-diagonal row sums $\varsigma_i := \sum_{j\neq i} Q_{ij}$ of Q. With $\widehat{G}$ and $\widehat{q}$ replaced by their targets JQJ and q, the resulting $\widehat{\kappa}_\circ$ satisfies, deterministically,

$$|\widehat{\kappa}_\circ - \kappa| \;\le\; \kappa\Big(2\sqrt{\mathrm{AM}(q)/\mathrm{HM}(q)} + \mathrm{AM}(q)/\mathrm{HM}(q)\Big).$$

Proved in Appendix G.1.

The hypothesis that makes the target meaningful cannot be tested. Any claim that a federated aggregate is closer to ground truth rests on it. The hypothesis that makes the estimator work can be estimated cheaply: $\widehat{\kappa}$ reuses the two half-averages the algorithm already holds and costs one $n \times n$ Gram product. The estimate is not consistent for κ: its population error is a rank-two term of the same order as κ itself, which is the price of κ not being identified. Where the estimate is reliable is therefore a question for measurement, and Section H.2 calibrates the threshold empirically. Without the linearity hypothesis the residual measures the disagreement at the iterate rather than at $x_0^\star$, and in our runs this deflates $\widehat{\kappa}$ (Section H.2). This suggests running CROWD with a guard.

Remark C.2 (Guarded CROWD: a screening heuristic, not a certificate). Compute $\widehat{\kappa}$ at the refresh rounds. If $\widehat{\kappa} < 1$, use CROWD and report $\frac{1+\widehat{\kappa}}{1-\widehat{\kappa}}$ as an estimate of the competitive ratio of Proposition 4.7. If $\widehat{\kappa} \ge 1$, the diagonal proxy is uninformative on this instance and the bound is vacuous; fall back to uniform weighting. The guard indicates which regime the deployment is likely in but does not certify it. Proposition C.1 shows that $\widehat{\kappa}$ is not consistent for κ: its population error is of the same order as κ itself. The reported ratio is therefore an estimate, not a bound. In place of a guarantee the guard comes with a measured error rate. It is cheap, and in the linear regime it gives the deployment an estimate of the quantity Assumption A4 bounds. Section H.1 measures $\widehat{\kappa}$ against κ on three mechanisms, and Section H.2 calibrates the threshold: the empirical probability that $\kappa \ge 1$ given $\widehat{\kappa} < 1$ is 4.2%. The restriction to a near-linear crowd matters. The estimator inherits the staleness of Lemma 4.2, so when probing happens far from the optimum $\widehat{\kappa}$ under-reports, by a factor three to five in Section H.2. That failure is detectable and partly repairable, and Remark C.3 does both. In the real-data study of Section H.3, where the probe is warm, $\widehat{\kappa} = 0.52$ against a realised $\kappa = 0.48$. The guard is also conservative: CROWD often still improves on uniform weighting when κ exceeds one, at $\kappa = 1.28$ in Section H.4, $\kappa \approx 2$ in Section H.5 and $\kappa \approx 2$ to 3.9 in Section H.6.

Remark C.3 (A staleness-aware guard, and what it costs). The restriction to a near-linear crowd can be checked on the trajectory the algorithm already produces. Split the probing window into two disjoint *time*-halves, compute $\widehat{\kappa}$ on each from its own parity split, and read the drift between them. A crowd that is effectively linear over the probing region returns the same value twice; a curved one

drifts, because the two halves measure the Gram at different points. Scaling that drift by how far the iterate still has to travel, which the probing displacement itself estimates, gives

$$\widehat{\kappa}_{\mathrm{corr}} = \widehat{\kappa}_{\mathrm{late}} + \left|\widehat{\kappa}_{\mathrm{late}} - \widehat{\kappa}_{\mathrm{early}}\right| \cdot \frac{\|x^{\mathrm{end}} - x^{\mathrm{start}}\|}{\|x^{\mathrm{end}} - x^{\mathrm{mid}}\|},$$

at the cost of one extra pair of accumulators and no extra samples. The rule is a heuristic. The supplementary materials measure it on two mechanisms and three probing budgets:

Mechanism	Probe	κ	$\widehat{\kappa}$	drift	$\widehat{\kappa}_{\mathrm{corr}}$
low rank	cold, short	2.19	1.48	0.054	1.88
low rank	cold, long	2.19	1.48	0.007	1.53
low rank	warm, long	2.19	1.48	0.001	1.48
full rank	cold, short	0.49	0.49	0.019	0.71
full rank	cold, long	0.49	0.49	0.004	0.57
full rank	warm, long	0.49	0.49	0.002	0.49

Three observations follow. The drift falls by one to two orders of magnitude as the probe warms and lengthens, so it is a working detector of a stale measurement. The correction is inert on a settled probe, leaving both estimates untouched, so it does not penalise a good measurement with a false refusal. And on a cold short probe it inflates in the safe direction, moving the low-rank case from 1.48 towards the true 2.19 while leaving the full-rank case at 0.71, still below the threshold. The gap does not close entirely: even on the best probe $\widehat{\kappa} = 1.48$ against $\kappa = 2.19$. The remaining gap comes from the change of evaluation point. The estimate is of the incoherence *at the probing point*, which is the quantity Lemma F.1 needs – the inversion is performed on the disagreement measured there – whereas Proposition 4.7 compares bias energies at the truth's own optimum. The practical rule therefore has two parts. First, probe after a burn-in and check the drift, which supports the inversion. Second, treat the bound of Proposition 4.7 as conditional on the drift being small, since only then are the two evaluation points interchangeable.

C.2 Correlated crowds

Independence is what makes the cross terms vanish in Proposition D.5 and what makes T the right map to invert. Without it the optimal weights are a minimum-variance portfolio, and the disagreement no longer determines them.

Proposition C.4 (Correlation destroys identifiability)**.** *For jointly Gaussian centred crowds with* $\mathrm{Cov}(\beta_i, \beta_j) = \Sigma_{ij} I_d/d$, $\mathbb{E}\|b(w)\|^2 = w^\top \Sigma w$ *for deterministic* w, *minimised over* $\{\mathbf{1}^\top w = 1\}$ *by the minimum-variance portfolio* $w \propto \Sigma^{-1}\mathbf{1}$ *(Markowitz, 1952), which is also the minimiser over* Δ_n *when* $\Sigma^{-1}\mathbf{1} \geq 0$. *The map* $\Sigma \mapsto J\Sigma J$ *has kernel exactly* $\{\mathbf{1}u^\top + u\mathbf{1}^\top : u \in \mathbb{R}^n\}$, *and for* $\Sigma = I_3$ *versus* $\Sigma' = I_3 + \mathbf{1}u^\top + u\mathbf{1}^\top$, $u = (0.1, 0, 0)^\top$ *– both positive definite – the disagreement residuals have identical laws while the optimal weights differ,* $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ *versus* $(0.267, 0.367, 0.367)$. *Under* Σ' *the uniform weights attain* 0.4000 *against the optimum* 0.3933, *and under* Σ *the weights optimal for* Σ' *attain* 0.3400 *against* $\frac{1}{3}$. *Hence no rule that is a function of the residuals is optimal for both.*

Proved in Appendix G.2.

The kernel is n-dimensional and meets the diagonal matrices only at zero when $n \geq 3$, since $\mathbf{1}u^\top + u\mathbf{1}^\top$ is diagonal only if $u_i + u_j = 0$ for all $i \neq j$. Independence restricts Σ to the diagonal family, where $J\Sigma J$ is therefore injective; that injectivity is the invertibility of T in Theorem 4.1.

C.3 The anytime algorithm

Theorem 4.4 analyses the two-phase schedule, in which the weights are computed once and then frozen. The experiments run the anytime version, in which the weights are refreshed and therefore depend on the trajectory they themselves generate. This subsection analyses the refresh schedule we recommend, a geometric grid.

Theorem C.5 (Anytime Crowd)**.** *Let the assumptions of Theorem 4.4 hold, fix* $\gamma > 1$, *and run Algorithm 1 with the refresh set* $\mathcal{R} = \{\lceil \gamma^j \rceil : j \geq j_0\} \cap [1, T/\gamma]$ *with* $j_0 \geq 1$ *and* $T \geq \gamma^{j_0+1}$, *so that* $\mathcal{R}$ *contains at least one refresh (repeated values of* $\lceil \gamma^j \rceil$ *are counted once), using at each refresh*

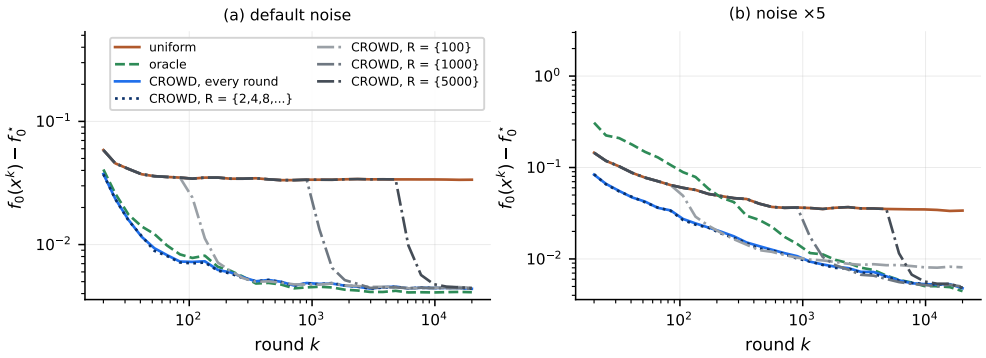


Figure 3: Excess risk on the truth for different refresh sets $\mathcal{R}$, synthetic instance, medians over 12 seeds. All schedules of one seed see the same gradient noise.

all rounds seen so far and the horizon of the coming segment as K, and on each segment the tuned stepsize of Lemma F.6 restarted with that segment's length and a fixed constant $\bar{R} > 0$ in place of R. Let $t_J \le T/\gamma$ be the last refresh, Δ_J the weight error there, $T_J = T - t_J \ge (1-\gamma^{-1})T$ the length of the final segment, $\bar{V} := \max_i \bar{\sigma}_i^2/b$ and $\bar{\ell} := 1 + \ln(e + \mu^2\bar{R}^2 T/\bar{V})$. Then, conditionally on the tilts, for every $\epsilon > 0$,

$$\mathbb{E}\big[f_0(x^T) - f_0^\star\big] \;\le\; \frac{(1+\epsilon^{-1})L^2}{2\mu}\Big(\widetilde{R}^2\Big(e^{-\mu T_J/(2L)} + \frac{\bar{V}}{\mu^2\bar{R}^2 T_J}\Big) + \frac{\bar{\ell}\bar{V}}{\mu^2 T_J}\Big) + \frac{(1+\epsilon)(1+\kappa)}{2\mu}\big(\widehat{\Psi}(bT_J) + 2\mathbb{E}\Delta_J\big),$$

where $\widetilde{R}^2 := \mathbb{E}\|x^{t_J} - x^\star_{w^{(J)}}\|^2$ is bounded by a constant that depends on the instance and on γ but not on T (Remark C.6), and Δ_J is built from $t_J \ge T/\gamma^2 - 1$ rounds. The bound is that of Theorem 4.4 with T replaced by T_J, so the anytime schedule costs a factor $(1-\gamma^{-1})^{-1}$ in the vanishing terms and nothing in the bias term.

Proved in Appendix G.3.

Remark C.6 (A bad early refresh cannot spoil the bound). Every $w^{(j)}$ lies in the simplex, so $V(w^{(j)}) \le \max_i q_i$ and, by Lemma D.3 and Lemma D.2(i), every target lies within $G := \sqrt{(1+\kappa)\max_i q_i/\mu}$ of $x_0^\star$, whatever the estimation errors Δ_j were. Appendix G.3 turns this into

$$\widetilde{R} \;\le\; \|x^0 - x_0^\star\| + G + (N_\star + 2)\,c_\star, \qquad c_\star := 2G + \sqrt{\frac{\bar{V}}{2L\mu}} + \sup_{t \ge T_\star}\sqrt{\frac{(1+\ln(e+\mu^2\bar{R}^2 t/\bar{V}))\bar{V}}{\mu^2 t}},$$

with $T_\star := \max\{4L\ln 8/\mu,\, 8\bar{V}/(\mu^2\bar{R}^2)\}$ and $N_\star := \lceil \log_\gamma(2T_\star/(\gamma-1))\rceil + 1$, none of which depends on T. A refresh that is badly wrong moves the target by a bounded amount.

Two further remarks. First, the result does not require Δ_j to be monotone in j: only the last refresh enters the bias term, and earlier refreshes affect the bound solely through $\widetilde{R}$. Second, a uniform grid would give $J = \Theta(T)$ segments and the restart cost of the tuned stepsize would no longer be negligible, which is why the geometric grid is the schedule we analyse and recommend.

Refresh schedules in practice. The synthetic runs of Section 5 refresh at every round with a decaying stepsize, a schedule the theorem does not cover; the real-data runs refresh once, after probing, as in Theorem 4.4. Figure 3 compares the schedules on the default synthetic instance and on the same instance with five times the noise (12 seeds). After $2 \cdot 10^4$ rounds, refreshing every round and refreshing on the doubling grid (14 refreshes) end at 1.08 and 1.08 times the oracle's exact floor under the default noise, and at 1.21 and 1.21 under the larger noise. A single refresh pays the bias of uniform weighting until it happens, and if it happens early it freezes a large estimation error. Under the larger noise a refresh at $k = 100$ ends at 1.89 times the floor, with an error in the energies of $0.31\,\bar{q}$ against $0.11\,\bar{q}$ for the doubling grid. A single refresh at $k = 1000$ ends at 1.17.

C.4 THE NON-LINEAR BOUND

The next result carries the end-to-end bound of Theorem 4.4 through a curved crowd.

Theorem C.7 (Non-linear crowd)**.** *Let Assumptions A1–A4 hold on a closed ball $\mathcal{X}$ of diameter $\bar{D}$, with $\lambda := \max_{w\in\Delta_n} \lambda_w \le \ell < \mu/2$, and write $L' = L + \lambda$ and $\Gamma^2_{\max} := (1+\kappa)\max_i q_i$. Assume the interior of $\mathcal{X}$ contains the closed ball of radius $\Gamma_{\max}/(\mu-\lambda)$ about $x_0^\star$. Run Algorithm 1 with $\mathcal{R} = \{K_0\}$ and uniform weights during probing, then T rounds of projected* SGD *on $F_{\widehat{w}}$ with the tuned stepsize of Lemma F.6 and $R = \bar{D}$, and put $V_\sigma := \sum_i \widehat{w}_i^2\bar{\sigma}_i^2/b$ and $\ell_T := 1 + \ln(e + (\mu-\lambda)^2\bar{D}^2T/V_\sigma)$. Then Lemma 4.2 applies with $\delta_i = \widetilde{\ell}_i\bar{D}$, $\widetilde{\ell}_i = \ell_i + \frac{1}{n}\sum_j \ell_j$, and, conditionally on the first K_0 rounds, with $R_0 = \bar{D}$ below,*

$$\mathbb{E}\big[f_0(x^{K_0+T}) - f_0^\star\big] \;\le\; \frac{LL'}{\mu-\lambda}\Big(R_0^2 e^{-(\mu-\lambda)T/(2L')} + \frac{\ell_T V_\sigma}{(\mu-\lambda)^2 T}\Big) \;+\; \frac{L(1+\kappa)}{(\mu-\lambda)^2}\big(\widehat{\Psi}(bT) + 2\Delta\big).$$

The sharp bias constant of Lemma D.2(ii) is available only for a linear crowd, so the bias term carries $L/(\mu-\lambda)^2$ here rather than the $1/(2\mu)$ of Theorem 4.4; at $\lambda = 0$ the two differ by the factor $2L/\mu$.

Proved in Appendix G.4.

Remark C.8 (Where the condition on ℓ/μ in Section 4.4 comes from). The condition is an order-of-magnitude reading of Theorem C.7. Its proof uses $\bar{D}$ only as a bound on $\|x^{k-1} - x_0^\star\|$ during probing (Step 1 of Appendix G.4), so if the probing iterates stay within distance r of $x_0^\star$, Lemma 4.2 applies with $\delta_i = \widetilde{\ell}_i r$. By Lemma 4.2(ii)–(iii) the non-linear contribution to Δ is then $C_n(2\max_i \widetilde{\zeta}_i\widetilde{\ell}_i r + \max_i \widetilde{\ell}_i^2 r^2)$, and $\widetilde{\ell}_i \le 2\ell$. Suppose, as in a well-started run, that probing takes place near the minimiser of the uniform objective, $r \le c\sqrt{\mathrm{AM}/n}/\mu$ (the radius Lemma D.2(i) gives for $x^\star_{w_{\mathrm{unif}}}$ when $\Gamma^2_{w_{\mathrm{unif}}} \approx \mathrm{AM}/n$), and that no single energy dominates, $\max_i \widetilde{\zeta}_i^2 \le c'\mathrm{AM}$. Then the first term is at most $4cc'^{1/2}C_n\ell\,\mathrm{AM}/(\mu\sqrt{n})$, and asking it to be at most $\epsilon\Psi_0 = \epsilon\mathrm{HM}/n$ gives $\ell/\mu \lesssim \epsilon\, n^{-1/2}\mathrm{HM}/\mathrm{AM}$. The quadratic term is smaller by a further factor of the same order. Without the two suppositions the condition changes accordingly; the theorem itself does not need them.

C.5 PARTIAL PARTICIPATION

Assumption A5 (Participation)**.** At round k a set $P_k \subseteq [n]$ with $|P_k| = m \ge 2$ reports, drawn uniformly at random and independently of $\mathcal{F}_{k-1}$ and of the round-k noises. The residual is centred *within* the participating set, $r_i^k = \widehat{g}_i^k - \frac{1}{m}\sum_{j\in P_k}\widehat{g}_j^k$ for $i \in P_k$.

Lemma E.1 applies verbatim with $[n]$ replaced by P_k: the residual still cancels $\nabla f_0(x^{k-1})$ identically, since the centring is again an average of the participants' reports. What changes is the linear map to be inverted and the effective sample size.

Proposition C.9 (Identifiability under uniform participation)**.** *Let Assumptions A1, A3 (linear tilts) and A5 hold, $n \ge 3$, and put $p := m/n$ and $\rho := \frac{(m-1)n}{m(n-1)} \in [\frac{1}{2}, 1]$.*

(i) (The disagreement shrinks by a known factor.) *For every i and k, $\mathbb{E}[r_i^k \mid \mathcal{F}_{k-1},\, i \in P_k] = \rho\,\widetilde{\beta}_i$, with $\widetilde{\beta}_i$ the full-crowd disagreement of Theorem 4.1.*

(ii) (An unbiased statistic.) *For the odd and even halves $\mathcal{K}_a$ of the probing rounds let $N_i^{(a)} := |\{k \in \mathcal{K}_a : i \in P_k\}|$ and, when both are positive, $\widehat{\beta}_i^{(a)} := \frac{1}{N_i^{(a)}}\sum_{k\in\mathcal{K}_a,\, i\in P_k} r_i^k$ and $\widehat{s}_i := \langle\widehat{\beta}_i^{(1)}, \widehat{\beta}_i^{(2)}\rangle$. Conditionally on the tilts and on the participation indicators of client i, $\mathbb{E}\widehat{s}_i = \rho^2\|\widetilde{\beta}_i\|^2$ and $\mathbb{E}(\widehat{s}_i - \rho^2\|\widetilde{\beta}_i\|^2)^2 \le 2\rho^2\|\widetilde{\beta}_i\|^2(\vartheta_{i,1}^2 + \vartheta_{i,2}^2) + 2\vartheta_{i,1}^2\vartheta_{i,2}^2$, where*

$$\vartheta_{i,a}^2 := \frac{1}{N_i^{(a)}}\Big(\psi_{m,n}\bar{q} + \frac{\tau_{i,P}^2}{b}\Big),$$

$\psi_{m,n} := \frac{n(n-m)}{m(n-1)(n-2)}$ and $\tau_{i,P}^2 := \sigma_i^2 + \frac{1}{m(n-1)}\sum_{j\ne i}\sigma_j^2$. With probability at least $1 - 2ne^{-pK_0/16}$, every $N_i^{(a)} \ge pK_0/4$.

(iii) (Weights.) $\widehat{q} := (T^{-1}\widehat{s}/\rho^2)_+$ *satisfies* $\max_i |\widehat{q}_i - q_i| \le C_n\big(\max_i |\widehat{s}_i/\rho^2 - \|\widetilde{\beta}_i\|^2| + 3\kappa\bar{q}\big)$.

The energies are therefore identified for every $m \ge 2$ *once* $n \ge 3$. *The price of partial participation is a factor of order* $1/p$ *in the number of reports, a factor* $\rho^{-2} \le 4$ *from rescaling, and the variance* $\psi_{m,n}\bar{q}$ *of drawing the participants, which vanishes at* $m = n$.

Proved in Appendix G.5.

Lemma 4.3 and Proposition 4.7 then apply verbatim with the Δ of (iii). The optimisation phase under partial participation has a sampling variance of its own, which we do not analyse. A per-round squared disagreement has a different mean, $\mathbb{E}[\|\widetilde{\beta}_i^P\|^2 \mid i \in P]$, because it also contains the variance of drawing P; cross-fitting over rounds with independently drawn participants removes that variance, so the map to invert is $\rho^2 T$. Participation that is observed and independent of the tilts is therefore harmless. Selection on the tilts themselves is a different matter.

Proposition C.10 (Selection on the tilts moves the target)**.** *Suppose that the clients that report are those in a set* $\mathcal{R}$, *and that their tilts are* $\beta_i = c + \beta_i'$, $i \in \mathcal{R}$, *with* $c \in \mathbb{R}^d$ *fixed and the* β_i' *independent and centred, as happens when tilts are drawn i.i.d. and selection depends on each client's own tilt. Then: (i) the transcript of any protocol has the same law under* $(f_0, (h_i)_{i\in\mathcal{R}})$ *and under* $(f_0 + \langle c, \cdot\rangle, (h_i - \langle c, \cdot\rangle)_{i\in\mathcal{R}})$, *and the second instance satisfies Assumption A1 on* $\mathcal{R}$; *(ii) for every* w *in the simplex on* $\mathcal{R}$, $b(w) = c + b'(w)$ *with* $b'(w) := \sum_i w_i\beta_i'$, *so no weighting removes* c; *(iii) the disagreements of the two instances coincide, so* CROWD *computes the same statistics under both, and they estimate the energies* $\|\beta_i'\|^2$ *of the second.*

Proved in Appendix G.6.

The mechanism is the one of Proposition C.4: a component shared by the reporting clients is invisible to the disagreement. Stragglers who are systematically the miscalibrated sites therefore shift the target the crowd is centred on, by an amount that no weighting rule built on the residuals can detect. The guarantee of Proposition C.9 requires participation to be ignorable, an assumption about the deployment that the data cannot check.

C.6 CLOSING THE LOOP: UNCONDITIONAL BAYES OPTIMALITY

Corollary 4.6 compares a conditional, fixed-centre upper bound with a worst-case, prior-averaged lower bound, and says so. This subsection reduces the asymmetry. The device is to compare *Bayes risks in one and the same formulation*. The centre is drawn from the Gaussian prior at a finite scale ρ. Both the floor and CROWD's risk are then unconditional expectations over the centre, the crowd, the oracle noise and the algorithm's randomness. Neither side is conditional or pointwise, and both are the same functional of the same experiment.

Throughout, fix $\mu > 0$, $n \ge 3$, energies $\zeta \in (0,\infty)^n$ with means AM, HM, noise levels $\sigma \in [0,\infty)^n$, and a batch size b; for each dimension d consider the Gaussian class of Theorem 4.5,

$$f_0 = \tfrac{\mu}{2}\|x - x_c\|^2, \qquad \beta_i \sim \mathcal{N}\big(0, \tfrac{\zeta_i^2}{d} I_d\big), \qquad \text{query noise } \mathcal{N}\big(0, \tfrac{\sigma_i^2}{d} I_d\big),$$

now with the centre itself random, $x_c \sim \mathcal{N}(x^0, \frac{\rho^2}{d} I_d)$, independent of everything else. For an algorithm A with output $\widehat{x}$ write $\mathfrak{B}_\rho(A) := \mathbb{E}[f_0(\widehat{x}) - f_0^\star]$ for its Bayes excess risk, the expectation running over x_c, the tilts, the noises and A's randomness. Step 4 of the proof of Theorem 4.5 establishes the exact finite-ρ floor *before* the limit $\rho \to \infty$ is taken. Every algorithm making at most K queries per client satisfies

$$\mathfrak{B}_\rho(A) \;\ge\; \mathfrak{B}_\rho^\star(K) := \frac{1}{2\mu}\Big(\frac{1}{\mu^2\rho^2} + \frac{1}{\Psi(K)}\Big)^{-1}. \tag{7}$$

The algorithm to be compared is the two-phase CROWD of Theorem 4.4, made uniform in the centre by a projection whose radius the prior supplies. Let CROWD* denote Algorithm 1 run on the ball $\mathcal{X} = B(x^0, D)$ of radius $D := 2\rho + \sqrt{2}\,\zeta_{\max}/\mu$, $\zeta_{\max} := \max_i \zeta_i$, with the projection $\Pi_{\mathcal{X}}$ applied in *every* round, probing included. Started at x^0, it runs K_0 probing rounds (K_0 even) with uniform weights, all n clients reporting in every round. It is then refreshed once by Algorithm 2 with horizon $K = T$, noise bounds $\bar{\sigma}_i = \sigma_i$ and floor $\varrho > 0$. Finally it runs T rounds of projected SGD on $F_{\widehat{w}}$ with the tuned stepsize of Lemma F.6 and $R = 2D$. Theorem 4.5 grants every algorithm μ, the σ_i and

T; besides these, CROWD* uses the prior scale ρ and $\zeta_{\max}$ through D. It is therefore a prior-aware variant of Algorithm 1, not Algorithm 1 itself.

Theorem C.11 (The loop closed)**.** *Consider* CROWD* *as just defined, on the Gaussian class with linear tilts. Fix $\delta \in (0,1)$ and let $\epsilon_0 \le \frac{1}{2}$, $\epsilon_1 \le 1$ be the quantities of Lemma D.4 at confidence δ, with $\epsilon_2 := (\epsilon_1+\epsilon_0)/(1-\epsilon_0) < 1$. Put $\bar V := \max_i \sigma_i^2/b$, $\ell_T := 1+\ln(e+4\mu^2D^2T/\bar V)$ (read $\ell_T\bar V$ as 0 when $\bar V = 0$), $s_i := (1-\frac{2}{n})\zeta_i^2 + \mathrm{AM}/n$, $\vartheta_i := \tau_i\sqrt{2/(bK_0)}$ and*

$$a(K_0) := \sum_{i=1}^n \big(2\sqrt{s_i}\,\vartheta_i + \sqrt{2}\,\vartheta_i^2\big).$$

Then for every $\epsilon_T > 0$,

$$\mathfrak{B}_\rho(\text{CROWD}^\star) \;\le\; \frac{(1+\epsilon_T)(1+\epsilon_2)(1+\epsilon_0)}{2\mu}\Big[\Psi(bT) + 2C_n a(K_0) + 6C_n\epsilon_2\mathrm{AM} + 2\varrho\Big] + (1+\epsilon_T^{-1})\,\frac{\mu}{2}\Big(4D^2e^{-T/2} + \frac{\ell_T\bar V}{\mu^2T}\Big) \;+\; 2\mu\big(D^2+\rho^2\big)\big(\delta + e^{-d/8}\big).$$

Proved in Appendix G.7.

Every term is an unconditional expectation, and each excess factor is priced. The crowd-realisation risk is paid by the tail term $2\mu(D^2+\rho^2)(\delta+e^{-d/8})$ rather than assumed away. The incoherence floor appears as $6C_n\epsilon_2\mathrm{AM}$ with ϵ_2 supplied by Lemma D.4, and the estimation error as the explicit $a(K_0)$. Uniformity in the centre is bought from the prior itself, through the projection radius D. Matching (7) is then a matter of sending the parameters to their limits in the right order, the prior scale ρ last.

Corollary C.12 (Unconditional matching in the diffuse-prior limit)**.** *Run* CROWD* *with $\varrho = 1/T$, $\epsilon_T = T^{-1/4}$ and $\delta = e^{-\sqrt{d}}$, and let $K = b(K_0+T)$ be the per-client budget shared by both sides. At every fixed prior scale ρ,*

$$\limsup_{d\to\infty}\ \limsup_{K_0\to\infty}\ \limsup_{T\to\infty}\ \frac{\mathfrak{B}_\rho(\text{CROWD}^\star)}{\mathfrak{B}^\star_\rho(b(K_0+T))} \;\le\; 1 + \frac{\Psi_0}{\mu^2\rho^2},$$

and the ratio is at least one at every parameter value by (7). Hence the ratio tends to one as the prior becomes diffuse, $\rho\to\infty$: for every $\epsilon \in (0,1]$ there is $\rho_0(\epsilon)$ and, for each $\rho \ge \rho_0$, thresholds d_0, K_0', T_0 such that $d \ge d_0$, $K_0 \ge K_0'$ and $T \ge T_0$ give $\mathfrak{B}_\rho(\text{CROWD}^\star) \le (1+\epsilon)\,\mathfrak{B}^\star_\rho(K)$.

Proved in Appendix G.7.

Four remarks on scope. First, at a fixed prior scale the two sides do not match. The gap $\Psi_0/(\mu^2\rho^2)$ is the information in the prior: the Bayes floor uses it, and CROWD*, which does not shrink towards x^0, does not. The matching is therefore a statement about the diffuse-prior limit, the regime of the minimax bound of Theorem 4.5. Second, what is matched is the bias part $\Psi_0/(2\mu)$ and the constant $1/(2\mu)$. At finite K_0 the upper side carries the estimation term $2C_na(K_0)$, of order $K_0^{-1/2}$, whereas $\Psi(bT)-\Psi_0$ is of order $1/T$. The finite-sample exchange rate σ_i^2/K inside $\Psi(K)$ is therefore not certified by this comparison. Third, the order of limits is forced. The horizon must grow at fixed prior scale, since a fixed number of steps cannot serve an arbitrarily diffuse prior, and $d/n\to\infty$ drives the incoherence floor to zero, as in Corollary 4.6. The finite-sample terms are explicit in the theorem. Fourth, the statement is about the Gaussian class of Theorem 4.5, the class the lower bound is proved for. For a single deployment the conditional, per-instance reading of Corollary 4.6 remains the operative statement.

D PROOFS FOR SECTION 2

Throughout, $\mathbb{E}[\cdot \mid B]$ denotes expectation conditional on the realised tilts. Recall $H_w = \sum_i w_ih_i$, $F_w = f_0 + H_w$, $b(w) = \nabla H_w(x_0^\star) = B^\top w$, $\Gamma_w = \|b(w)\|$, $V(w) = \sum_i w_i^2 q_i$.

D.1 RIGIDITY OF A MEAN-ZERO CONVEX CROWD

Proposition D.1 (Rigidity)**.** *Let $h_1,\dots,h_n$ satisfy Assumption A1 on an open convex set. If each h_i is almost surely convex, then almost surely every h_i is affine.*

The proof works with gradients only, so it needs no integrability of h_i itself: Assumption A1 already presupposes that $\nabla h_i(x)$ is integrable for every x. The idea is that convexity makes a certain gradient pairing non-negative while the centring makes its mean zero, which forces it to vanish.

Proof. Let U be the open convex set, fix i, and write $h = h_i$.

Step 1: the monotonicity defect vanishes pointwise. For $x, y \in U$ put $M(x,y) := \langle \nabla h(x) - \nabla h(y),\, x - y\rangle$. A differentiable convex function has a monotone gradient, so $M(x,y) \geq 0$ almost surely. By linearity of expectation and Assumption A1, $\mathbb{E}M(x,y) = \langle \mathbb{E}\nabla h(x) - \mathbb{E}\nabla h(y),\, x - y\rangle = 0$. A non-negative random variable with zero mean vanishes almost surely, so $M(x,y) = 0$ almost surely, for each fixed pair.

Step 2: simultaneously for all pairs. Let $U_0 \subseteq U$ be countable and dense. By countable additivity there is an event of probability one on which $M(x,y) = 0$ for all $x, y \in U_0$. On that event, continuity of ∇h ($h \in C^2$ by Assumption A1; C^1 would suffice) makes M continuous on $U \times U$, so $M(x,y) = 0$ for all $x, y \in U$.

Step 3: a vanishing defect forces affinity. Fix $x_0 \in U$ and $x \in U$, and put $g(t) := h(x_0 + t(x - x_0))$ for $t \in [0,1]$. Then $g'(t) - g'(0) = \langle \nabla h(x_0 + t(x - x_0)) - \nabla h(x_0),\, x - x_0\rangle = M(x_0 + t(x - x_0), x_0)/t = 0$ for $t \in (0,1]$. Hence g is affine and $h(x) = h(x_0) + \langle \nabla h(x_0),\, x - x_0\rangle$ for every $x \in U$: h is affine on U. Since i was arbitrary and there are finitely many clients, the conclusion holds for all of them on one event of probability one. □

The same argument with a deterministic crowd shows: if $h_1, \ldots, h_n$ are differentiable and convex and $\sum_i \lambda_i \nabla h_i$ is constant for some $\lambda_i > 0$, then every h_i is affine, since $\sum_i \lambda_i M_i(x,y) = 0$ with $M_i \geq 0$ forces each $M_i(x,y) = 0$.

D.2 The bias-transfer lemma

The claims quoted in Section 2 are the linear-tilt cases of (ii) and (iii) below, together with the displacement bound of (i).

Lemma D.2 (Bias transfer). *Let Assumptions A1–A2 hold, $w \in \Delta_n$, $\mathcal{X} = \mathbb{R}^d$ and $\lambda_w < \mu$, and write $\Gamma_w := \|b(w)\|$.*

(i) F_w is $(\mu - \lambda_w)$-strongly convex and $(L + \lambda_w)$-smooth, and its minimiser obeys $\|x_w^\star - x_0^\star\| \leq \Gamma_w/(\mu - \lambda_w)$.

(ii) For every x, the excess risk on the truth is at most $\frac{2L}{\mu - \lambda_w}[F_w(x) - \min F_w] + \frac{L\Gamma_w^2}{(\mu - \lambda_w)^2}$, and in the linear-tilt case the bias constant can be made sharp: for every $\epsilon > 0$,

$$f_0(x) - f_0^\star \;\leq\; (1 + \epsilon)\,\frac{\Gamma_w^2}{2\mu} \;+\; \Big(1 + \tfrac{1}{\epsilon}\Big)\frac{L}{\mu}\big[F_w(x) - \min F_w\big].$$

(iii) If $\lambda_w < \mu/2$ then $f_0(x_w^\star) - f_0^\star \geq \frac{1}{2L}(\frac{\mu - 2\lambda_w}{\mu - \lambda_w})^2 \Gamma_w^2$, which is $\Gamma_w^2/(2L)$ in the linear case.

Proof. (i) $\nabla^2 F_w = \nabla^2 f_0 + \nabla^2 H_w$ and $\|\nabla^2 H_w\|_{\mathrm{op}} \leq \lambda_w$ give the moduli. For the displacement, $\nabla F_w(x_0^\star) = \nabla f_0(x_0^\star) + \nabla H_w(x_0^\star) = b(w)$, so by $(\mu - \lambda_w)$-strong convexity of F_w, $\|x_0^\star - x_w^\star\| \leq \|\nabla F_w(x_0^\star)\|/(\mu - \lambda_w) = \Gamma_w/(\mu - \lambda_w)$.

(ii) By L-smoothness of f_0 and $\nabla f_0(x_0^\star) = 0$, $f_0(x) - f_0^\star \leq \frac{L}{2}\|x - x_0^\star\|^2 \leq L\|x - x_w^\star\|^2 + L\|x_w^\star - x_0^\star\|^2$, and $\|x - x_w^\star\|^2 \leq \frac{2}{\mu - \lambda_w}[F_w(x) - \min F_w]$ by strong convexity. Combining with (i) gives the claim.

For the linear case ($\lambda_w = 0$, $H_w = \langle b, \cdot\rangle$) we argue through the gradient instead, which is what makes the constant sharp. Strong convexity gives the Polyak–Łojasiewicz inequality $f_0(x) - f_0^\star \leq \|\nabla f_0(x)\|^2/(2\mu)$, and $\nabla f_0(x) = \nabla F_w(x) - b(w)$, so by Young's inequality, for any $\epsilon > 0$,

$$\|\nabla f_0(x)\|^2 \leq (1 + \epsilon)\Gamma_w^2 + \Big(1 + \tfrac{1}{\epsilon}\Big)\|\nabla F_w(x)\|^2.$$

Smoothness of F_w (constant L, since $\lambda_w = 0$) gives $\|\nabla F_w(x)\|^2 \le 2L[F_w(x) - \min F_w]$, and dividing by 2μ yields the claim.

(iii) $\nabla f_0(x_w^\star) = -\nabla H_w(x_w^\star)$ and, by λ_w-Lipschitz continuity of ∇H_w together with (i),

$$\|\nabla H_w(x_w^\star)\| \;\ge\; \Gamma_w - \lambda_w\|x_w^\star - x_0^\star\| \;\ge\; \Gamma_w\Big(1 - \frac{\lambda_w}{\mu - \lambda_w}\Big) = \Gamma_w\frac{\mu - 2\lambda_w}{\mu - \lambda_w}.$$

For convex L-smooth f_0 the descent lemma at $y^+ = y - \nabla f_0(y)/L$ gives $f_0(y) - f_0^\star \ge \|\nabla f_0(y)\|^2/(2L)$; apply at $y = x_w^\star$. □

D.3 The diagonal proxy

The next two lemmas are used wherever the bias energy is replaced by its diagonal surrogate $V(w)$, and wherever the incoherence of Assumption A4 is bounded for a Gaussian crowd.

Lemma D.3. *Under Assumption A4,* $(1-\kappa)V(w) \le \|b(w)\|^2 \le (1+\kappa)V(w)$ *for every* $w \in \mathbb{R}^n$.

Proof. $\|b(w)\|^2 = w^\top Q w = V(w) + w^\top(Q - D)w$ and, with $z = D^{1/2}w$, $|w^\top(Q-D)w| = |z^\top D^{-1/2}(Q-D)D^{-1/2}z| \le \kappa\|z\|^2 = \kappa V(w)$. □

D.4 Incoherence of a Gaussian crowd

Lemma D.4. *Let* $\beta_i \sim \mathcal{N}(0, \frac{\zeta_i^2}{d} I_d)$ *independently,* $\delta \in (0,1)$, *and put*

$$\epsilon_0 := 2\sqrt{\frac{\log(4n/\delta)}{d}} + \frac{2\log(4n/\delta)}{d}, \qquad \epsilon_1 := C\Big(\sqrt{\frac{n}{d}} + \sqrt{\frac{\log(4/\delta)}{d}}\Big),$$

with $C := K^2 C_V$, *where* C_V *is the absolute constant of Vershynin (2018, Thm. 4.6.1), taken to be at least one (enlarging it only weakens that theorem's bound, and it is needed so that* ϵ_1/K^2 *dominates the theorem's* $C_V\sqrt{n/d} + t/\sqrt{d}$*), and* $K \ge 1$ *the sub-Gaussian norm of a standard Gaussian vector. If* $\epsilon_0 \le \frac{1}{2}$ *and* $\epsilon_1 \le 1$, *then with probability at least* $1-\delta$, $\max_i |q_i/\zeta_i^2 - 1| \le \epsilon_0$ *and* $\kappa \le (\epsilon_1 + \epsilon_0)/(1-\epsilon_0)$. *Both conditions hold when* $d \ge C'(n + \log(4n/\delta))$ *for an absolute constant* C'.

Proof. Write $\beta_i = \zeta_i u_i$ with $u_i \sim \mathcal{N}(0, I_d/d)$ independent, so that $q_i/\zeta_i^2 = \|u_i\|^2$ and $d\|u_i\|^2 \sim \chi^2_d$.

Norms. The Laurent–Massart inequality (Laurent & Massart, 2000) gives $\mathbb{P}(|\chi^2_d/d - 1| > 2\sqrt{t/d} + 2t/d) \le 2e^{-t}$. With $t = \log(4n/\delta)$ and a union bound over i, the event $E_1 = \{\max_i |\|u_i\|^2 - 1| \le \epsilon_0\}$ has probability at least $1 - \delta/2$.

Gram matrix. Let $Z \in \mathbb{R}^{n\times d}$ have the rows $\sqrt{d}\,u_i$, so Z has independent standard normal entries and $M_0 := \frac{1}{d}ZZ^\top$ has entries $\langle u_i, u_j\rangle$. The rows of $Z^\top$ are independent $\mathcal{N}(0, I_n)$ vectors, so Vershynin (2018, Thm. 4.6.1) with $t = \sqrt{\log(4/\delta)}$ gives, with probability at least $1 - 2e^{-t^2} = 1 - \delta/2$, $\|M_0 - I_n\|_{\mathrm{op}} \le K^2\max(\epsilon_1', \epsilon_1'^2)$ for $\epsilon_1' := \epsilon_1/K^2$. Since $K \ge 1$, this is at most $\max(\epsilon_1, \epsilon_1^2) = \epsilon_1$, using $\epsilon_1 \le 1$; call this event E_2.

Normalisation. Since $Q_{ij} = \zeta_i\zeta_j\langle u_i, u_j\rangle$ and $D_{ii} = \zeta_i^2\|u_i\|^2$, we have $D^{-1/2}QD^{-1/2} = \Lambda M_0 \Lambda$ with $\Lambda := \mathrm{diag}(\|u_i\|^{-1})$, whose diagonal is one. Hence

$$D^{-1/2}(Q-D)D^{-1/2} = \Lambda M_0\Lambda - I = \Lambda(M_0 - I)\Lambda + (\Lambda^2 - I).$$

On E_1, $\|\Lambda\|_{\mathrm{op}}^2 = \max_i\|u_i\|^{-2} \le 1/(1-\epsilon_0)$ and $\|\Lambda^2 - I\|_{\mathrm{op}} = \max_i|1 - \|u_i\|^2|/\|u_i\|^2 \le \epsilon_0/(1-\epsilon_0)$. On $E_1 \cap E_2$, which has probability at least $1-\delta$, the triangle inequality gives $\kappa \le (\epsilon_1 + \epsilon_0)/(1-\epsilon_0)$.

Dimension. Let $C' := \max(64, 4C^2)$ and $d \ge C'(n + \log(4n/\delta))$. Then $\log(4n/\delta)/d \le \frac{1}{64}$, so $\epsilon_0 \le \frac{2}{8} + \frac{2}{64} < \frac{1}{2}$; and $n/d \le \frac{1}{4C^2}$, $\log(4/\delta)/d \le \frac{1}{4C^2}$, so $\epsilon_1 \le C(\frac{1}{2C} + \frac{1}{2C}) = 1$. □

D.5 THE ORACLE WEIGHTS

Proposition D.5 (Exact bias identity, oracle weights, and the gain)**.** *Under Assumption A1, for every $w \in \Delta_n$ independent of the tilts, $\mathbb{E}\Gamma_w^2 = \sum_i w_i^2 \zeta_i^2$. For any $c \in \mathbb{R}^n_{>0}$ the minimum of $\sum_i w_i^2 c_i$ over Δ_n, and also over the affine hull $\mathbf{1}^\top w = 1$, equals $(\sum_i c_i^{-1})^{-1}$ and is attained at $w_i^\star \propto c_i^{-1} > 0$. With $c_i = \zeta_i^2$ this gives $\Psi_0 = \mathrm{HM}/n$ against AM/n for uniform weights. Moreover* (i) *the ratio is $\mathrm{AM}/\mathrm{HM} \geq 1$, with equality iff all ζ_i are equal;* (ii) *$\Psi_0 \leq \min_i \zeta_i^2$;* (iii) *the ratio is unbounded.*

Proof. Identity. $\mathbb{E}\|\sum_i w_i \beta_i\|^2 = \sum_{i,j} w_i w_j \mathbb{E}\langle \beta_i, \beta_j \rangle$; for $i \neq j$ independence and $\mathbb{E}\beta_i = \mathbb{E}\nabla h_i(x_0^\star) = 0$ give $\mathbb{E}\langle \beta_i, \beta_j \rangle = \langle \mathbb{E}\beta_i, \mathbb{E}\beta_j \rangle = 0$, leaving $\sum_i w_i^2 \zeta_i^2$.

Optimum. For $\mathbf{1}^\top w = 1$, Cauchy–Schwarz gives $1 = (\sum_i w_i \sqrt{c_i} \cdot c_i^{-1/2})^2 \leq (\sum_i w_i^2 c_i)(\sum_i c_i^{-1})$, so $\sum_i w_i^2 c_i \geq (\sum_i c_i^{-1})^{-1}$ with equality iff $w_i \propto c_i^{-1}$. That candidate is strictly positive, hence lies in the relative interior of Δ_n: the affine and simplex-constrained problems have the same solution and value. Strict convexity gives uniqueness.

(i) The ratio of the two values is $(\mathrm{AM}/n)/(\mathrm{HM}/n) = \mathrm{AM}/\mathrm{HM} \geq 1$ by the inequality between the arithmetic and harmonic means, with equality iff all ζ_i^2 coincide. *(ii)* $\Psi_0 = (\sum_j \zeta_j^{-2})^{-1}$ and $\sum_j \zeta_j^{-2} \geq \zeta_{\min}^{-2}$ give $\Psi_0 \leq \zeta_{\min}^2$. *(iii)* Take $\zeta_1^2 = R$ and $\zeta_j^2 = 1$ for $j \geq 2$. Then $\mathrm{AM} = (R+n-1)/n$ and $\mathrm{HM} = n/(R^{-1}+n-1)$, so $\mathrm{AM}/\mathrm{HM} = (R+n-1)(R^{-1}+n-1)/n^2 \geq (n-1)R/n^2 \to \infty$ as $R \to \infty$. □

With $n = 21$ and $R = 100$, uniform weighting gives $\mathrm{AM}/n = 120/441 = 0.272$, the value of a uniform crowd of 3.7 unit-energy members: one miscalibrated site among twenty unit-energy ones makes the uniform crowd behave like a crowd of fewer than four.

D.6 THE TRANSFER LEMMA ON A CONSTRAINED DOMAIN

Theorem C.7 runs on a bounded domain, where the minimiser of F_w may lie on the boundary. Parts (i) and the general bound of (ii) survive.

Lemma D.6 (Constrained bias transfer)**.** *Let Assumptions A1–A2 hold on a closed convex $\mathcal{X}$ whose interior contains $x_0^\star$, let $w \in \Delta_n$ with $\lambda_w < \mu$, and let $x_w^\star$ minimise F_w over $\mathcal{X}$. Then $\|x_w^\star - x_0^\star\| \leq \Gamma_w/(\mu - \lambda_w)$, and for every $x \in \mathcal{X}$, $f_0(x) - f_0^\star \leq \frac{2L}{\mu - \lambda_w}[F_w(x) - F_w(x_w^\star)] + \frac{L\Gamma_w^2}{(\mu-\lambda_w)^2}$.*

Proof. Put $m := \mu - \lambda_w$; F_w is m-strongly convex on $\mathcal{X}$ as in Lemma D.2(i). First-order optimality of $x_w^\star$ over $\mathcal{X}$ gives $\langle \nabla F_w(x_w^\star), y - x_w^\star \rangle \geq 0$ for every $y \in \mathcal{X}$.

Displacement. Strong monotonicity of ∇F_w and optimality with $y = x_0^\star$ give

$$m\|x_0^\star - x_w^\star\|^2 \leq \langle \nabla F_w(x_0^\star) - \nabla F_w(x_w^\star), x_0^\star - x_w^\star \rangle \leq \langle \nabla F_w(x_0^\star), x_0^\star - x_w^\star \rangle \leq \Gamma_w \|x_0^\star - x_w^\star\|,$$

since $\nabla F_w(x_0^\star) = \nabla f_0(x_0^\star) + b(w) = b(w)$, the truth's minimiser being interior.

Excess risk. Strong convexity and optimality give $F_w(x) - F_w(x_w^\star) \geq \langle \nabla F_w(x_w^\star), x - x_w^\star \rangle + \frac{m}{2}\|x - x_w^\star\|^2 \geq \frac{m}{2}\|x - x_w^\star\|^2$. Since $\nabla f_0(x_0^\star) = 0$ and f_0 is L-smooth on the segment $[x_0^\star, x] \subseteq \mathcal{X}$, $f_0(x) - f_0^\star \leq \frac{L}{2}\|x - x_0^\star\|^2 \leq L\|x - x_w^\star\|^2 + L\|x_w^\star - x_0^\star\|^2$, and the two displays combine to the claim. □

D.7 HARMONIC SUPERADDITIVITY

The lower bound of Section 4.3 splits a harmonic value into a bias part and a sampling part, and needs the following.

Lemma D.7 (Harmonic superadditivity)**.** *$H(c) := (\sum_i c_i^{-1})^{-1}$ is concave and positively homogeneous of degree one on $\mathbb{R}^n_{>0}$; hence $H(a+b) \geq H(a) + H(b)$, with equality if and only if b is proportional to a.*

Proof. By Proposition D.5, $H(c) = \min_{\mathbf{1}^\top w = 1} \phi_c(w)$ with $\phi_c(w) := \sum_i w_i^2 c_i$, and the minimiser $w^c \propto c^{-1}$ is unique. As a pointwise minimum of functions linear in c, H is concave, and $H(tc) =$

$tH(c)$ for $t > 0$ is immediate. For superadditivity and its equality case, write $w := w^{a+b}$ and use $\phi_{a+b} = \phi_a + \phi_b$:

$$H(a+b) = \phi_a(w) + \phi_b(w) \;\geq\; H(a) + H(b),$$

with equality iff $\phi_a(w) = H(a)$ and $\phi_b(w) = H(b)$, that is, by uniqueness of the minimisers, iff $w^{a+b} = w^a = w^b$. Since w^c is the normalisation of $(c_1^{-1}, \dots, c_n^{-1})$, $w^a = w^b$ holds iff $a^{-1} \propto b^{-1}$, i.e. iff $b = ta$ for some $t > 0$; and then $w^{a+b} = w^a$ as well. □

The defect can be large. With $n = 2$, $a = (1, 10)$ and $b = (10, 1)$ one has $H(a+b) = 5.50$ against $H(a) + H(b) = 1.82$, so the two sides differ by a factor three; they coincide when the bias and noise profiles are proportional.

E Proofs for Section 3

This section proves the two impossibilities of Proposition 3.1 and establishes the property of the residual (3) that the algorithm rests on.

E.1 Proof of Proposition 3.1(i)

For $c \in \mathbb{R}^d$ let I_c denote the instance with truth $f_0 + \langle c, \cdot \rangle$ and perturbations $h_i - \langle c, \cdot \rangle$, the same noise laws, and the same realised crowd otherwise; I_0 is the original instance.

Proof. Client risks. For every i and x, $\big(f_0 + \langle c, \cdot \rangle\big)(x) + \big(h_i - \langle c, \cdot \rangle\big)(x) = f_0(x) + h_i(x) = f_i(x)$, so I_c and I_0 have the same tuple $(f_1, \dots, f_n)$.

Transcript. A protocol's transcript is produced from (a) the algorithm's private randomness U, independent of the instance, (b) the reports, whose conditional law given the past depends on the instance only through the queried client's ∇f_i at the queried point, by Assumption A3, and (c) measurable maps of the two. We show by induction on the round k that the joint law of (U, reports up to k) is the same under I_c and I_0. For $k = 0$ only U is present. If the claim holds up to $k-1$, the round-k query points are the same measurable function of that history under both instances, and the conditional law of the round-k reports given the history is the same because the ∇f_i agree; hence the joint law up to k agrees.

Targets and biases. On $\mathcal{X} = \mathbb{R}^d$ the minimiser of $f_0 + \langle c, \cdot \rangle$ solves $\nabla f_0(x) = -c$, which no solution of $\nabla f_0(x) = 0$ satisfies when $c \neq 0$; so the targets differ. For linear tilts the tilts of I_c at its own minimiser are $\beta_i - c$, and $\sum_i w_i(\beta_i - c) = b(w) - c$ because $\sum_i w_i = 1$. □

The statement is about the realised crowd, which is what a protocol faces. The shifted instance stays inside the model class. For Gaussian linear tilts every realisation $(\beta_i - c)_i$ lies in the support of the centred law, so I_c is itself a possible realisation of an instance that satisfies Assumption A1. At fixed n no amount of data selects between the two.

E.2 Proof of Proposition 3.1(ii)

The deviation of client i from the crowd mean at x is $\widetilde{\nabla h_i}(x) := \nabla h_i(x) - \frac{1}{n}\sum_j \nabla h_j(x)$, and the corrected report is $\widehat{g}_i^{\,\mathrm{corr}} := \widehat{g}_i - \widetilde{\nabla h_i}(x)$.

Proof. By Assumption A3, $\mathbb{E}[\widehat{g}_i \mid x] = \nabla f_0(x) + \nabla h_i(x)$, so $\mathbb{E}[\widehat{g}_i^{\,\mathrm{corr}} \mid x] = \nabla f_0(x) + \frac{1}{n}\sum_j \nabla h_j(x) = \nabla F_{w_{\mathrm{unif}}}(x)$ for every i. For any $w \in \Delta_n$ the aggregate therefore has conditional mean $\sum_i w_i \nabla F_{w_{\mathrm{unif}}}(x) = \nabla F_{w_{\mathrm{unif}}}(x)$: it is an unbiased gradient of $F_{w_{\mathrm{unif}}}$, whatever w is. Its bias vector at $x_0^\star$ is $\bar{\beta} = b(w_{\mathrm{unif}})$, whose expected energy is AM/n by Proposition D.5 with uniform weights, and $\mathrm{AM}/n > \mathrm{HM}/n = \Psi_0$ unless all ζ_i coincide. □

E.3 The residual is a maximal invariant

Lemma E.1 (Residual representation; maximal invariant)**.** *Under Assumptions A1 and A3, the residuals $r_i^k = \widehat{g}_i^k - \frac{1}{n}\sum_j \widehat{g}_j^k = \widetilde{\nabla h_i}(x^{k-1}) + \widetilde{\varepsilon}_i^k$ do not depend on $\nabla f_0(x^{k-1})$, whatever the query point and whatever f_0 is; the array (r_i^k) is a maximal invariant of $(\widehat{g}_i^k)$ under per-round common shifts; and $\mathbb{E}[\widetilde{\varepsilon}_i^k \mid \mathcal{F}_{k-1}] = 0$, $\mathbb{E}[\|\widetilde{\varepsilon}_i^k\|^2 \mid \mathcal{F}_{k-1}] \le \tau_i^2/b$.*

Proof. Representation. By Assumption A3, $\widehat{g}_i^k = \nabla f_0(x^{k-1}) + \nabla h_i(x^{k-1}) + \varepsilon_i^k$. Averaging over i and subtracting gives $r_i^k = \nabla h_i(x^{k-1}) - \frac{1}{n}\sum_j \nabla h_j(x^{k-1}) + \varepsilon_i^k - \bar{\varepsilon}^k = \widetilde{\nabla h_i}(x^{k-1}) + \widetilde{\varepsilon}_i^k$, with $\widetilde{\varepsilon}_i^k := \varepsilon_i^k - \bar{\varepsilon}^k$, in which $\nabla f_0(x^{k-1})$ has cancelled identically; no property of x^{k-1} or of f_0 was used. For linear tilts $\widetilde{\nabla h_i} \equiv \widetilde{\beta}_i$.

Maximal invariance. If $\widehat{g}_i'^k = \widehat{g}_i^k + c^k$ for all i, k then the row means shift by c^k and the residuals are unchanged. Conversely, if two arrays have the same residuals then $\widehat{g}_i'^k - \widehat{g}_i^k = \bar{\widehat{g}}'^k - \bar{\widehat{g}}^k =: c^k$ does not depend on i. Hence the residual array separates exactly the orbits of the group, which is the definition of a maximal invariant (Lehmann & Romano, 2005, Ch. 6).

Moments. $\widetilde{\varepsilon}_i^k = (1 - \frac{1}{n})\varepsilon_i^k - \frac{1}{n}\sum_{j\neq i}\varepsilon_j^k$ has conditional mean zero. By conditional independence across clients and zero conditional means, the cross terms vanish and

$$\begin{aligned}\mathbb{E}\big[\|\widetilde{\varepsilon}_i^k\|^2 \mid \mathcal{F}_{k-1}\big] &= \Big(1 - \tfrac{1}{n}\Big)^2 \mathbb{E}\big[\|\varepsilon_i^k\|^2 \mid \mathcal{F}_{k-1}\big] + \tfrac{1}{n^2}\sum_{j\neq i}\mathbb{E}\big[\|\varepsilon_j^k\|^2 \mid \mathcal{F}_{k-1}\big] \\ &\le \frac{\sigma_i^2}{b} + \frac{1}{n^2}\sum_j \frac{\sigma_j^2}{b} = \frac{\tau_i^2}{b}.\end{aligned}$$

□

F Proofs for Section 4

The subsections below follow the order of Section 4. Identification is treated first, then the estimation error and the end-to-end bound, then the lower bound, then the non-linear crowd.

F.1 Proof of Theorem 4.1

The proof runs in five steps. Step 1 computes the expected disagreement energy of one client, which is where independence removes the cross terms. Step 2 sums that identity over clients to recover the total energy, and so inverts the map in closed form. Step 3 reads the ℓ_∞ condition number off the rows of the inverse. Step 4 checks that the quantity being inverted is a functional of the observable law. Step 5 treats the degenerate case $n = 2$.

Proof. Step 1: the forward map. By Assumption A1, $\mathbb{E}\langle\beta_i, \beta_j\rangle = 0$ for $i \neq j$ and $\mathbb{E}\|\beta_i\|^2 = \zeta_i^2$. Hence, with $Z := \sum_j \zeta_j^2$,

$$\mathbb{E}\langle\beta_i, \bar{\beta}\rangle = \frac{1}{n}\Big(\zeta_i^2 + \sum_{j\neq i}\mathbb{E}\langle\beta_i, \beta_j\rangle\Big) = \frac{\zeta_i^2}{n}, \qquad \mathbb{E}\|\bar{\beta}\|^2 = \frac{1}{n^2}\sum_j \zeta_j^2 = \frac{Z}{n^2},$$

so that

$$s_i = \mathbb{E}\|\beta_i - \bar{\beta}\|^2 = \zeta_i^2 - \frac{2\zeta_i^2}{n} + \frac{Z}{n^2} = \Big(1 - \frac{2}{n}\Big)\zeta_i^2 + \frac{1}{n^2}\sum_j \zeta_j^2 = (T\zeta^2)_i.$$

Step 2: inversion. Summing over i, $\sum_i s_i = (1 - \frac{2}{n})Z + \frac{Z}{n} = Z\frac{n-1}{n}$, so $Z = \frac{n}{n-1}\sum_i s_i$, and therefore

$$\zeta_i^2 = \frac{n}{n-2}\Big(s_i - \frac{Z}{n^2}\Big) = \frac{n}{n-2}\Big(s_i - \frac{1}{n(n-1)}\sum_j s_j\Big),$$

which is well defined for $n \ge 3$. This exhibits a two-sided inverse of T, hence T is a bijection.

Step 3: conditioning. The ith row of T^{-1} has entries $\frac{n}{n-2}\big(\delta_{ij} - \frac{1}{n(n-1)}\big)$, whose absolute values sum to

$$\frac{n}{n-2}\Big[\Big(1 - \frac{1}{n(n-1)}\Big) + \frac{n-1}{n(n-1)}\Big] = \frac{n}{n-2} \cdot \frac{n(n-1) - 1 + (n-1)}{n(n-1)} = \frac{n^2-2}{(n-1)(n-2)}.$$

Since $\|A\|_{\infty\to\infty}$ is the maximum absolute row sum and all rows are permutations of each other, $\|T^{-1}\|_{\infty\to\infty} = C_n$ as claimed. The numerical values follow, and $C_n = 1 + 3/n + O(n^{-2}) \to 1$.

Step 4: identifiability. For linear tilts Lemma E.1 gives $r_i^k = \widetilde{\beta}_i + \widetilde{\varepsilon}_i^k$ with $\widetilde{\beta}$ measurable with respect to $\mathcal{F}_0$ and $(\widetilde{\varepsilon}_i^k)_k$ a martingale difference sequence. For any two rounds $k < l$,

$$\mathbb{E}\langle r_i^k,\, r_i^l\rangle = \mathbb{E}\|\widetilde{\beta}_i\|^2 + \mathbb{E}\langle\widetilde{\beta}_i,\, \widetilde{\varepsilon}_i^k\rangle + \mathbb{E}\langle\widetilde{\beta}_i + \widetilde{\varepsilon}_i^k,\, \mathbb{E}[\widetilde{\varepsilon}_i^l \mid \mathcal{F}_{l-1}]\rangle = s_i,$$

because the middle term vanishes on conditioning on $\mathcal{F}_{k-1}$ and the last on conditioning on $\mathcal{F}_{l-1}$. So s_i is a functional of the joint law of the residuals of two rounds, whatever the query points and without knowledge of the noise level; $\zeta^2 = T^{-1}s$ is then identified, and so is $w_i^\star \propto \zeta_i^{-2}$.

Step 5: failure at $n = 2$. For $n = 2$, $(Tq)_i = 0 \cdot q_i + \frac{1}{4}(q_1 + q_2)$, so T has rank one and ζ_1^2, ζ_2^2 are not separately identified. □

F.2 The realised inversion error

Lemma F.1 (Realised inversion error)**.** *For* arbitrary *vectors $\beta_1, \dots, \beta_n$ with $q_i = \|\beta_i\|^2 > 0$ and incoherence κ, and with $\bar{q} = \frac{1}{n}\sum_j q_j$,*

$$\|\widetilde{\beta}_i\|^2 = (Tq)_i + \eta_i, \qquad |\eta_i| \le 3\kappa\bar{q} \quad \textit{for every } i,$$

so $\|T^{-1}(\|\widetilde{\beta}\|^2) - q\|_\infty \le 3C_n\kappa\bar{q}$. No distributional assumption is used.

Proof. Expanding as in the previous proof but without expectations,

$$\|\widetilde{\beta}_i\|^2 = q_i - \frac{2}{n}\Big(q_i + \sum_{j\neq i}\langle\beta_i,\, \beta_j\rangle\Big) + \frac{1}{n^2}\Big(\sum_j q_j + \sum_{j\neq l}\langle\beta_j,\, \beta_l\rangle\Big) = (Tq)_i + \eta_i,$$

with $\eta_i = -\frac{2}{n}\sum_{j\neq i}\langle\beta_i,\, \beta_j\rangle + \frac{1}{n^2}\sum_{j\neq l}\langle\beta_j,\, \beta_l\rangle$. Write $M := D^{-1/2}(Q - D)D^{-1/2}$, $\|M\|_{\mathrm{op}} \le \kappa$. Then

$$\Big|\sum_{j\neq i}\langle\beta_i,\, \beta_j\rangle\Big| = \big|e_i^\top(Q-D)\mathbf{1}\big| = \big|(D^{1/2}e_i)^\top M(D^{1/2}\mathbf{1})\big| \le \kappa\sqrt{q_i}\,\sqrt{\textstyle\sum_j q_j} = \kappa\sqrt{q_i}\,\sqrt{n\bar{q}},$$

and $\big|\sum_{j\neq l}\langle\beta_j,\, \beta_l\rangle\big| = |\mathbf{1}^\top(Q-D)\mathbf{1}| \le \kappa n\bar{q}$. Using $q_i \le n\bar{q}$,

$$|\eta_i| \le \frac{2}{n}\kappa\sqrt{q_i}\,\sqrt{n\bar{q}} + \frac{\kappa n\bar{q}}{n^2} \le \frac{2\kappa}{n}\cdot n\bar{q} + \frac{\kappa\bar{q}}{n} \le 3\kappa\bar{q}.$$

The final claim follows by applying T^{-1} and $\|T^{-1}\eta\|_\infty \le C_n\|\eta\|_\infty$. □

F.3 A Gaussian refinement of the inversion error

Lemma F.2 (Gaussian refinement of Lemma F.1)**.** *If $\beta_i \sim \mathcal{N}(0, \frac{\zeta_i^2}{d} I_d)$ independently and $d \ge C'(n + \log(12n/\delta))$ with C' as in Lemma D.4, then there is an absolute constant C such that with probability at least $1 - \delta$, $\max_i |\eta_i| \le C\,\bar{q}\sqrt{\log(6n/\delta)/d}$.*

Proof. Write $\bar{\zeta}^2 := \frac{1}{n}\sum_j \zeta_j^2$ and recall $\eta_i = -\frac{2}{n}X_i + \frac{1}{n^2}Y$ with $X_i := \langle\beta_i,\, \sum_{j\neq i}\beta_j\rangle$ and $Y := \sum_{j\neq l}\langle\beta_j,\, \beta_l\rangle$ from the proof of Lemma F.1. We use three events, each of probability at least $1 - \delta/3$.

Event E_0. Lemma D.4 with $\delta/3$ in place of δ gives $\max_i |q_i/\zeta_i^2 - 1| \le \frac{1}{2}$, hence $\frac{1}{2}\bar{\zeta}^2 \le \bar{q} \le \frac{3}{2}\bar{\zeta}^2$.

Event E_1: the linear terms. Conditionally on β_i, $\sum_{j\neq i}\beta_j$ is $\mathcal{N}(0, \frac{1}{d}\sum_{j\neq i}\zeta_j^2 I_d)$, so X_i is centred Gaussian with variance $q_i\sum_{j\neq i}\zeta_j^2/d \le q_i n\bar{\zeta}^2/d$. The Gaussian tail and a union bound over i

give, with probability at least $1-\delta/3$ and for all i, $|X_i| \le \sqrt{q_i n\bar{\zeta}^2/d}\,\sqrt{2\log(6n/\delta)}$. On E_0, $q_i \le \frac{3}{2}\zeta_i^2 \le \frac{3}{2}n\bar{\zeta}^2$, so $\frac{2}{n}|X_i| \le 2\sqrt{3/2}\,\bar{\zeta}^2\sqrt{2\log(6n/\delta)/d}$.

Event E_2: the quadratic term. Stack $g := (\sqrt{d}\beta_j/\zeta_j)_j$, a standard Gaussian vector in $\mathbb{R}^{nd}$. Then $Y = g^\top A g$ where A has the $d\times d$ blocks $A_{jl} = \frac{\zeta_j\zeta_l}{d}I_d$ for $j \ne l$ and $A_{jj} = 0$. Its norms are $\|A\|_F^2 = \frac{1}{d}\sum_{j\ne l}\zeta_j^2\zeta_l^2 \le (n\bar{\zeta}^2)^2/d$ and $\|A\|_{\mathrm{op}} \le \frac{1}{d}\|\zeta\zeta^\top - \mathrm{diag}(\zeta^2)\|_{\mathrm{op}} \le \frac{1}{d}\|\zeta\|^2 = n\bar{\zeta}^2/d$, since $-\mathrm{diag}(\zeta^2) \preceq \zeta\zeta^\top - \mathrm{diag}(\zeta^2) \preceq \zeta\zeta^\top$. Since $\mathbb{E}Y = \mathrm{tr}\,A = 0$, the Hanson–Wright inequality (Vershynin, 2018, Thm. 6.2.1) gives, with probability at least $1-\delta/3$, $|Y| \le C_{\mathrm{HW}} n\bar{\zeta}^2\big(\sqrt{\log(6/\delta)/d} + \log(6/\delta)/d\big)$, which is at most $2C_{\mathrm{HW}}n\bar{\zeta}^2\sqrt{\log(6/\delta)/d}$ because $\log(6/\delta) \le d$ under the dimension condition.

Conclusion. On $E_0\cap E_1\cap E_2$, which has probability at least $1-\delta$, $|\eta_i| \le \frac{2}{n}|X_i| + \frac{1}{n^2}|Y| \le C''\bar{\zeta}^2\sqrt{\log(6n/\delta)/d}$ for an absolute C'', and $\bar{\zeta}^2 \le 2\bar{q}$ on E_0. □

F.4 Proof of Lemma 4.2

We prove the following version, which contains Lemma 4.2; in (i), $8\widetilde{\zeta}_i^2\tau_i^2/(bK_0) + 8\tau_i^4/(bK_0)^2 = 4\widetilde{\zeta}_i^2\vartheta_i^2 + 2\vartheta_i^4$.

Lemma F.3 (Estimation error, full version of Lemma 4.2)**.** *Let Assumption A3 hold with arbitrary tilts, $n \ge 3$, K_0 even, and suppose the disagreement measured at the query points differs from $\widetilde{\beta}_i$ by at most δ_i almost surely, for all $k \le K_0$. Put $\vartheta_i := \tau_i\sqrt{2/(bK_0)}$ and $\widetilde{\zeta}_i := \|\widetilde{\beta}_i\|$.*

- *(i)* (Linear crowd, $\delta_i = 0$.) *$\widehat{s}_i$ is exactly unbiased for $\|\widetilde{\beta}_i\|^2$ given the tilts, for every adaptive query sequence, with conditional variance at most $8\widetilde{\zeta}_i^2\tau_i^2/(bK_0) + 8\tau_i^4/(bK_0)^2$; the bound is dimension-free.*
- *(ii)* (General.) $\mathbb{E}|\widehat{s}_i - \|\widetilde{\beta}_i\|^2| \le 2\widetilde{\zeta}_i(\delta_i+\vartheta_i) + (\delta_i+\vartheta_i)^2$.
- *(iii)* (Weights.) *With $\widehat{q} = (T^{-1}\widehat{s})_+$ and Assumption A4, $\Delta := \max_i|\widehat{q}_i - q_i| \le C_n(\max_i|\widehat{s}_i - \|\widetilde{\beta}_i\|^2| + 3\kappa\bar{q})$, which converts into a high-probability bound by Chebyshev and a union bound, improving to a $\sqrt{\log(dn/\delta)}$ dependence under sub-Gaussian noise (Remark F.5).*

Notation for the proof. Let $\mathcal{F}_k$ be the history through round k, including the tilts and the private randomness, so that $\widetilde{\beta}$ is $\mathcal{F}_0$-measurable. By Lemma E.1, which uses only Assumption A3 and the decomposition $f_i = f_0 + h_i$, the $\widetilde{\varepsilon}_i^k$ form a martingale difference sequence with $\mathbb{E}[\|\widetilde{\varepsilon}_i^k\|^2 \mid \mathcal{F}_{k-1}] \le \tau_i^2/b$. *Adaptive query points are allowed throughout.* The halves $\mathcal{K}_1, \mathcal{K}_2$ are the odd and the even probing rounds, each of size $K_0/2$. Write

$$\nu_i^{(a)} := \frac{2}{K_0}\sum_{k\in\mathcal{K}_a}\widetilde{\varepsilon}_i^k, \qquad u_i^{(a)} := \frac{2}{K_0}\sum_{k\in\mathcal{K}_a}\big(\widetilde{\nabla h_i}(x^{k-1}) - \widetilde{\beta}_i\big), \qquad m_i^{(a)} := u_i^{(a)} + \nu_i^{(a)},$$

so that $\widehat{\beta}_i^{(a)} = \widetilde{\beta}_i + m_i^{(a)}$ and

$$\widehat{s}_i - \|\widetilde{\beta}_i\|^2 = \langle\widetilde{\beta}_i,\, m_i^{(1)} + m_i^{(2)}\rangle + \langle m_i^{(1)},\, m_i^{(2)}\rangle. \tag{8}$$

Each summand of $u_i^{(a)}$ has norm at most δ_i by hypothesis, so $\|u_i^{(a)}\| \le \delta_i$ on every trajectory; the proof of (ii) uses only the weaker $(\mathbb{E}\|u_i^{(a)}\|^2)^{1/2} \le \delta_i$. In the linear case $u_i^{(a)} = 0$.

We use one fact about martingale sums twice.

Lemma F.4 (Cross products of martingale sums over disjoint rounds)**.** *Let (X_k) and (Y_k) be $\mathbb{R}^d$-valued martingale difference sequences with respect to $(\mathcal{F}_k)$, with $\mathbb{E}[\|X_k\|^2 \mid \mathcal{F}_{k-1}] \le v_X$ and $\mathbb{E}[\|Y_k\|^2 \mid \mathcal{F}_{k-1}] \le v_Y$, and let $\mathcal{K}_1, \mathcal{K}_2$ be disjoint sets of rounds. Put $A := \sum_{k\in\mathcal{K}_1}X_k$ and $B := \sum_{l\in\mathcal{K}_2}Y_l$. Then $\mathbb{E}\|A\|^2 \le |\mathcal{K}_1|v_X$, $\mathbb{E}\langle A,\,B\rangle = 0$ and $\mathbb{E}\langle A,\,B\rangle^2 \le |\mathcal{K}_1||\mathcal{K}_2|v_Xv_Y$.*

Proof. The first claim is orthogonality of martingale increments. For the other two, order the rounds of $\mathcal{K}_1\cup\mathcal{K}_2$ increasingly and put $A_t := \sum_{k\in\mathcal{K}_1,k<t}X_k$, $B_t := \sum_{l\in\mathcal{K}_2,l<t}Y_l$, both $\mathcal{F}_{t-1}$-measurable.

Every pair $(k,l) \in \mathcal{K}_1 \times \mathcal{K}_2$ has $k \neq l$ and is counted once, at its later round, so

$$\langle A,\, B\rangle = \sum_t D_t, \qquad D_t := \begin{cases} \langle A_t,\, Y_t\rangle, & t \in \mathcal{K}_2, \\ \langle X_t,\, B_t\rangle, & t \in \mathcal{K}_1. \end{cases}$$

Each D_t has conditional mean zero given $\mathcal{F}_{t-1}$, so $\mathbb{E}\langle A,\, B\rangle = 0$ and, by orthogonality of the D_t, $\mathbb{E}\langle A,\, B\rangle^2 = \sum_t \mathbb{E}D_t^2 \le \sum_{t\in\mathcal{K}_2} v_Y \mathbb{E}\|A_t\|^2 + \sum_{t\in\mathcal{K}_1} v_X \mathbb{E}\|B_t\|^2$. With $\mathbb{E}\|A_t\|^2 \le v_X|\{k \in \mathcal{K}_1 : k < t\}|$ and the analogue for B_t, the right side is $v_X v_Y$ times the number of pairs (k,l), i.e. $|\mathcal{K}_1||\mathcal{K}_2| v_X v_Y$. □

With $X_k = \frac{2}{K_0}\widetilde{\varepsilon}_i^k$ this gives $\mathbb{E}\|\nu_i^{(a)}\|^2 \le \frac{K_0}{2}\cdot\frac{4}{K_0^2}\cdot\frac{\tau_i^2}{b} = \vartheta_i^2$, with $\vartheta_i^2 = 2\tau_i^2/(bK_0)$.

The three parts are proved in order, each from the decomposition (8). Part (i) uses the disjointness of the halves to kill both terms in expectation and then Lemma F.4 for the variance. Part (ii) drops every independence claim between trajectory and noise and bounds the terms absolutely. Part (iii) applies T^{-1} and adds the realised inversion error of Lemma F.1.

Proof. (i) Linear crowd. Here $m_i^{(a)} = \nu_i^{(a)}$. Each summand of $\langle \widetilde{\beta}_i,\, \nu_i^{(1)} + \nu_i^{(2)}\rangle$ is $\frac{2}{K_0}\langle \widetilde{\beta}_i,\, \widetilde{\varepsilon}_i^k\rangle$, whose conditional mean given $\mathcal{F}_{k-1}$ is zero because $\widetilde{\beta}_i$ is $\mathcal{F}_0$-measurable; and $\mathbb{E}[\langle \nu_i^{(1)},\, \nu_i^{(2)}\rangle \mid \mathcal{F}_0] = 0$ by Lemma F.4, applied conditionally on $\mathcal{F}_0$. Hence $\mathbb{E}[\widehat{s}_i \mid \mathcal{F}_0] = \|\widetilde{\beta}_i\|^2$ exactly.

For the variance, $\nu_i^{(1)} + \nu_i^{(2)}$ is a martingale sum over all K_0 rounds, so $\mathbb{E}[\|\nu_i^{(1)} + \nu_i^{(2)}\|^2 \mid \mathcal{F}_0] \le 4\tau_i^2/(bK_0)$ and, by Cauchy–Schwarz with $\widetilde{\zeta}_i = \|\widetilde{\beta}_i\|$ fixed given $\mathcal{F}_0$, $\mathbb{E}[\langle \widetilde{\beta}_i,\, \nu_i^{(1)} + \nu_i^{(2)}\rangle^2 \mid \mathcal{F}_0] \le 4\widetilde{\zeta}_i^2\tau_i^2/(bK_0)$. Lemma F.4 with $v_X = v_Y = \frac{4}{K_0^2}\cdot\frac{\tau_i^2}{b}$ and $|\mathcal{K}_1| = |\mathcal{K}_2| = K_0/2$ gives $\mathbb{E}[\langle \nu_i^{(1)},\, \nu_i^{(2)}\rangle^2 \mid \mathcal{F}_0] \le \frac{K_0^2}{4}\cdot\frac{16\tau_i^4}{b^2K_0^4} = \frac{4\tau_i^4}{(bK_0)^2}$. Applying $(u+v)^2 \le 2u^2 + 2v^2$ to (8) yields the stated bound $8\widetilde{\zeta}_i^2\tau_i^2/(bK_0) + 8\tau_i^4/(bK_0)^2$. No dimension appears because Assumption A3 bounds $\mathbb{E}\|\varepsilon\|^2$ directly.

(ii) General. The $u_i^{(a)}$ depend on the trajectory, which depends on the past noise, so no independence between $u_i^{(a)}$ and $\nu_i^{(a)}$ is available and we bound absolutely. By Cauchy–Schwarz applied to (8),

$$\left|\widehat{s}_i - \|\widetilde{\beta}_i\|^2\right| \le \widetilde{\zeta}_i\left(\|m_i^{(1)}\| + \|m_i^{(2)}\|\right) + \|m_i^{(1)}\|\|m_i^{(2)}\|.$$

Minkowski's inequality in L^2 gives $\|m_i^{(a)}\|_{L^2} \le \|u_i^{(a)}\|_{L^2} + \|\nu_i^{(a)}\|_{L^2} \le \delta_i + \vartheta_i$, using only L^2 bounds on each piece. Then $\mathbb{E}\|m_i^{(a)}\| \le \delta_i + \vartheta_i$ by Jensen and $\mathbb{E}[\|m_i^{(1)}\|\|m_i^{(2)}\|] \le \|m_i^{(1)}\|_{L^2}\|m_i^{(2)}\|_{L^2} \le (\delta_i+\vartheta_i)^2$ by Cauchy–Schwarz, so $\mathbb{E}|\widehat{s}_i - \|\widetilde{\beta}_i\|^2| \le 2\widetilde{\zeta}_i(\delta_i+\vartheta_i) + (\delta_i+\vartheta_i)^2$. The same argument applies conditionally on $\mathcal{F}_0$, and on any $\mathcal{F}_{k_0}$ when the probing rounds come after k_0.

(iii) Weights. Lemma F.1 gives $\|\widetilde{\beta}\|^2 = Tq + \eta$ with $\|\eta\|_\infty \le 3\kappa\bar{q}$, so that

$$T^{-1}\widehat{s} - q = T^{-1}\Big(\big(\widehat{s} - \|\widetilde{\beta}\|^2\big) + \eta\Big),$$

and $\|T^{-1}\|_{\infty\to\infty} = C_n$ by Theorem 4.1. Clipping at zero does not increase the error: the positive part is the projection onto $[0,\infty)$, a 1-Lipschitz map fixing $q_i \ge 0$.

High probability. In case (i), Chebyshev's inequality with threshold $\sqrt{n\,\mathrm{Var}_i/\delta}$, Var_i the variance bound of (i), and a union bound over i give $\max_i|\widehat{s}_i - \|\widetilde{\beta}_i\|^2| \le \max_i\sqrt{n\,\mathrm{Var}_i/\delta}$ with probability at least $1-\delta$. In case (ii), Markov's inequality with threshold $n\mathbb{E}|\cdot|/\delta$ and a union bound give the analogous statement. A sharper form of case (ii), used in Theorem C.7, applies Chebyshev to the noise alone: $\mathbb{P}(\|\nu_i^{(a)}\| > \vartheta_i\sqrt{2n/\delta}) \le \delta/(2n)$, so with probability at least $1-\delta$, simultaneously for all i and a, $\|m_i^{(a)}\| \le \delta_i + \vartheta_i\sqrt{2n/\delta}$ when $\|u_i^{(a)}\| \le \delta_i$ surely, and then $|\widehat{s}_i - \|\widetilde{\beta}_i\|^2| \le 2\widetilde{\zeta}_i(\delta_i + \vartheta_i\sqrt{2n/\delta}) + (\delta_i + \vartheta_i\sqrt{2n/\delta})^2$. □

Remark F.5 (Sub-Gaussian improvement). Suppose the centred noises are conditionally norm-sub-Gaussian, $\mathbb{P}(\|\widetilde{\varepsilon}_i^k\| \ge t \mid \mathcal{F}_{k-1}) \le 2e^{-bt^2/(2\tau_i^2)}$. The concentration inequality for vector martingales

with norm-sub-Gaussian increments of Jin et al. (2019) then gives, with probability at least $1-\delta$ and for all i and a simultaneously, $\|\nu_i^{(a)}\| \le C\tau_i\sqrt{\log(8dn/\delta)/(bK_0)}$ with an absolute C. On that event the bound of case (ii) with $\delta_i = 0$ gives $|\widehat{s}_i - \|\widetilde{\beta}_i\|^2| \le 2\widetilde{\zeta}_i C\tau_i\sqrt{\log(8dn/\delta)/(bK_0)} + C^2\tau_i^2\log(8dn/\delta)/(bK_0)$, so the factors $\sqrt{n/\delta}$ and n/δ of the moment version become $\sqrt{\log(dn/\delta)}$ and $\log(dn/\delta)$. We keep the moment version in the main text because it assumes nothing beyond Assumption A3.

F.5 Proof of Lemma 4.3

The argument uses only that the weights lie in the simplex, so it holds verbatim with Δ_n replaced by any closed convex subset $\mathcal{W} \subseteq \Delta_n$, with $\widehat{w}$ and $w^\star$ the minimisers over $\mathcal{W}$, for instance a simplex with capped coordinates.

Proof. For $w \in \Delta_n$, $\sum_i w_i^2 \le (\sum_i w_i)^2 = 1$, so $|\sum_i w_i^2 c_i - \sum_i w_i^2 \widehat{c}_i| \le \sum_i w_i^2 |c_i - \widehat{c}_i| \le \Delta$ for any such w. Let $w^\star$ minimise $\sum_i w_i^2 c_i$ and $\widehat{w}$ minimise $\sum_i w_i^2 \widehat{c}_i$ over the same feasible set; over Δ_n both are given in closed form by Proposition D.5, and $\widehat{w}_i \propto \widehat{c}_i^{-1}$. Then

$$\sum_i \widehat{w}_i^2 c_i \le \sum_i \widehat{w}_i^2 \widehat{c}_i + \Delta \le \sum_i (w_i^\star)^2 \widehat{c}_i + \Delta \le \sum_i (w_i^\star)^2 c_i + 2\Delta,$$

the middle step by optimality of $\widehat{w}$ for $\widehat{c}$. For the specialisation, take $c_i = q_i + \bar{\sigma}_i^2/(bT)$: then $V(\widehat{w}) \le \sum_i \widehat{w}_i^2 c_i$ because $c_i \ge q_i$, and $\min_{w\in\Delta_n} \sum_i w_i^2 c_i = \widehat{\Psi}(bT)$ by Proposition D.5. Finally consider the floored costs $\widehat{c}_i = \max\{(\widehat{q}_i)_+ + \bar{\sigma}_i^2/(bT), \varrho\}$ of Algorithm 2, which are positive by construction – the floor is what guarantees positivity, since $(\widehat{q}_i)_+ = 0$ is attainable and $\bar{\sigma}_i = 0$ is admissible. Write $\widehat{c}_i^{\,0} := (\widehat{q}_i)_+ + \bar{\sigma}_i^2/(bT)$; then $|\widehat{c}_i^{\,0} - c_i| = |(\widehat{q}_i)_+ - q_i| \le |\widehat{q}_i - q_i| \le \Delta$ because the positive part is a 1-Lipschitz projection onto a set containing q_i. If the floor is inactive, $\widehat{c}_i = \widehat{c}_i^{\,0}$; if it is active, $\widehat{c}_i = \varrho > \widehat{c}_i^{\,0}$ and $|\widehat{c}_i - c_i| \le \max\{\varrho - c_i,\, c_i - \widehat{c}_i^{\,0}\} \le \max\{\varrho, \Delta\}$, using $c_i \ge 0$. In both cases $\|\widehat{c} - c\|_\infty \le \Delta + \varrho$, and the first display applies with $\Delta + \varrho$ in place of Δ. □

F.6 A stochastic-approximation lemma

Lemma F.6 (Constant-step projected SGD). *Let F be μ_F-strongly convex and L_F-smooth on a closed convex $\mathcal{X}$ with minimiser $x^\star$ over $\mathcal{X}$, and $x_t = \Pi_{\mathcal{X}}(x_{t-1} - \eta g_t)$ with $\mathbb{E}[g_t \mid \mathcal{F}_{t-1}] = \nabla F(x_{t-1})$ and $\mathbb{E}[\|g_t - \nabla F(x_{t-1})\|^2 \mid \mathcal{F}_{t-1}] \le V$. If $\eta \le 1/L_F$, then $a_t := \mathbb{E}[\|x_t - x^\star\|^2 \mid \mathcal{F}_0]$ obeys $a_t \le (1-\mu_F\eta)a_{t-1} + \eta^2 V$, hence $a_T \le e^{-\mu_F\eta T}a_0 + \eta V/\mu_F$. For any $R > 0$ let $\ell := \ln(e + \mu_F^2 R^2 T/V)$ and $\eta := \min\{\frac{1}{2L_F}, \frac{\ell}{\mu_F T}\}$. Then*

$$a_T \;\le\; a_0 e^{-\mu_F T/(2L_F)} + \Big(\frac{a_0}{R^2} + \ell\Big)\frac{V}{\mu_F^2 T},$$

which is at most $a_0 e^{-\mu_F T/(2L_F)} + (1+\ell)V/(\mu_F^2 T)$ when $a_0 \le R^2$.

Proof. One step. The minimiser is a fixed point of the projected gradient map, $x^\star = \Pi_{\mathcal{X}}(x^\star - \eta\nabla F(x^\star))$, by the first-order optimality condition. Nonexpansiveness of $\Pi_{\mathcal{X}}$ then gives $\|x_t - x^\star\| \le \|x_{t-1} - x^\star - \eta(g_t - \nabla F(x^\star))\|$. Squaring, taking conditional expectations, and writing $e := x_{t-1} - x^\star$ and $\Delta_F := \nabla F(x_{t-1}) - \nabla F(x^\star)$,

$$\mathbb{E}[\|x_t - x^\star\|^2 \mid \mathcal{F}_{t-1}] \le \|e\|^2 - 2\eta\langle \Delta_F, e\rangle + \eta^2\big(\|\Delta_F\|^2 + V\big).$$

For a μ_F-strongly convex L_F-smooth function, $\langle \Delta_F, e\rangle \ge \mu_F\|e\|^2$ and $\langle \Delta_F, e\rangle \ge \|\Delta_F\|^2/L_F$ (Nesterov, 2018, Sec. 2.1); averaging the two gives $\langle \Delta_F, e\rangle \ge \frac{\mu_F}{2}\|e\|^2 + \frac{1}{2L_F}\|\Delta_F\|^2$. Hence the right side is at most $(1-\mu_F\eta)\|e\|^2 - \eta(\frac{1}{L_F} - \eta)\|\Delta_F\|^2 + \eta^2 V \le (1-\mu_F\eta)\|e\|^2 + \eta^2 V$ for $\eta \le 1/L_F$. Taking expectations given $\mathcal{F}_0$ gives the recursion.

Unrolling. $a_T \le (1-\mu_F\eta)^T a_0 + \eta^2 V\sum_{t<T}(1-\mu_F\eta)^t \le e^{-\mu_F\eta T}a_0 + \eta V/\mu_F$.

Tuned step. If $\eta = \ell/(\mu_F T)$ then $e^{-\mu_F\eta T}a_0 = a_0 e^{-\ell} \le a_0 V/(\mu_F^2 R^2 T)$ and $\eta V/\mu_F = \ell V/(\mu_F^2 T)$. Otherwise $\eta = 1/(2L_F) < \ell/(\mu_F T)$, so $e^{-\mu_F\eta T}a_0 = a_0 e^{-\mu_F T/(2L_F)}$ and $\eta V/\mu_F < \ell V/(\mu_F^2 T)$. In both cases the stated bound holds. □

The bound is linear in a_0, so it can be averaged over the conditioning: if only $\mathbb{E}a_0 \le R^2$ is known, the same display holds for $\mathbb{E}\|x_T - x^\star\|^2$ with a_0 replaced by R^2.

F.7 Proof of Theorem 4.4

The proof runs in three steps. Step 1 conditions on the probing phase, which freezes the weights and makes Phase II an ordinary stochastic-approximation run on a fixed objective. Step 2 applies the SGD lemma to that objective. Step 3 converts accuracy on $F_{\widehat{w}}$ into accuracy on the truth by the sharp form of the transfer lemma, and bounds the bias term by the diagonal proxy. Throughout, $\mathbb{E}_{K_0}$ denotes expectation conditional on $\mathcal{F}_{K_0}$.

Proof. Step 1: freezing the weights. $\widehat{w}$, Δ, x^{K_0} and R_0 are $\mathcal{F}_{K_0}$-measurable. During Phase II the aggregate $\sum_i \widehat{w}_i \widehat{g}_i^k$ has conditional mean $\nabla F_{\widehat{w}}(x^{k-1})$ and, by conditional independence across clients and $\sigma_i \le \bar{\sigma}_i$, conditional variance at most $\sum_i \widehat{w}_i^2 \sigma_i^2/b \le V_\sigma$. Since the tilts are linear, $\lambda_{\widehat{w}} = 0$ and $F_{\widehat{w}}$ is μ-strongly convex and L-smooth.

Step 2: optimisation. Lemma F.6, applied conditionally on $\mathcal{F}_{K_0}$ with $a_0 = \|x^{K_0} - x^\star_{\widehat{w}}\|^2 \le R_0^2$, gives $\mathbb{E}_{K_0}\|x^{K_0+T} - x^\star_{\widehat{w}}\|^2 \le R_0^2 e^{-\mu T/(2L)} + \ell_T V_\sigma/(\mu^2 T)$, and L-smoothness gives $F_{\widehat{w}}(x) - \min F_{\widehat{w}} \le \frac{L}{2}\|x - x^\star_{\widehat{w}}\|^2$.

Step 3: transfer. The sharp form of Lemma D.2(ii), a deterministic inequality, holds pointwise at $x = x^{K_0+T}$; taking $\mathbb{E}_{K_0}$ gives, for every $\epsilon > 0$,

$$\mathbb{E}_{K_0}[f_0(x^{K_0+T}) - f_0^\star] \le \frac{1+\epsilon}{2\mu}\|b(\widehat{w})\|^2 + \Big(1 + \tfrac{1}{\epsilon}\Big)\frac{L}{\mu}\cdot\frac{L}{2}\,\mathbb{E}_{K_0}\|x^{K_0+T} - x^\star_{\widehat{w}}\|^2.$$

Lemma D.3 and Lemma 4.3 bound the first term through $\|b(\widehat{w})\|^2 \le (1+\kappa)V(\widehat{w}) \le (1+\kappa)(\widehat{\Psi}(bT) + 2\Delta)$, and Step 2 bounds the second. This is the statement of Theorem 4.4. □

F.8 Proof of Corollary 4.6

Proof of Corollary 4.6. Upper bound. Work conditionally on the probing phase, on the event named in the hypotheses, and apply Theorem 4.4 with $\epsilon = \epsilon_T := T^{-1/4}$. The bias term is $\frac{(1+\epsilon_T)(1+\kappa)}{2\mu}(\widehat{\Psi}(bT) + 2\Delta)$, at most $\frac{(1+\epsilon_T)(1+\epsilon_1)(1+\epsilon_2)}{2\mu}\widehat{\Psi}(bT)$ under the two hypotheses. The other term is $(1 + T^{1/4})\frac{L^2}{2\mu}\big(R_0^2 e^{-\mu T/(2L)} + \ell_T V_\sigma/(\mu^2 T)\big)$. At a fixed centre, R_0, V_σ and K_0 do not depend on T and $\ell_T = O(\log T)$, so this term is $O(T^{-3/4}\log T)$; it vanishes, more slowly than the horizon term $O(T^{-1}\log T)$ of Theorem 4.4. Any schedule with $\epsilon_T \to 0$ and $\epsilon_T^{-1} = o(T/\log T)$ works the same way; ϵ appears only in the analysis, so the algorithm does not depend on it.

Lower bound. Crowd draws $b(K_0 + T)$ samples per client, and a batch of b samples returns a gradient whose noise has second moment σ_i^2/b, which a query-based algorithm can reproduce by averaging b queries. Theorem 4.5 with $K = b(K_0 + T)$ therefore bounds the minimax risk of the problem Crowd faces by $\Psi(b(K_0 + T))/(2\mu)$ from below.

The ratio. As $T \to \infty$, $\widehat{\Psi}(bT) \to \widehat{\Psi}_0$ and $\Psi(b(K_0 + T)) \to \Psi_0$, since $(\sum_i (a_i + t_i)^{-1})^{-1}$ is continuous in $t \ge 0$. Hence the ratio of the upper bound to the lower bound tends to $(1 + \epsilon_1)(1 + \epsilon_2)\widehat{\Psi}_0/\Psi_0$. For the last factor, on the event of Lemma D.4, $(1 - \epsilon_0)\zeta_i^2 \le q_i \le (1 + \epsilon_0)\zeta_i^2$ for every i, and $H(c) = (\sum_i c_i^{-1})^{-1}$ is non-decreasing in each coordinate and positively homogeneous, so $(1 - \epsilon_0)\Psi_0 \le \widehat{\Psi}_0 \le (1 + \epsilon_0)\Psi_0$. □

The comparison is deliberately asymmetric, in two ways, and the corollary's statement in the main text names both. First, it is pointwise in the centre: the upper bound holds for every fixed x_c as $T \to \infty$, while the lower bound is a supremum over x_c. A fixed number of steps from a fixed starting point cannot be uniformly accurate over $x_c \in \mathbb{R}^d$. Since R_0 is the only quantity through which the centre enters, the upper bound is uniform over centres in any fixed ball for T large. This is the form in which an iterative method can approach a minimax floor over an unbounded class. Second, it is conditional on the realisation: the upper bound is against the realised benchmark $\widehat{\Psi}$ on the event named in the hypotheses, while the lower bound averages over the crowd prior. Concavity of H

(Lemma D.7) gives $\mathbb{E}\widehat{\Psi}_0 \le \Psi_0$ by Jensen, and Lemma D.4 bounds the probability of the event. The corollary itself therefore claims per-instance, asymptotic sharpness. Appendix C.6 integrates it into an unconditional statement, through Theorem C.11 and Corollary C.12. The ingredients are a projection that makes the transient uniform, tail control off the incoherence event, and the finite-ρ Bayes floor established in Step 4 of the proof of Theorem 4.5.

F.9 The oracle weights exhaust the disagreement

Proposition F.7 (The oracle weights exhaust the disagreement)**.** *With $w_i^\star = \Psi_0/\zeta_i^2$:* (i) $\mathbb{E}\langle b(w^\star),\, \widetilde{\beta}_j\rangle = 0$ *for every j;* (ii) $\min_{v\in\mathbb{R}^n} \mathbb{E}\|\bar{\beta} - \sum_i v_i\widetilde{\beta}_i\|^2 = \Psi_0$*, attained by $\bar{\beta} - b(w^\star)$;* (iii) *in the Gaussian model $\mathbb{E}[\bar{\beta} \mid \widetilde{\beta}] = \bar{\beta} - b(w^\star)$ and the residual variance is exactly Ψ_0.*

Proof. (i) With $w_i^\star = \Psi_0/\zeta_i^2$ (so $\sum_i w_i^\star = 1$), independence and $\mathbb{E}\beta_i = 0$ give, for each j,

$$\mathbb{E}\langle b(w^\star),\, \widetilde{\beta}_j\rangle = \mathbb{E}\Big\langle \sum_i w_i^\star\beta_i,\ \beta_j - \frac{1}{n}\sum_l \beta_l\Big\rangle = w_j^\star\zeta_j^2 - \frac{1}{n}\sum_i w_i^\star\zeta_i^2 = \Psi_0 - \frac{1}{n}\cdot n\Psi_0 = 0.$$

(ii) For $v\in\mathbb{R}^n$ put $w_i := \frac{1}{n} - v_i + \bar{v}$, where $\bar{v}$ is the mean of v. Since only the centred part of v acts on the disagreements,

$$\bar{\beta} - \sum_i v_i\widetilde{\beta}_i = \sum_i w_i\beta_i = b(w), \qquad \mathbf{1}^\top w = 1.$$

Conversely, given w with $\mathbf{1}^\top w = 1$, the choice $v_i = \frac{1}{n} - w_i$ has $\bar{v} = 0$ and reproduces w. The correspondence $v \leftrightarrow w$ is therefore onto $\{\mathbf{1}^\top w = 1\}$, and by Proposition D.5 the minimal value is $\min_{\mathbf{1}^\top w=1}\sum_i w_i^2\zeta_i^2 = \Psi_0$, attained at $w = w^\star$.

(iii) Write $\bar{\beta} = (\bar{\beta} - b(w^\star)) + b(w^\star)$. The first summand is $\sum_i v_i^\star\widetilde{\beta}_i$ with $v_i^\star = \frac{1}{n} - w_i^\star$ by (ii), a function of $\widetilde{\beta}$. For the second, take coordinates $k,l\in[d]$ and a client j. The coordinates of the β_i are independent across k, so $\mathbb{E}[b(w^\star)_k\widetilde{\beta}_{j,l}] = 0$ for $k\neq l$; for $k = l$ the covariance structure $\frac{\zeta_i^2}{d}I_d$ gives $\mathbb{E}[b(w^\star)_k\widetilde{\beta}_{j,k}] = \frac{1}{d}\mathbb{E}\langle b(w^\star),\, \widetilde{\beta}_j\rangle = 0$ by (i). So $b(w^\star)$ is uncorrelated with every coordinate of $\widetilde{\beta}$, and since $(b(w^\star), \widetilde{\beta})$ is jointly Gaussian it is independent of $\widetilde{\beta}$. Hence $\mathbb{E}[\bar{\beta} \mid \widetilde{\beta}] = \bar{\beta} - b(w^\star)$, and the conditional law of $\bar{\beta}$ given $\widetilde{\beta}$ is Gaussian with covariance that of $b(w^\star)$, namely $\frac{1}{d}\sum_i (w_i^\star)^2\zeta_i^2 I_d = \frac{\Psi_0}{d}I_d$, constant in $\widetilde{\beta}$. □

F.10 Proof of Theorem 4.5

The proof runs in five steps. Step 1 replaces the minimax risk by a Bayes risk with a diffuse prior on the centre, and shows that the transcript is a function of a fixed array that does not depend on the query points. Step 2 reduces that array to a per-client average. Step 3 computes the posterior variance of the centre in the resulting Gaussian location model. Step 4 converts that variance into excess risk and lets the prior become flat. Step 5 repeats the computation with the exact disagreements handed to the algorithm, and shows the bound survives.

Proof. Fix $\rho > 0$ and put the prior $x_c \sim \mathcal{N}(0, \frac{\rho^2}{d}I_d)$, independent of β and of all noises. Since the minimax risk over $x_c\in\mathbb{R}^d$ dominates the Bayes risk for this prior, it suffices to lower bound the latter.

Step 1: reduction to a location model. The algorithm interacts with client i only through at most K gradient queries; the m-th query at a point x returns $\nabla f_i(x) + \varepsilon_i^m = \mu(x - x_c) + \beta_i + \varepsilon_i^m$, with the ε_i^m independent $\mathcal{N}(0, \frac{\sigma_i^2}{d}I_d)$. The algorithm chooses x and knows μ, so the transcript is a measurable function of the private seed U and of the array

$$\xi_i^m := -\mu x_c + \beta_i + \varepsilon_i^m, \qquad i\in[n],\ m\in[K],$$

and conversely (by induction over rounds, the algorithm's state after each report is a measurable function of U and the ξ's seen so far, and adding the known μx recovers the report). Note that

the law of ξ_i^m does not depend on the query point: adaptivity provides no information. Hence $\sigma(\text{transcript}) \subseteq \sigma(U, \{\xi_i^m\})$, and it suffices to bound the Bayes risk of estimating x_c from $(U, \{\xi_i^m\})$.

Step 2: sufficiency. Conditionally on (x_c, β_i) the vectors $\xi_i^1, \dots, \xi_i^K$ are i.i.d. Gaussian with mean $-\mu x_c + \beta_i$ and covariance $\frac{\sigma_i^2}{d} I_d$; hence $\bar{\xi}_i = \frac{1}{K}\sum_m \xi_i^m$ is sufficient, and

$$\bar{\xi}_i = -t + \theta_i, \qquad t := \mu x_c, \qquad \theta_i := \beta_i + \bar{\varepsilon}_i \sim \mathcal{N}\Big(0, \frac{v_i}{d} I_d\Big), \qquad v_i := \zeta_i^2 + \frac{\sigma_i^2}{K},$$

with $\theta_1, \dots, \theta_n$ independent of each other and of t, and $t \sim \mathcal{N}(0, \frac{\mu^2\rho^2}{d} I_d)$.

Step 3: Gaussian Bayes risk. All quantities decouple across the d coordinates. For a single coordinate, t has prior variance $P := \mu^2\rho^2/d$ and we observe n independent measurements $-t + \theta_i$ with noise variances v_i/d; the posterior of t is Gaussian with variance $\big(P^{-1} + \sum_i d/v_i\big)^{-1}$. The Bayes risk for squared error is the posterior variance, and independence of U from everything else cannot reduce it. Summing over d coordinates,

$$\mathbb{E}\|\widehat{t} - t\|^2 \;\geq\; d\Big(\frac{d}{\mu^2\rho^2} + \sum_i \frac{d}{v_i}\Big)^{-1} = \Big(\frac{1}{\mu^2\rho^2} + \sum_i \frac{1}{v_i}\Big)^{-1}.$$

Step 4: back to the objective. $f_0(\widehat{x}) - \min f_0 = \frac{\mu}{2}\|\widehat{x} - x_c\|^2 = \frac{1}{2\mu}\|\mu\widehat{x} - t\|^2$, and $\mu\widehat{x}$ is an admissible estimator of t. Hence the Bayes risk is at least $\frac{1}{2\mu}\big(\frac{1}{\mu^2\rho^2} + \sum_i v_i^{-1}\big)^{-1}$; letting $\rho \to \infty$ gives $\Psi(K)/(2\mu)$.

Step 5: the algorithm may even be told $\widetilde{\beta}$. Condition on $\widetilde{\beta}$. By Proposition F.7(iii), $\bar{\beta} = m(\widetilde{\beta}) + e$ with $m(\widetilde{\beta}) := \bar{\beta} - b(w^\star)$ known once $\widetilde{\beta}$ is, and $e := b(w^\star) \sim \mathcal{N}(0, \frac{\Psi_0}{d} I_d)$ independent of $\widetilde{\beta}$, of t and of the noise. Subtracting the known $\widetilde{\beta}_i + m(\widetilde{\beta})$ from $\bar{\xi}_i$ leaves $z_i = -t + e + \bar{\varepsilon}_i$, $i \in [n]$, with $\bar{\varepsilon}_i \sim \mathcal{N}(0, \frac{\sigma_i^2}{Kd} I_d)$ independent. The z_i share the mean $e - t$, so their precision-weighted average y is sufficient for $e - t$, and $y = e - t + \bar{\varepsilon}$ with $\bar{\varepsilon} \sim \mathcal{N}(0, \frac{s^2}{d} I_d)$, $s^2 := (\sum_i K/\sigma_i^2)^{-1}$ (if some $\sigma_i = 0$ then $s^2 = 0$ and the argument is unchanged). Per coordinate, t has prior variance P and y is t observed through independent Gaussian noise of variance $(\Psi_0 + s^2)/d$, so $\mathrm{Var}(t \mid y) = P(\Psi_0 + s^2)/(dP + \Psi_0 + s^2) \to (\Psi_0 + s^2)/d$ as $P \to \infty$, and the Bayes risk is at least $(\Psi_0 + s^2)/(2\mu)$ in the limit, as in Step 4. By Lemma D.7 applied with $a_i = \zeta_i^2$ and $b_i = \sigma_i^2/K$,

$$\Psi_0 + s^2 \;=\; H(a) + H(b) \;\leq\; H(a + b) \;=\; \Psi(K),$$

with equality exactly when ζ^2 and σ^2 are proportional. Handing the algorithm the disagreements therefore lowers the floor, by the harmonic defect $\Psi(K) - \Psi_0 - s^2$, and never below $\Psi_0/(2\mu)$. □

Remark F.8. The instance used has $L = \mu$, $\mathcal{X} = \mathbb{R}^d$ and Gaussian everything, so the obstruction is neither conditioning, nor constraint geometry, nor heavy tails. It is the diffuse nuisance x_c, i.e. exactly the unidentifiability of Proposition 3.1(i). Step 5 separates the part of the risk the disagreement can remove from the part it cannot. Exact knowledge of $\widetilde{\beta}$ lowers the floor from $\Psi(K)/(2\mu)$ to $(\Psi_0 + s^2)/(2\mu)$. The drop is zero when the bias and noise profiles are proportional and positive otherwise. The consensus error Ψ_0 survives in every case, since it is the one component the disagreement does not determine. Proposition F.7 is the companion statement for weighting rules, where no such side information is available and the floor is Ψ_0 itself.

F.11 Proof of Proposition 4.7

Proof. Every ingredient is deterministic. Lemma 4.3 applied with $c = q$ and the estimate $\widehat{c}$, $\|\widehat{c} - q\|_\infty \leq \Delta$, gives

$$V(\widehat{w}) = \sum_i \widehat{w}_i^2 q_i \;\leq\; \min_{w \in \Delta_n} \sum_i w_i^2 q_i + 2\Delta. \tag{9}$$

Let $w^{\mathrm{opt}} \in \arg\min_{w\in\Delta_n} \|b(w)\|^2 = \arg\min_{w\in\Delta_n} w^\top Q w$, which exists by compactness. Then $\min_{w\in\Delta_n} \sum_i w_i^2 q_i \leq V(w^{\mathrm{opt}}) \leq \frac{1}{1-\kappa}\|b(w^{\mathrm{opt}})\|^2$ by the lower half of Lemma D.3, which needs $\kappa < 1$. Combining with the upper half, $\|b(\widehat{w})\|^2 \leq (1+\kappa)V(\widehat{w})$, and (9),

$$\|b(\widehat{w})\|^2 \leq (1+\kappa)\Big(\frac{\|b(w^{\mathrm{opt}})\|^2}{1-\kappa} + 2\Delta\Big) = \frac{1+\kappa}{1-\kappa} \min_{w\in\Delta_n} \|b(w)\|^2 + 2(1+\kappa)\Delta,$$

which is (6). For the gain over uniform weighting, Proposition D.5 applied to the realised vector q gives $\min_w \sum_i w_i^2 q_i = (\sum_i q_i^{-1})^{-1} = \mathrm{HM}(q)/n$, so $\|b(\widehat{w})\|^2 \le (1+\kappa)(\mathrm{HM}(q)/n + 2\Delta)$, while $\|b(w_{\mathrm{unif}})\|^2 \ge (1-\kappa)V(w_{\mathrm{unif}}) = (1-\kappa)\mathrm{AM}(q)/n$; dividing gives the claim.

The size of Δ. CROWD uses $\widehat{c}_i = (\widehat{q}_i)_+ + \bar{\sigma}_i^2/(bT)$, so $\Delta \le \max_i |\widehat{q}_i - q_i| + \max_i \bar{\sigma}_i^2/(bT)$. The first term is bounded by Lemma 4.2(iii), which assumes Assumption A3 and nothing about the tilts beyond Assumption A4; the second vanishes with the horizon. □

Remark F.9 (What is and is not assumption-free). Proposition 4.7 uses Assumption A3 (about the sampling noise) and Assumption A4 (a property of the realised tilts), and nothing else: in particular no independence, no zero mean, no linearity, and no distribution over crowds. The crowd hypotheses enter only the *interpretation* of the benchmark: under them $\min_w \|b(w)\|^2$ concentrates around Ψ_0, and $\Psi_0/(2\mu)$ is the lower bound of Theorem 4.5.

G PROOFS FOR APPENDIX C

The subsections follow the order of Appendix C.

G.1 PROOF OF PROPOSITION C.1

Proof. (i) This is Proposition 3.1(i). The optimal weights of the two instances are computed from the energies $\|\beta_i\|^2$ and $\|\beta_i - c\|^2$, which differ for generic c.

(ii), non-identifiability. Under the shift of (i) with linear tilts the tilts of the second instance at its own minimiser are $\beta_i - c$, so its Gram matrix is $Q_{ij} - \langle\beta_i, c\rangle - \langle\beta_j, c\rangle + \|c\|^2$, as stated. The transcript law is the same, and κ changes in general, so κ is not a function of that law. The disagreements $\widetilde{\beta}_i = \beta_i - \bar{\beta}$ are invariant, and so is $JQJ = (\langle\widetilde{\beta}_i, \widetilde{\beta}_j\rangle)_{ij}$.

(ii), unbiasedness. For linear tilts $\widehat{\beta}_i^{(a)} = \widetilde{\beta}_i + \nu_i^{(a)}$ in the notation of Appendix F.4. Expanding, $\mathbb{E}[\langle\widehat{\beta}_i^{(1)}, \widehat{\beta}_j^{(2)}\rangle \mid B] = \langle\widetilde{\beta}_i, \widetilde{\beta}_j\rangle + \mathbb{E}[\langle\widetilde{\beta}_i, \nu_j^{(2)}\rangle + \langle\nu_i^{(1)}, \widetilde{\beta}_j\rangle \mid B] + \mathbb{E}[\langle\nu_i^{(1)}, \nu_j^{(2)}\rangle \mid B]$. The middle terms vanish because each is a sum of martingale increments against an $\mathcal{F}_0$-measurable vector, and the last vanishes by Lemma F.4 applied to the martingale difference sequences $(\frac{2}{K_0}\widetilde{\varepsilon}_i^k)_k$ and $(\frac{2}{K_0}\widetilde{\varepsilon}_j^k)_k$ over the disjoint halves. Symmetrising gives $\mathbb{E}[\widehat{G} \mid B] = JQJ$.

(ii), the correction. Write $\varsigma := (Q-D)\mathbf{1}$, so $Q\mathbf{1} = q + \varsigma$ and $\mathbf{1}^\top Q\mathbf{1} = \mathbf{1}^\top q + \mathbf{1}^\top\varsigma$. Expanding JQJ,

$$JQJ = Q - \tfrac{1}{n}\big((q+\varsigma)\mathbf{1}^\top + \mathbf{1}(q+\varsigma)^\top\big) + \tfrac{\mathbf{1}^\top q + \mathbf{1}^\top\varsigma}{n^2}\mathbf{1}\mathbf{1}^\top.$$

The population version of $\widehat{Q}$ adds back $\frac{1}{n}(q\mathbf{1}^\top + \mathbf{1}q^\top) - \frac{\mathbf{1}^\top q}{n^2}\mathbf{1}\mathbf{1}^\top$, so $\widehat{Q}_\circ = Q + E$ with $E := -\frac{1}{n}(\varsigma\mathbf{1}^\top + \mathbf{1}\varsigma^\top) + \frac{\mathbf{1}^\top\varsigma}{n^2}\mathbf{1}\mathbf{1}^\top$; under Assumption A1 each ς_i has conditional mean zero given β_i, which is the sense in which the correction sets the row sums to their mean. With $\widehat{D} = D$, $\widehat{\kappa}_\circ = \|D^{-1/2}(Q - D + E)D^{-1/2}\|_{\mathrm{op}}$, so $|\widehat{\kappa}_\circ - \kappa| \le \|D^{-1/2}ED^{-1/2}\|_{\mathrm{op}}$. Put $u := D^{-1/2}\mathbf{1}$, with $\|u\|^2 = \sum_i q_i^{-1} = n/\mathrm{HM}(q)$, and $r := D^{-1/2}\varsigma = MD^{1/2}\mathbf{1}$, $M := D^{-1/2}(Q-D)D^{-1/2}$, so that $\|r\| \le \kappa\|D^{1/2}\mathbf{1}\| = \kappa\sqrt{n\mathrm{AM}(q)}$ and $|\mathbf{1}^\top\varsigma| = |(D^{1/2}\mathbf{1})^\top M(D^{1/2}\mathbf{1})| \le \kappa n\mathrm{AM}(q)$. Then $D^{-1/2}ED^{-1/2} = -\frac{1}{n}(ru^\top + ur^\top) + \frac{\mathbf{1}^\top\varsigma}{n^2}uu^\top$ and

$$\|D^{-1/2}ED^{-1/2}\|_{\mathrm{op}} \le \frac{2\|r\|\|u\|}{n} + \frac{|\mathbf{1}^\top\varsigma|\|u\|^2}{n^2} \le 2\kappa\sqrt{\mathrm{AM}(q)/\mathrm{HM}(q)} + \kappa\,\mathrm{AM}(q)/\mathrm{HM}(q). \quad \square$$

G.2 PROOF OF PROPOSITION C.4

The proof runs in four steps. Step 1 evaluates the expected bias energy of a correlated crowd and identifies the optimal weights. Step 2 shows that the two candidate covariances have the same disagreement law, so no shift-invariant rule can tell them apart. Step 3 checks that both are legitimate covariances. Step 4 computes the two optima and their values, which differ.

Proof. Step 1: the objective. With $\mathrm{Cov}(\beta_i, \beta_j) = \Sigma_{ij} I_d/d$ and $\mathbb{E}\beta = 0$,

$$\mathbb{E}\|b(w)\|^2 = \sum_{i,j} w_i w_j \, \mathbb{E}\langle \beta_i,\, \beta_j \rangle = \sum_{i,j} w_i w_j \, \mathrm{tr}\left(\Sigma_{ij} I_d/d\right) = w^\top \Sigma w.$$

Minimising over $\{\mathbf{1}^\top w = 1\}$ gives $w \propto \Sigma^{-1}\mathbf{1}$ with value $(\mathbf{1}^\top \Sigma^{-1}\mathbf{1})^{-1}$, by the same Cauchy–Schwarz (in the Σ-inner product) as in Proposition D.5.

Step 2: same disagreement law. $\widetilde{\beta} = (J \otimes I_d)\beta$ has covariance $J\Sigma J \otimes I_d/d$. Since $J\mathbf{1} = 0$, $J(\mathbf{1}u^\top + u\mathbf{1}^\top)J = (J\mathbf{1})u^\top J + Ju\,(\mathbf{1}^\top J) = 0$, so $J\Sigma' J = J\Sigma J$. Hence $\widetilde{\beta}$ has the same law under both models, and by Lemma E.1 so does the residual array (r_i^k) (the noise law is unchanged). Any rule that is a measurable function of the residuals therefore has the same output law under Σ and Σ'.

Step 2′: the kernel is exactly this family. Let Σ be symmetric with $J\Sigma J = 0$. For $z \perp \mathbf{1}$, $Jz = z$, so $J\Sigma z = 0$ and $\Sigma z = \alpha(z)\mathbf{1}$ for a linear functional α on $\mathbf{1}^\perp$; extend it by $\alpha(\mathbf{1}) = 0$ and write $\alpha(z) = a^\top z$ with $a \perp \mathbf{1}$. Then $\Sigma J = \mathbf{1}a^\top$, and with $v := \Sigma\mathbf{1}/n$, $\Sigma = \Sigma J + \Sigma\frac{\mathbf{1}\mathbf{1}^\top}{n} = \mathbf{1}a^\top + v\mathbf{1}^\top$. Symmetry gives $\mathbf{1}a^\top + v\mathbf{1}^\top = a\mathbf{1}^\top + \mathbf{1}v^\top$, i.e. $\mathbf{1}(a-v)^\top = (a-v)\mathbf{1}^\top$; a rank-one matrix $xy^\top$ equal to $yx^\top$ has $y \propto x$, so $a - v = \gamma\mathbf{1}$ for a scalar γ. Hence $\Sigma = \mathbf{1}a^\top + a\mathbf{1}^\top - \gamma\mathbf{1}\mathbf{1}^\top = \mathbf{1}u^\top + u\mathbf{1}^\top$ with $u := a - \frac{\gamma}{2}\mathbf{1}$. The map $u \mapsto \mathbf{1}u^\top + u\mathbf{1}^\top$ is injective: applying the image to $\mathbf{1}$ gives $nu + (\mathbf{1}^\top u)\mathbf{1}$, which vanishes only if $u = \gamma'\mathbf{1}$ for a scalar γ', and then the image $2\gamma'\mathbf{1}\mathbf{1}^\top$ vanishes only if $\gamma' = 0$. So the kernel has dimension n.

Step 3: positive definiteness. $\Sigma = I_3 \succ 0$; $\Sigma' = \left(\begin{smallmatrix} 1.2 & 0.1 & 0.1 \\ 0.1 & 1 & 0 \\ 0.1 & 0 & 1 \end{smallmatrix}\right)$ is symmetric and strictly diagonally dominant with positive diagonal, hence $\succ 0$.

Step 4: different optima. For Σ we have $\Sigma^{-1}\mathbf{1} = \mathbf{1}$, so $w^\star(\Sigma) = \mathbf{1}/3$. For Σ', solving $\Sigma' x = \mathbf{1}$ gives $x_2 = x_3 = 1 - 0.1x_1$ from the last two rows and, from the first,

$$x_1\big(1 + 2t - 2t^2\big) = 1 - 2t \quad \text{with } t = 0.1, \qquad \text{so} \qquad x_1 = 0.6780, \quad x_2 = x_3 = 0.9322.$$

Normalising gives $w^\star(\Sigma') = (0.2667, 0.3667, 0.3667)$, interior to the simplex, so the two optima differ. Finally the values differ strictly: $w^\star(\Sigma)^\top \Sigma' w^\star(\Sigma) = \frac{1}{9}\sum_{ij}\Sigma'_{ij} = \frac{3.6}{9} = 0.4$, whereas $(\mathbf{1}^\top \Sigma'^{-1}\mathbf{1})^{-1} = 1/2.5424 = 0.3933$. Hence a rule that is optimal under Σ (necessarily outputting $\mathbf{1}/3$, up to the shared output law) is suboptimal under Σ' by a factor 1.017; in the other direction $w^\star(\Sigma')^\top \Sigma w^\star(\Sigma') = 0.3400$ against $1/3$, a factor 1.020. □

G.3 Proof of Theorem C.5

The proof runs in four steps. Step 1 records the measurability fact that makes each segment an ordinary run with a fixed objective. Step 2 locates the last refresh. Step 3 proves the display by applying the SGD lemma to the final segment. Step 4 proves the bound of Remark C.6 on $\widetilde{R}$ by a recursion over the segments. Everything is conditional on the tilts.

Proof. Step 1: segmentation. Number the distinct refreshes $t_1 < \cdots < t_J$; the hypothesis $T \geq \gamma^{j_0+1}$ gives $J \geq 1$, so the last refresh and Δ_J are defined. Repeated grid values $\lceil \gamma^j \rceil$, possible for γ close to one and absent for $\gamma \geq 2$, are counted once, so no segment has length zero; the bound on the number $N_\star$ of short segments in Step 4 counts distinct refreshes and is unaffected. Put $t_0 := 0$, $t_{J+1} := T$; segment j is the rounds $t_j < k \leq t_{j+1}$, run with weights $w^{(j)}$ ($w^{(0)}$ uniform). The weights $w^{(j)}$ are $\mathcal{F}_{t_j}$-measurable, being computed from rounds before the segment starts. Conditionally on $\mathcal{F}_{t_j}$ the objective $F_{w^{(j)}}$ is therefore fixed, the aggregate is conditionally unbiased for its gradient with conditional variance at most $\sum_i (w_i^{(j)})^2 \sigma_i^2/b \leq \bar{V}$, and Lemma F.6 applies to the segment.

Step 2: the last refresh. The grid points are $\lceil \gamma^j \rceil \leq T/\gamma$, so $t_J \leq T/\gamma$ and $T_J = T - t_J \geq (1-\gamma^{-1})T$. The next grid point exceeds T/γ, so $\gamma^{j+1} + 1 > T/\gamma$ for the last index j, whence $t_J \geq \gamma^j > T/\gamma^2 - 1$.

Step 3: the final segment. The refresh at t_J splits the t_J rounds seen so far by parity; when t_J is odd the two halves differ in size by one, and Lemma F.3 holds with $|\mathcal{K}_a|$ in place of $K_0/2$, which changes ϑ_i by a factor at most $\sqrt{t_J/(t_J-1)}$ and leaves $\mathbb{E}\Delta_J$ of the same order. Conditionally on $\mathcal{F}_{t_J}$, apply Lemma F.6 to segment J with $\mu_F = \mu$, $L_F = L$, $V = \bar{V}$, $R = \bar{R}$, horizon T_J and

$a_0 = \|x^{t_J} - x^\star_{w^{(J)}}\|^2$. Its logarithm is $\ln(e + \mu^2\bar{R}^2 T_J/\bar{V}) \le \bar{\ell}$, so

$$\mathbb{E}\big[\|x^T - x^\star_{w^{(J)}}\|^2 \,\big|\, \mathcal{F}_{t_J}\big] \le a_0\Big(e^{-\mu T_J/(2L)} + \frac{\bar{V}}{\mu^2\bar{R}^2 T_J}\Big) + \frac{\bar{\ell}\bar{V}}{\mu^2 T_J}.$$

The sharp form of Lemma D.2(ii) holds pointwise, and Lemmas D.3 and 4.3, with $K = T_J$ at the last refresh, give $\|b(w^{(J)})\|^2 \le (1+\kappa)(\widehat{\Psi}(bT_J) + 2\Delta_J)$. Combining as in Appendix F.7 and taking expectations over $\mathcal{F}_{t_J}$ gives the display, with $\mathbb{E}a_0 = \widetilde{R}^2$ and $\mathbb{E}\Delta_J$ in place of a_0 and Δ_J.

Step 4: the bound on $\widetilde{R}$. Put $\rho_j := (\mathbb{E}\|x^{t_j} - x^\star_{w^{(j)}}\|^2)^{1/2}$ and $s_j := t_{j+1} - t_j$, so that $\widetilde{R} = \rho_J$. By Remark C.6 every target lies within G of $x_0^\star$, so $\rho_0 \le \|x^0 - x_0^\star\| + G$, and by Minkowski's inequality in L^2, $\rho_{j+1} \le (\mathbb{E}\|x^{t_{j+1}} - x^\star_{w^{(j)}}\|^2)^{1/2} + 2G$. Two bounds on the first term are available for segment j.

- *Any segment.* The recursion $a_t \le (1-\mu\eta)a_{t-1} + \eta^2\bar{V}$ of Lemma F.6 keeps $a_t \le \max(a_0, \eta\bar{V}/\mu) \le a_0 + \bar{V}/(2L\mu)$, since $\eta \le 1/(2L)$; so $\rho_{j+1} \le \rho_j + \sqrt{\bar{V}/(2L\mu)} + 2G \le \rho_j + c_\star$.
- *Long segments.* If $s_j \ge T_\star$ then $e^{-\mu s_j/(2L)} \le \frac{1}{64}$ and $\bar{V}/(\mu^2\bar{R}^2 s_j) \le \frac{1}{8}$, so the tuned bound of Lemma F.6 gives $\mathbb{E}\|x^{t_{j+1}} - x^\star_{w^{(j)}}\|^2 \le \frac{1}{4}\rho_j^2 + \ell(s_j)\bar{V}/(\mu^2 s_j)$ with $\ell(s) := 1 + \ln(e + \mu^2\bar{R}^2 s/\bar{V})$, hence $\rho_{j+1} \le \rho_j/2 + c_\star$.

For $j \ge 1$ the segment runs between consecutive grid points, so $s_j \ge \gamma^{j'}(\gamma-1) - 1$ with j' its grid index, which is at least $T_\star$ once $\gamma^{j'} \ge 2T_\star/(\gamma-1)$, using $T_\star \ge 1$. So all but the first $N_\star$ segments, segment 0 included, are long. Over the short ones $\rho \le \rho_0 + N_\star c_\star$. Over the long ones, if $\rho_j \le \Lambda$ with $\Lambda \ge 2c_\star$ then $\rho_{j+1} \le \Lambda/2 + c_\star \le \Lambda$. With $\Lambda := \rho_0 + (N_\star + 2)c_\star$ both regimes stay below Λ, which is the bound of Remark C.6. □

G.4 Proof of Theorem C.7

The proof runs in three steps. Step 1 bounds the staleness of the residual by the diameter of the domain times the Lipschitz constant of the centred gradient, which is what Lemma 4.2 needs. Step 2 runs the optimisation phase with the moduli of the weighted objective. Step 3 transfers to the truth with the constrained transfer lemma, keeping the two smoothness constants apart.

Proof. Step 1: staleness. On $\mathcal{X}$, $\|\nabla^2 h_i\|_{\mathrm{op}} \le \ell_i$, so $\widetilde{\nabla h_i} = \nabla h_i - \frac{1}{n}\sum_j \nabla h_j$ is $\widetilde{\ell}_i$-Lipschitz with $\widetilde{\ell}_i = \ell_i + \frac{1}{n}\sum_j \ell_j$. Every query point lies in $\mathcal{X}$, as does $x_0^\star$, so for every $k \le K_0$ and every i,

$$\|\widetilde{\nabla h_i}(x^{k-1}) - \widetilde{\beta}_i\| \le \widetilde{\ell}_i\|x^{k-1} - x_0^\star\| \le \widetilde{\ell}_i\bar{D} =: \delta_i$$

surely. Lemma 4.2(ii)–(iii) therefore applies with these δ_i, and its high-probability form gives, with probability at least $1-\delta$, $\Delta \le C_n\big(\max_i[2\widetilde{\zeta}_i(\delta_i + \vartheta_i\sqrt{2n/\delta}) + (\delta_i + \vartheta_i\sqrt{2n/\delta})^2] + 3\kappa\bar{q}\big)$.

Step 2: optimisation. Conditionally on $\mathcal{F}_{K_0}$ the weights $\widehat{w}$ are fixed, $\nabla^2 F_{\widehat{w}} = \nabla^2 f_0 + \nabla^2 H_{\widehat{w}}$ with $\|\nabla^2 H_{\widehat{w}}\|_{\mathrm{op}} \le \lambda$, so $F_{\widehat{w}}$ is $(\mu-\lambda)$-strongly convex and L'-smooth on $\mathcal{X}$, and the aggregate has conditional variance at most V_σ. Lemma F.6 with $\mu_F = \mu - \lambda$, $L_F = L'$, $R = \bar{D}$ and $a_0 \le \bar{D}^2$ (both points lie in $\mathcal{X}$) gives

$$\mathbb{E}_{K_0}\|x^{K_0+T} - x^\star_{\widehat{w}}\|^2 \le \bar{D}^2 e^{-(\mu-\lambda)T/(2L')} + \frac{\ell_T V_\sigma}{(\mu-\lambda)^2 T},$$

where $x^\star_{\widehat{w}}$ minimises $F_{\widehat{w}}$ over $\mathcal{X}$.

Step 3: transfer. Lemma D.6 applies because $x_0^\star$ is interior and $\lambda < \mu$; it gives, pointwise, $f_0(x) - f_0^\star \le \frac{2L}{\mu-\lambda}[F_{\widehat{w}}(x) - F_{\widehat{w}}(x^\star_{\widehat{w}})] + \frac{L}{(\mu-\lambda)^2}\|b(\widehat{w})\|^2$. Two smoothness constants occur: L is that of the truth, used by the lemma, while L' is that of $F_{\widehat{w}}$, which converts iterate accuracy into function accuracy. By Lemma D.6, Lemma D.3 and $V(\widehat{w}) \le \max_i q_i$, $\|x^\star_{\widehat{w}} - x_0^\star\| \le \Gamma_{\widehat{w}}/(\mu-\lambda) \le \Gamma_{\max}/(\mu-\lambda)$, so $x^\star_{\widehat{w}}$ lies in the interior of $\mathcal{X}$ by hypothesis and $\nabla F_{\widehat{w}}(x^\star_{\widehat{w}}) = 0$. L'-smoothness then gives $F_{\widehat{w}}(x) - F_{\widehat{w}}(x^\star_{\widehat{w}}) \le \frac{L'}{2}\|x - x^\star_{\widehat{w}}\|^2$, and

$$\mathbb{E}_{K_0}[f_0 - f_0^\star] \le \frac{2L}{\mu-\lambda}\cdot\frac{L'}{2}\,\mathbb{E}_{K_0}\|x - x^\star_{\widehat{w}}\|^2 + \frac{L}{(\mu-\lambda)^2}\|b(\widehat{w})\|^2,$$

and $\|b(\widehat{w})\|^2 \le (1+\kappa)V(\widehat{w}) \le (1+\kappa)(\widehat{\Psi}(bT)+2\Delta)$ by Lemmas D.3 and 4.3. Collecting terms gives the display. □

G.5 Proof of Proposition C.9

Proof. (i) Fix k and i and condition on $\mathcal{F}_{k-1}$ and on $i \in P_k$. By Assumption A5 the other $m-1$ members of P_k form a uniform $(m-1)$-subset S of $[n]\setminus\{i\}$, independent of the past and of the round's noise. For linear tilts, $r_i^k = \beta_i - \frac{1}{m}\sum_{j\in P_k}\beta_j + \widetilde{\varepsilon}_i^{P,k}$ with $\widetilde{\varepsilon}_i^{P,k} := \varepsilon_i^k - \frac{1}{m}\sum_{j\in P_k}\varepsilon_j^k$, whose conditional mean is zero. Since $\mathbb{E}\sum_{j\in S}\beta_j = \frac{m-1}{n-1}\sum_{j\neq i}\beta_j = \frac{m-1}{n-1}(n\bar\beta - \beta_i)$,

$$\mathbb{E}\Big[\beta_i - \tfrac{1}{m}\sum_{j\in P_k}\beta_j\Big] = \Big(1 - \tfrac{1}{m} + \tfrac{m-1}{m(n-1)}\Big)\beta_i - \tfrac{(m-1)n}{m(n-1)}\bar\beta = \rho(\beta_i - \bar\beta),$$

because $1 - \frac{1}{m} + \frac{m-1}{m(n-1)} = \frac{m-1}{m}\cdot\frac{n}{n-1} = \rho$. The bounds on ρ follow from $\rho = \frac{m-1}{m}\cdot\frac{n}{n-1}$, which is increasing in m, equals $n/(2(n-1)) > \frac{1}{2}$ at $m=2$ and 1 at $m=n$.

(ii) Condition on the tilts and on the indicators $(\mathbf{1}\{i\in P_k\})_k$. The P_k are independent across rounds, independent of the noise, and given $i\in P_k$ the set $P_k\setminus\{i\}$ is a uniform $(m-1)$-subset of $[n]\setminus\{i\}$. So, enlarging $\mathcal{F}_0$ by the indicators, which are independent of everything else, the facts used in (i) still hold at every round with $i\in P_k$. On such a round decompose $r_i^k = \rho\widetilde{\beta}_i + \chi_i^k + \widetilde{\varepsilon}_i^{P,k}$ with $\chi_i^k := \beta_i - \frac{1}{m}\sum_{j\in P_k}\beta_j - \rho\widetilde{\beta}_i$. By (i), $\mathbb{E}[\chi_i^k \mid \mathcal{F}_{k-1}] = 0$, and $\widetilde{\varepsilon}_i^{P,k}$ has conditional mean zero given P_k; so $X_k := \chi_i^k + \widetilde{\varepsilon}_i^{P,k}$ is a martingale difference sequence over the rounds with $i\in P_k$, and the two terms are orthogonal. For the first, the variance of the sum of a simple random sample of size $m-1$ from the $n-1$ vectors $(\beta_j)_{j\neq i}$ is $\frac{(m-1)(n-m)}{(n-1)(n-2)}\sum_{j\neq i}\|\beta_j - \bar\beta_{-i}\|^2 \le \frac{(m-1)(n-m)}{(n-1)(n-2)}\,n\bar q$ (Cochran, 1977, Thm. 2.2), so $\mathbb{E}[\|\chi_i^k\|^2 \mid \mathcal{F}_{k-1}] \le \frac{(m-1)(n-m)n}{m^2(n-1)(n-2)}\bar q \le \psi_{m,n}\bar q$. For the second, conditional independence of the noise across clients gives $\mathbb{E}[\|\widetilde{\varepsilon}_i^{P,k}\|^2 \mid \mathcal{F}_{k-1}] \le \frac{1}{b}\big((1-\frac{1}{m})^2\sigma_i^2 + \frac{m-1}{m^2(n-1)}\sum_{j\neq i}\sigma_j^2\big) \le \tau_{i,P}^2/b$. Hence $\widehat{\beta}_i^{(a)} = \rho\widetilde{\beta}_i + m_i^{(a)}$ with $m_i^{(a)}$ the average of $N_i^{(a)}$ martingale increments, $\mathbb{E}\|m_i^{(a)}\|^2 \le \vartheta_{i,a}^2$. The decomposition (8) with $\rho\widetilde{\beta}_i$ in place of $\widetilde{\beta}_i$ and Lemma F.4 over the disjoint halves give $\mathbb{E}\widehat{s}_i = \rho^2\|\widetilde{\beta}_i\|^2$, $\mathbb{E}\langle\rho\widetilde{\beta}_i,\, m_i^{(1)} + m_i^{(2)}\rangle^2 \le \rho^2\|\widetilde{\beta}_i\|^2(\vartheta_{i,1}^2 + \vartheta_{i,2}^2)$ and $\mathbb{E}\langle m_i^{(1)},\, m_i^{(2)}\rangle^2 \le \vartheta_{i,1}^2\vartheta_{i,2}^2$, and $(u+v)^2 \le 2u^2 + 2v^2$ gives the variance bound. Finally $N_i^{(a)} \sim \mathrm{Bin}(K_0/2, p)$, and the multiplicative Chernoff bound gives $\mathbb{P}(N_i^{(a)} < pK_0/4) \le e^{-pK_0/16}$; a union bound over i and a finishes.

(iii) As in Lemma 4.2(iii), with $\widehat{s}/\rho^2$ in place of $\widehat{s}$. □

G.6 Proof of Proposition C.10

Proof. (i) The transcript is generated by the reporting clients only, and Proposition 3.1(i), applied to the instance restricted to $\mathcal{R}$, shows that the shift by c leaves its law unchanged. In the shifted instance the tilts are $\beta_i - c = \beta_i'$, independent and centred, which is Assumption A1 on $\mathcal{R}$ for linear tilts. *(ii)* $\sum_i w_i\beta_i = \sum_i w_i(c+\beta_i') = c + b'(w)$ because $\sum_i w_i = 1$. *(iii)* $\beta_i - \bar\beta = \beta_i' - \bar\beta'$ for every $i\in\mathcal{R}$, so the residuals, and every statistic built from them, coincide. In the shifted instance the tilts are β_i', whose energies are $\|\beta_i'\|^2$, and the analysis of Section 4 applies to it. □

G.7 Proofs of Theorem C.11 and Corollary C.12

Throughout, $\mathfrak{G} := \sigma(x_c, \beta_1, \dots, \beta_n)$ and $\mathcal{F}_k := \mathfrak{G} \vee \sigma(U, \text{noises through round } k)$, U the algorithm's private randomness. The oracle noises are independent of (x_c, β), so Assumption A3 holds conditionally on $\mathfrak{G}$ with $\bar\sigma_i = \sigma_i$, the tilts are $\mathcal{F}_0$-measurable, and Lemmas E.1, F.3 and F.6 apply conditionally. Write $q_i = \|\beta_i\|^2$, $\bar q = \mathrm{AM}(q)$, and $\mathrm{exc} := f_0(x^{K_0+T}) - f_0^\star = \frac{\mu}{2}\|x^{K_0+T} - x_c\|^2$.

Proof of Theorem C.11. Define two events. G is the event of Lemma D.4 at confidence δ: $\max_i|q_i/\zeta_i^2 - 1| \le \epsilon_0$ and $\kappa \le \epsilon_2$, with $\mathbb{P}(G^c) \le \delta$; $G\in\sigma(\beta)$. And $E_c := \{\|x_c - x^0\| \le 2\rho\}$; since $d\|x_c - x^0\|^2/\rho^2 \sim \chi_d^2$, the Laurent–Massart bound gives $\mathbb{P}(E_c^c) \le e^{-d/4}$, and $E_c \in \sigma(x_c)$ is independent of G.

Step 1: off the good event. The output lies in $\mathcal{X} = B(x^0, D)$, so on every outcome $\|x^{K_0+T} - x_c\| \le D + \|x_c - x^0\|$ and $\mathrm{exc} \le \mu(D^2 + \|x_c - x^0\|^2)$. Independence of G and x_c gives $\mathbb{E}[\mathbf{1}_{G^c}\,\mathrm{exc}] \le \mu\delta(D^2 + \rho^2)$. By Cauchy–Schwarz and $\mathbb{E}\|x_c - x^0\|^4 = \rho^4(1 + 2/d) \le 3\rho^4$,

$$\mathbb{E}[\mathbf{1}_{E_c^c}\,\mathrm{exc}] \le \mu\big(D^2 e^{-d/4} + \sqrt{3}\,\rho^2 e^{-d/8}\big) \le \sqrt{3}\,\mu\,(D^2 + \rho^2)\,e^{-d/8}.$$

Since $\mathbf{1}_{(G\cap E_c)^c} \le \mathbf{1}_{G^c} + \mathbf{1}_{E_c^c}$,

$$\mathbb{E}\big[\mathbf{1}_{(G\cap E_c)^c}\,\mathrm{exc}\big] \;\le\; 2\mu\,(D^2 + \rho^2)\big(\delta + e^{-d/8}\big). \tag{10}$$

Step 2: geometry on the good event. On G, $\max_i \|\beta_i\| \le \sqrt{(1+\epsilon_0)}\,\zeta_{\max} \le \sqrt{2}\,\zeta_{\max}$, so $\|b(w)\| \le \sqrt{2}\,\zeta_{\max}$ for every $w \in \Delta_n$. The tilts are linear and f_0 is the μ-quadratic, so $F_{\widehat{w}}$ is μ-strongly convex and μ-smooth with unconstrained minimiser $x^\star_{\widehat{w}} = x_c - b(\widehat{w})/\mu$; on $G \cap E_c$, $\|x^\star_{\widehat{w}} - x^0\| \le 2\rho + \sqrt{2}\,\zeta_{\max}/\mu = D$, so $x^\star_{\widehat{w}} \in \mathcal{X}$ and the minimiser of $F_{\widehat{w}}$ over $\mathcal{X}$ is $x^\star_{\widehat{w}}$.

Step 3: conditional bound. $\mathbf{1}_{G\cap E_c}$ is $\mathcal{F}_{K_0}$-measurable. Conditionally on $\mathcal{F}_{K_0}$, Phase II is projected SGD on $F_{\widehat{w}}$ over $\mathcal{X}$ with unbiased gradients of conditional variance at most $\sum_i \widehat{w}_i^2\sigma_i^2/b \le \bar{V}$. Because CROWD* projects onto $\mathcal{X}$ in every round, probing included, $x^{K_0} \in \mathcal{X}$; on $G \cap E_c$ also $x^\star_{\widehat{w}} \in \mathcal{X}$ by Step 2, and $\mathcal{X}$ has diameter $2D$. Lemma F.6 with $\mu_F = L_F = \mu$ and $a_0 = \|x^{K_0} - x^\star_{\widehat{w}}\|^2 \le (2D)^2 = R^2$ therefore gives

$$\mathbb{E}_{K_0}\|x^{K_0+T} - x^\star_{\widehat{w}}\|^2 \;\le\; 4D^2 e^{-T/2} + \frac{\ell_T \bar{V}}{\mu^2 T}.$$

Lemma D.2(ii) is a pointwise inequality on $\mathbb{R}^d$ for a linear crowd, with $\min F_{\widehat{w}}$ the unconstrained minimum. On $G \cap E_c$ that minimum is attained at $x^\star_{\widehat{w}} \in \mathcal{X}$, where $\nabla F_{\widehat{w}}(x^\star_{\widehat{w}}) = 0$, so μ-smoothness gives $F_{\widehat{w}}(x) - F_{\widehat{w}}(x^\star_{\widehat{w}}) \le \frac{\mu}{2}\|x - x^\star_{\widehat{w}}\|^2$ at the output $x = x^{K_0+T} \in \mathcal{X}$. (Off $G\cap E_c$ the unconstrained minimiser may leave $\mathcal{X}$, and the display below is neither claimed nor used there.) For every $\epsilon_T > 0$, on $G \cap E_c$,

$$\mathbb{E}_{K_0}[\mathrm{exc}] \;\le\; \frac{1+\epsilon_T}{2\mu}\,\Gamma^2_{\widehat{w}} + \big(1 + \epsilon_T^{-1}\big)\frac{\mu}{2}\Big(4D^2 e^{-T/2} + \frac{\ell_T \bar{V}}{\mu^2 T}\Big). \tag{11}$$

Step 4: the bias term in expectation. On G, Assumption A4 holds with $\kappa \le \epsilon_2 < 1$, so Lemma D.3 and Lemma 4.3 with the floored costs and $c_i = q_i + \sigma_i^2/(bT)$ give $\Gamma^2_{\widehat{w}} \le (1+\kappa)\big(\widehat{\Psi}(bT) + 2(\Delta + \varrho)\big)$ with Δ as in Lemma F.3(iii). Three bounds hold on G. First, $q_i + \sigma_i^2/(bT) \le (1+\epsilon_0)(\zeta_i^2 + \sigma_i^2/(bT))$, so monotonicity and homogeneity of H give $\widehat{\Psi}(bT) \le (1+\epsilon_0)\Psi(bT)$. Second, $\Delta \le C_n(M_s + 3\kappa\bar{q})$ with $M_s := \max_i |\widehat{s}_i - \|\widetilde{\beta}_i\|^2|$, and $3\kappa\bar{q}\,\mathbf{1}_G \le 3\epsilon_2(1+\epsilon_0)\mathrm{AM}$. Third, the estimation term is bounded without conditioning on G, since $\mathbf{1}_{G\cap E_c} M_s \le M_s$. Lemma F.3(i) applies conditionally on $\mathfrak{G}$: the crowd is linear ($\delta_i = 0$), K_0 is even, all n clients report in every probing round, and projected query points are adaptive query points, which the lemma allows. It gives $\mathbb{E}[|\widehat{s}_i - \|\widetilde{\beta}_i\|^2| \mid \mathfrak{G}] \le (4\widetilde{\zeta}_i^2\vartheta_i^2 + 2\vartheta_i^4)^{1/2} \le 2\widetilde{\zeta}_i\vartheta_i + \sqrt{2}\,\vartheta_i^2$ with ϑ_i deterministic. Taking the unconditional expectation and using $\mathbb{E}\widetilde{\zeta}_i \le (\mathbb{E}\widetilde{\zeta}_i^2)^{1/2} = \sqrt{s_i}$, where the equality is the unconditional identity of Theorem 4.1, $\mathbb{E}[\mathbf{1}_{G\cap E_c} M_s] \le \mathbb{E}M_s \le \sum_i \mathbb{E}|\widehat{s}_i - \|\widetilde{\beta}_i\|^2| \le a(K_0)$. (Summing over i bounds the maximum crudely; a union bound would replace the factor n by about $\sqrt{n}$.) Combining, and using $1 + \kappa \le 1 + \epsilon_2$ and $1 \le 1 + \epsilon_0$ where convenient,

$$\mathbb{E}\big[\mathbf{1}_{G\cap E_c}(1+\kappa)\big(\widehat{\Psi}(bT) + 2\Delta + 2\varrho\big)\big] \;\le\; (1+\epsilon_2)(1+\epsilon_0)\Big[\Psi(bT) + 2C_n a(K_0) + 6C_n\epsilon_2\mathrm{AM} + 2\varrho\Big].$$

Step 5: assembly. Multiply (11) by $\mathbf{1}_{G\cap E_c}$, take full expectations, insert Step 4, and add (10). □

Proof of Corollary C.12. The ratio is at least one at every parameter value by (7). For the upper limit, note first that $\Psi(K)$ is non-increasing in K with limit Ψ_0, so $\mathfrak{B}^\star_\rho(K) \ge \mathfrak{B}^\star_\rho(\infty) := \big((\mu^2\rho^2)^{-1} + \Psi_0^{-1}\big)^{-1}/(2\mu) > 0$ for every K; dividing the bound of Theorem C.11 by $\mathfrak{B}^\star_\rho(\infty)$ therefore bounds the ratio at every T and K_0. Hold ρ, d, K_0 fixed and let $T \to \infty$ with $\varrho = 1/T$ and $\epsilon_T = T^{-1/4}$: the transient of Theorem C.11 vanishes ($\ell_T = O(\log T)$), $\Psi(bT) \to \Psi_0$ and $\varrho \to 0$, so

$$\limsup_{T\to\infty} \frac{\mathfrak{B}_\rho(\mathrm{CROWD}^\star)}{\mathfrak{B}^\star_\rho(b(K_0+T))} \;\le\; \frac{(1+\epsilon_2)(1+\epsilon_0)\big[\Psi_0 + 2C_n a(K_0) + 6C_n\epsilon_2\mathrm{AM}\big] + 4\mu^2(D^2+\rho^2)(\delta + e^{-d/8})}{\big((\mu^2\rho^2)^{-1} + \Psi_0^{-1}\big)^{-1}}.$$

Now let $K_0 \to \infty$: $a(K_0) \to 0$ since every $\vartheta_i \to 0$. Then let $d \to \infty$ with $\delta = e^{-\sqrt{d}}$: $\log(4n/\delta) = \log 4n + \sqrt{d}$, so the ϵ_0 and ϵ_1 of Lemma D.4 tend to zero and $\epsilon_2 \to 0$. The hypotheses of Lemma D.4, $\epsilon_0 \le \frac{1}{2}$ and $\epsilon_1 \le 1$, hold once $d \ge C'(n + \log 4n + \sqrt{d})$, which is true for all large d, and then $\epsilon_2 < 1$ as Theorem C.11 requires; D does not depend on d, so the tail term vanishes. The bound tends to $\Psi_0/\big((\mu^2\rho^2)^{-1} + \Psi_0^{-1}\big)^{-1} = 1 + \Psi_0/(\mu^2\rho^2)$, which is the displayed limit at fixed ρ and tends to one as $\rho \to \infty$. For the $(1+\epsilon)$ form with $\epsilon \in (0, 1]$: choose ρ_0 with $\Psi_0/(\mu^2\rho_0^2) \le \epsilon/4$; for each $\rho \ge \rho_0$ choose d_0 and K_0' so that the $T \to \infty$ limit of the bound is at most $(1+\epsilon/4)^2$; then choose T_0 so that the bound itself is at most $(1+\epsilon/4)^3$ for $T \ge T_0$, which is possible because it converges and the denominator was replaced by its T-independent lower bound $\mathfrak{B}_\rho^\star(\infty)$. Finally $(1+\epsilon/4)^3 \le 1 + \frac{3}{4}\epsilon + \frac{3}{16}\epsilon^2 + \frac{1}{64}\epsilon^3 \le 1 + \epsilon$ for $\epsilon \le 1$. □

H Experimental details and further figures

This section records how the runs of Section 5 were produced and reports the ones that did not fit there. Every figure comes from a seeded script. The supplementary materials contain the code, the recorded output of every run holding each quoted number, a README with per-experiment instructions and runtimes, and a document mapping the quoted numbers to the statements they test. The synthetic instances match the model by construction: their purpose is to check constants, exponents and floors, not to argue about realism. The studies on real scans begin in Section H.2.

Default instance. $n = 20$, $d = 150$, $f_0(x) = \frac{1}{2}(x - x_c)^\top A(x - x_c)$ with $A = Q\Lambda Q^\top$, Q random orthogonal and Λ geometric in $[\mu, L] = [0.5, 5]$; $x_c \sim \mathcal{N}(0, 4I_d/d)$; $\log \zeta_i \sim \mathrm{U}[\log 0.2, \log 3]$, $\beta_i \sim \mathcal{N}(0, \zeta_i^2 I_d/d)$; $\log \sigma_i \sim \mathrm{U}[\log 0.5, \log 3]$, Gaussian gradient noise of second moment σ_i^2/b, $b = 1$.

Rules compared. *Uniform*; *size-proportional* ($w_i \propto n_i$, $n_i \sim \mathrm{U}\{20, \dots, 199\}$ drawn independently of β, so size and quality are unrelated); *inverse-noise* ($w_i \propto \sigma_i^{-2}$); *best single client* ($w = \delta_{i^\star}$, $i^\star = \arg\min_i \|\beta_i\|^2$, chosen with oracle knowledge); *oracle* ($w_i \propto \|\beta_i\|^{-2}$); and CROWD in its anytime form, refreshing at every round. All take the same stochastic-gradient steps and differ only in w.

Figure-specific settings.

- **Figure 4**: 12 seeds, $K = 4000$, $\eta = 1/(2L)$; reports the exact realised bias energy $\|B^\top w^k\|^2$.
- **Figure 2a**: 12 seeds, $K = 20000$, $\eta_k = \min\{1/(2L), 2/(\mu(k + 2L/\mu))\}$; the decaying step makes the sampling term vanish so the bias floors become visible. Dotted lines are $\frac{1}{2}b(w)^\top A^{-1} b(w)$, which for a quadratic f_0 is the exact excess risk of $\arg\min F_w$ – a strengthening of Lemma D.2(iii) holding with equality.
- **Figure 5**: 12 seeds; panel (a) reports $\max_i |\widehat{s}_i - \|\widetilde{\beta}_i\|^2|$ at $d = 60$, panel (b) $\max_i |\widehat{q}_i - \|\beta_i\|^2|$ for $d \in \{60, 240, 960\}$. The query point is fixed and random, which is legitimate because Lemma E.1 makes the residual independent of it.
- **Figure 2d**: 24 seeds, $K = 3000$, $A = I$, $\sigma_i \equiv 1$, $\log \zeta_i \sim \mathrm{U}[-s/2, s/2]$ with spread $s \in \{0, 0.5, \dots, 3\}$.
- **Figure 6**: 6 seeds, $K = 3000$, $n = 20$. Panel (a): $d = 100$, $h_i(x) = \langle \beta_i, x\rangle - \frac{s_i\lambda}{\omega^2}\sum_j \cos(\omega z_j)$ with $z = x - x_0^\star$, $\omega = 3$ and random signs $s_i = \pm 1$, so $\nabla h_i(x_0^\star) = \beta_i$ exactly and $\|\nabla^2 h_i\|_{\mathrm{op}} \le \lambda$; λ/μ is swept over $\{0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8\}$. Panel (b): $d = 400$, $\beta_i = \zeta_i(\sqrt{1-\rho}\,\xi_i + \sqrt{\rho}\,\chi_i)$ where χ_i is a shared factor within a pair of clients (*pairwise*) or common to all (*common factor*); the optimal w for the realised Q is computed by projected gradient on the simplex.

Reported statistics. Trajectory plots show medians and interquartile bands over seeds. Log–log slopes are least-squares fits over the plotted grid. Figure 5(a) is fitted over the last nine grid points, to avoid the small-K regime where the quadratic remainder of Lemma 4.2(i) dominates.

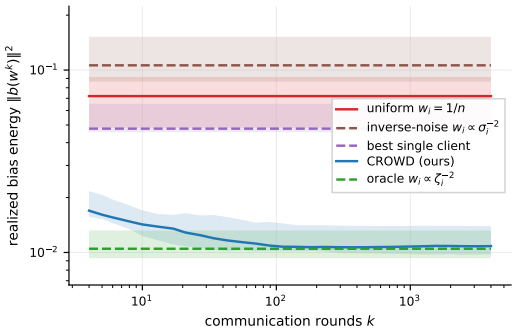


Figure 4: Realised bias energy $\|b(w^k)\|^2$ against the round index on the default instance ($n = 20$, $d = 150$, quadratic f_0 with spectrum geometric in $[0.5, 5]$, $\log \zeta_i$ and $\log \sigma_i$ uniform with $\zeta_i \in [0.2, 3]$ and $\sigma_i \in [0.5, 3]$, $b = 1$; medians and interquartile bands over 12 seeds), all rules taking identical stochastic gradients and differing only in w. CROWD reaches the oracle level within 3.4% in a few tens of rounds; the oracle-selected best single client is $4.6\times$ worse, as Proposition D.5(ii) implies. The companion excess-risk trajectories, with the exact floors $\frac{1}{2} b(w)^\top A^{-1} b(w)$, are Figure 2a.

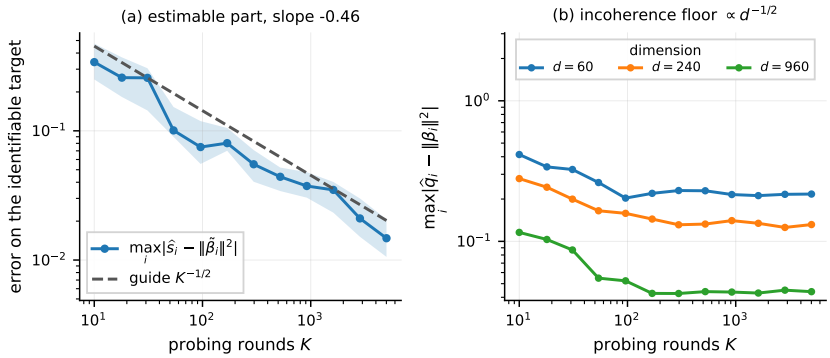


Figure 5: Estimator accuracy on the default instance (12 seeds), error against the number of probing rounds K_0 on log–log axes. (a) On the identifiable target $\|\widetilde{\beta}_i\|^2$ the error decays at the predicted rate: measured log–log slope -0.46 against the $-1/2$ of Lemma 4.2(i). (b) On q_i the error stalls at the incoherence floor of Lemma F.1, which does not decay with K at all; the floor falls by $4.9\times$ when d grows by $16\times$, against the $4.0\times$ that the $d^{-1/2}$ of Lemma F.2 predicts. The two individual $4\times$ steps give $1.6\times$ and $3.0\times$ against $2.0\times$ each, so the floor is of the predicted size and direction but not cleanly power-law at these dimensions.

What the baselines are, under their usual names. Three of the rules of Section 5 implement published families. *Inverse-noise* is precision-weighted federated learning (Reyes et al., 2021; Tölle et al., 2024). *No-inversion* is the scoring step of truth discovery (Chen et al., 2026), which by Lemma F.1 estimates $(Tq)_i$ rather than q_i. *Common-ζ^2* is the DerSimonian & Laird (1986) fit. SCAFFOLD, FEDPROX and their relatives change the optimiser rather than the weights, and by Lemma D.2(iii) every convergent optimiser stalls at the floor set by w. Figure 8 checks this with $E = 5$ local steps per round (8 seeds). With uniform weights, FEDAVG, FEDPROX and SCAFFOLD all end at 8.65 times the oracle floor, against 8.67 for the exact floor of uniform weighting. With CROWD's weights, computed from the first local gradient of each round, FEDAVG and SCAFFOLD both end at 1.05. We also run a quadratic crowd whose client Hessians differ by centred perturbations of norm 0.4μ, with a constant local step so that client drift is present. There the three optimisers end at 8.63–8.66 with uniform weights, and both end at 1.07 with CROWD's weights. The drift-correcting optimiser and the weights act on different errors, and they compose.

H.1 DIAGNOSTICS: WHEN IS A CROWD INCOHERENT, AND WHAT BREAKS THE ESTIMATOR

(a) Incoherence is a property of the heterogeneity mechanism. Assumption A4 is the one hypothesis of this paper that concerns the realised instance rather than the model, so the question

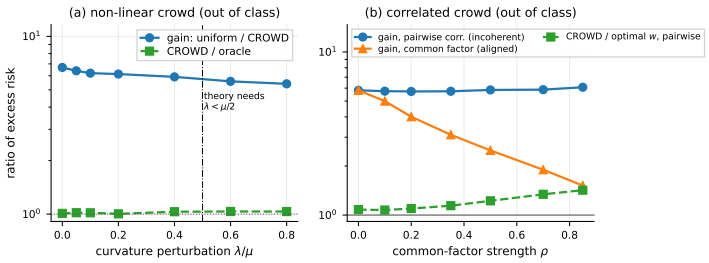


Figure 6: CROWD applied *outside* its class: $n = 20$, $K = 3000$, 6 seeds, all rules taking identical gradients. *(a)* Non-linear crowds $\nabla h_i(x) = \beta_i \pm \frac{\lambda}{\omega}\sin(\omega(x - x_0^\star))$: the gain over uniform weights falls only from $6.7\times$ to $5.4\times$ as λ grows to 0.8μ, and CROWD stays within 1.00–$1.03\times$ of the oracle – including where Theorem C.7 does not apply. *(b)* Correlated crowds: under pairwise correlation the gain stays between $5.7\times$ and $6.1\times$, even where κ exceeds one, and the competitive ratio stays in $[1.07, 1.42]$, while a common factor – the invisible kernel of Proposition C.4 – erodes it smoothly.

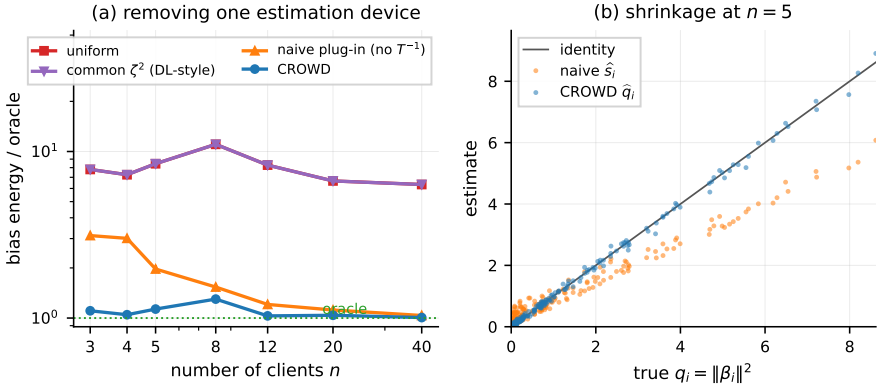


Figure 7: The contribution of each of CROWD's two estimation devices, on the default instance with the crowd size n swept over $\{3, \ldots, 40\}$ (12 seeds; medians). *(a)* Realised bias energy relative to the oracle. Fitting a *single* heterogeneity variance in the manner of random-effects meta-analysis is indistinguishable from uniform weighting at every n (7.80 versus 7.80 at $n = 3$): the gain comes from the per-client second moments. Dropping the inversion T^{-1} and using the raw disagreement energy costs $3.1\times$ at $n = 3$ and $1.04\times$ at $n = 40$, decreasing with n as C_n does. *(b)* Estimated versus true q_i at $n = 5$: the naive statistic is shrunk towards the crowd mean with measured slope 0.634 against the $1 - 2/n + 1/n^2 = 0.64$ that Lemma F.1 predicts for a fit pooled over seeds, while CROWD's slope is 0.995.

is which mechanisms produce it. We compare three, all with the same convex model (ℓ_2-logistic regression on a frozen random feature map, parameter dimension $d = 240$) and the same crowd sizes.

- *Shared pipeline, label noise.* Every client runs the same feature map on exchangeable data and differs only through its label channel. Measured κ rises from 1.96 at $n = 5$ to 21.0 at $n = 40$.
- *Site-specific instruments.* Each client additionally applies its own per-feature gain and offset before the shared map, drawn once per client – the motivating deployment, different machines at different hospitals. Measured κ falls to 0.82 at $n = 5$ and 1.66 at $n = 10$.
- *Isotropic Gaussian tilts*, the model reference: 0.19 at $n = 5$ and 0.81 at $n = 40$, consistent with Lemma D.4.

The pattern has a structural explanation, and it is the main limitation of the method. Label-channel heterogeneity is *low rank*: with c classes the channel deviations live in a space of dimension at most c^2, and all tilts factor through one design matrix, so a crowd larger than that dimension cannot be incoherent whatever d is. Instrument heterogeneity has as many independent causes as there are measured features, which pushes the threshold out. The condition $\kappa < 1$ therefore compares n with the *number of independent causes of heterogeneity*, and Lemma D.4 is the special case in which that

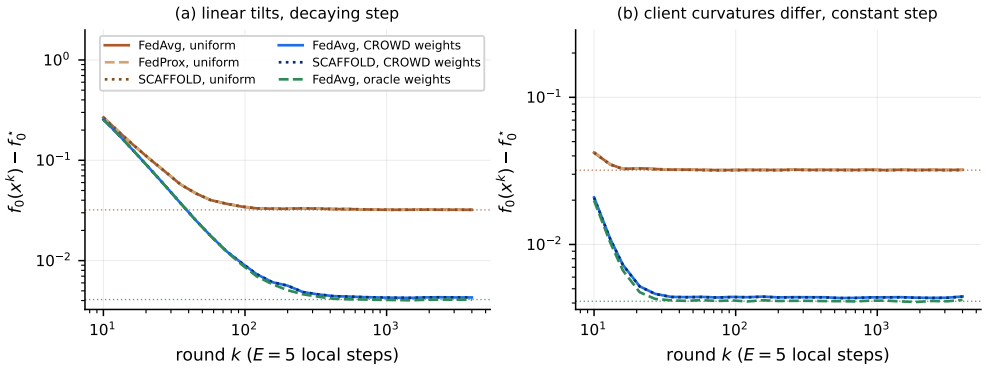


Figure 8: Optimisers with local steps against weights ($n = 20$, $E = 5$, 8 seeds, medians). Dotted lines are the exact floors of uniform (upper) and oracle (lower) weights. (a) Linear tilts, decaying step. (b) Client Hessians differ, constant local step.

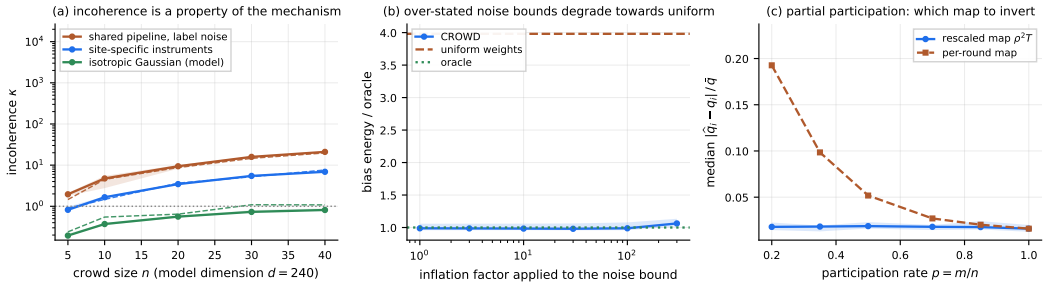


Figure 9: Diagnostics. (a) Incoherence κ of the realised tilts under three mechanisms of heterogeneity, at fixed model dimension $d = 240$ (5 seeds, medians and interquartile bands; dashed lines are the estimate $\widehat{\kappa}$ of Proposition C.1, computed from residuals alone). (b) CROWD with the noise bound $\bar{\sigma}_i^2$ inflated by a factor ρ, bias energy relative to the oracle, 12 seeds. (c) Median error of the estimated energies under uniform partial participation, inverting the rescaled map of Proposition C.9 or the per-round map, 16 seeds.

number is d. A standard benchmark with a single frozen feature map and synthetic label noise sits at the wrong end of this scale, which is why we do not present one as supporting evidence.

The dashed curves are the estimate $\widehat{\kappa}$ of Proposition C.1, which uses no knowledge of the tilts. It tracks κ across all three mechanisms and over two orders of magnitude. For Gaussian tilts it over-states κ at the larger crowds, 1.09 against 0.73 at $n = 30$, which errs towards refusing. So the guard of Remark C.2 is implementable: a deployment can estimate which regime it is in before deciding to weight at all.

(b) Over-stated noise bounds. The bound $\bar{\sigma}_i$ is the one input a practitioner must supply, and the natural failure is to set it far too large. Inflating $\bar{\sigma}_i^2$ by up to $100\times$ leaves the bias energy within 2% of the oracle, and a $300\times$ inflation costs 6%, against $3.98\times$ for uniform weighting on the same instances. An inflated noise term is common to the cost of every client up to the spread of σ_i, so it shrinks the weights towards uniform, and Proposition 4.7 bounds that loss. Under-stating $\bar{\sigma}_i$ is the dangerous direction, and there the guarantee is the one of Lemma 4.3 with Δ enlarged by the shortfall.

(c) Partial participation. At each round a uniformly drawn set of $m = pn$ clients reports, the residual is centred within it, and each client's two half-averages are taken over the rounds in which it participated. Inverting the rescaled full-crowd map $\rho^2 T$ of Proposition C.9 gives a median error of 0.016 at $p = 1$ and between 0.018 and 0.019 for every p down to 0.2, so at this probing budget the loss from partial participation is small. Inverting instead the map of the per-round squared disagreement, which also contains the variance of drawing the participants, leaves an error that grows

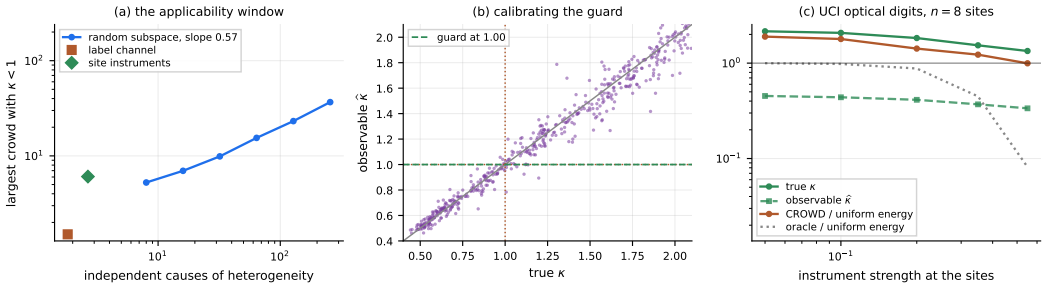


Figure 10: The applicability window. (a) The largest crowd for which the realised κ still lies below one, against the number of independent causes of heterogeneity, for tilts drawn from a random subspace of that dimension (7 seeds); the two markers are the label-channel and site-instrument mechanisms of Section H.1, placed at their measured effective rank. (b) Estimated $\widehat{\kappa}$ against true κ on 541 instances with $n = 20$, $d = 240$ and 2500 probing rounds, in the linear regime. (c) UCI optical digits with $n = 8$ sites and per-site instrument distortion of varying strength (4 seeds): true and estimated incoherence, and the bias energy of CROWD and of the oracle weights relative to uniform weighting.

as participation falls, 0.020, 0.027, 0.052, 0.098 and 0.193 at $p = 0.85, 0.7, 0.5, 0.35, 0.2$. That error is a bias of the per-round map and does not decrease as K_0 grows.

H.2 THE APPLICABILITY WINDOW, THE GUARD, AND A RUN ON REAL DATA

(a) How wide is the window. Section H.1 argued that Assumption A4 is a condition on crowd size against the number of independent causes of heterogeneity. Panel (a) turns that into a number. For tilts drawn from a random subspace of dimension r – exactly r independent causes – the largest crowd with $\kappa < 1$ grows from 5.3 at $r = 8$ to 36.7 at $r = 256$, with fitted exponent 0.57. The two mechanisms sit at the left of that curve: measured effective rank 1.8 for label-channel heterogeneity with threshold $n^\star = 1.5$, and 2.7 for site instruments with $n^\star = 6.1$. The window is therefore narrow: with the site instruments of Section H.1 incoherence already fails at ten clients, and the whole-array federation of Section H.3 keeps $\kappa < 1$ up to sixteen sites and loses it by twenty-two. A federation in the regime the theory describes needs heterogeneity with many independent causes, and identifying such a federation is the practical question this paper leaves open.

(b) How tightly the guard can be set. The decision $\widehat{\kappa} < 1$ is taken where the estimate matters most, so panel (b) sweeps instances with κ near one. In the linear regime the mean absolute error of $\widehat{\kappa}$ for $\kappa \in [0.8, 1.2]$ is 0.048, its ninetieth percentile is 0.11, and the probability that $\kappa \geq 1$ given $\widehat{\kappa} < 1$ is 4.2%. No margin is needed there: the plain threshold already holds the false-accept rate below five per cent.

(c) Real data, and a failure of the guard. We run the protocol on UCI optical digits (1797 real scans, ten classes). One frozen random feature map is shared by all sites, the model is ℓ_2-logistic regression, and there are eight sites of unequal size. Each site has its own instrument: its scanner is miscalibrated on its own subset of pixels, a minority of sites badly. Sweeping the distortion strength gives the following.

- *Incoherence fails throughout.* The realised κ runs from 2.16 at the weakest distortion to 1.34 at the strongest, never below one. The crowd is small ($n = 8$) and the parameter space large, so this is not a dimension effect; the tilts share the digit structure.
- *The oracle gain is large.* The ratio AM/HM of the realised energies grows from 1.42 to 11.85, and the oracle weights reach 0.08 times the uniform bias energy at the strongest setting.
- CROWD *does not realise it.* CROWD attains $1.90, 1.79, 1.42, 1.23, 1.00$ times the uniform bias energy across the sweep – never better than uniform. This is what Lemma 4.2(iii) predicts once $\kappa > 1$: the floor $3\kappa\bar{q}$ exceeds the spread of the q_i, and the inverted statistic carries no usable signal.
- *The guard fails here.* The estimate $\widehat{\kappa}$ reports 0.45 down to 0.34 across the sweep, against a true κ between 2.16 and 1.34. This is an under-estimate by a factor three to five, on the wrong side

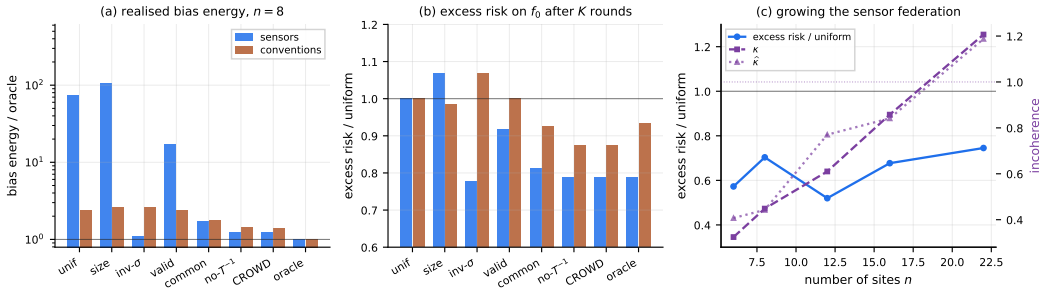


Figure 11: UCI optical digits under two constructed federations ($n = 8$, 8 seeds, medians). (a) Realised bias energy relative to the oracle weights, log scale. (b) Excess risk on f_0 after $K = 2500$ rounds, relative to uniform weighting; below one is a gain. (c) The same ratio as the sensor federation grows (4 seeds per n), with the measured and estimated incoherence on the right axis.

of the threshold. The cause is the staleness of Lemma 4.2. On a strongly non-linear crowd the residual measured at the iterate is not the disagreement at $x_0^\star$, and the same bias that inflates δ_i deflates the off-diagonal Gram entries. Proposition C.1 is proved for the linear crowd, and panel (b) checks it there. On this non-linear crowd $\widehat{\kappa}$ under-reports by a factor we do not control, and Remark C.2 must be read with that restriction.

- *The standard construction is worse.* With an IID split and label noise as the only heterogeneity, the same data give $\kappa = 2.25$, AM/HM $= 1.23$, and CROWD at 2.36 times the uniform energy: the oracle gain is small and the estimate is uninformative.

The three requirements of the method pull against each other on a small benchmark. It needs many independent causes, so that $\kappa < 1$; a wide spread of site quality, so that AM/HM is large; and perturbations small enough for the first-order model to hold. On UCI digits with eight sites we could not satisfy all three at once, and the configuration above is reported as it stands.

H.3 THREE FEDERATIONS ON THE SAME REAL SCANS

What is real and what is constructed. The scans are real: 1797 scanned 8×8 images, one frozen random feature map shared by all sites, ℓ_2-regularised logistic regression, so the learning problem is the convex one the theory assumes. The *federation* is not found in nature. We partition one dataset into sites and impose a distortion on each. The independence of the tilts across sites, their vanishing crowd mean and the spread of quality are consequences of that construction, not measurements of an observed federation. So is f_0: we define it as the risk averaged over forty draws from the population law of the distortions. Assumption A1 then holds by construction, up to the Monte Carlo error of those draws, and excess risk is evaluated exactly on them. The study therefore tests a conditional claim: that the method behaves as predicted *if* heterogeneity is arranged this way. Whether federations with $\kappa < 1$ occur in practice is a question a constructed federation cannot settle. On real data CROWD and the oracle weight by $\widehat{q}_i + \sigma_i^2/n_i$ and $q_i + \sigma_i^2/n_i$, the second term being the variance of a site's mean gradient over its cohort, in place of the $\bar{\sigma}_i^2/(bK)$ of Algorithm 2. Three constructions follow, on the same scans and the same model.

Mechanism	κ	$\widehat{\kappa}$	AM/HM	size	inv-σ	valid.	CROWD
Few pixels + label channel	1.34–2.16	0.34–0.45	1.4–11.9	—	—	—	$\geq 1^\dagger$
Whole sensor arrays	0.48	0.52	86.9	1.069	0.778	0.918	**0.787**
Annotation conventions	0.48	0.53	2.2	0.985	1.069	1.000	**0.875**

Table 2: The three federations built on the same 1797 real scans. The last four columns are the *excess* risk on f_0 after $K = 2500$ rounds, relative to uniform weighting: below one is a gain, and the oracle weights reach 0.789 and 0.934 in the second and third rows. Row one is the failure of Section H.2(c); the other two are described below. †Bias energy relative to uniform weighting, 1.00–1.90; those runs do not record the risk.

Whole sensor arrays: incoherence holds and the estimator is accurate. Each site has its own detector array, miscalibrated across *all* of its pixels with a per-pixel gain and offset, and sites differ in

severity over two decades. This is the mechanism the motivating application is usually *described* as having – different machines at different hospitals – reproduced on real images. The realised incoherence is $\kappa = 0.48$ and the estimate $\widehat{\kappa} = 0.52$, so the guard of Remark C.2 says proceed, correctly. The crowd-hypothesis ratio $\|\bar{\beta}\|^2/(\bar{q}/n)$, close to one for a centred independent crowd, is 1.00; the Spearman correlation between $\widehat{q}$ and the realised energies is 0.95, and $\mathrm{AM}/\mathrm{HM} = 86.9$. Uniform weighting sits at 72.9 times the oracle bias energy and CROWD at 1.22.

We report *excess* risk $f_0(x) - f_0^\star$, not total risk: the total contains $f_0^\star$, which no weighting can change, and ratios of totals understate every effect. The end-of-run ratios are those of Section 5; at the limit points of the weighted objectives, computed exactly, CROWD and the oracle reach 0.53 and 0.54 of the uniform excess. This federation cannot show the separation of contribution (1): a badly miscalibrated sensor inflates its site's gradient variance as well as its bias, so inverse-noise weighting lands level with CROWD (0.52 at the limit). The mechanism confounds the two costs.

Annotation conventions: bias and noise pulled apart. To separate them we need a source of bias that does not scramble the covariates, and cohort sizes that do not follow it. Every site now carries a detector array of the *same* moderate severity, which supplies the independent causes and keeps the gradient variances comparable. Sites differ instead through their annotation convention: a site-specific confusion channel applied to the labels, with strength spread over two decades. The cohort sizes are then assigned *against* quality, so the most badly conventioned site receives the largest cohort. Such a site is quiet and biased, and every rule that reads precision must prefer it. Precision weighting is then *worse than not weighting at all* and CROWD beats the oracle (Section 5); size-proportional weighting reaches 0.985, the no-inversion variant 0.874 and the common-ζ^2 fit 0.926. At the limit points the ordering is the same: CROWD 0.86, the oracle 0.92, inverse-noise 1.01. The oracle weights minimise the realised bias energy, a first-order proxy for the excess risk, and on this instance the proxy is loose (Section H.4). The federation lies inside the region $\kappa < 1$, with $\kappa = 0.48$ and $\widehat{\kappa} = 0.53$, so the guard says proceed. The result is consistent with contribution (1): the rule that reads precision does worse than uniform weighting, and CROWD does better.

Accuracy. Held-out accuracy, measured on clean labels, moves with the risk in both federations. It rises from 0.8469 under uniform weighting to 0.8570 under CROWD in the sensor federation, and from 0.5632 to 0.5884 in the convention one, where inverse-noise weighting falls to 0.5231. The theorems bound the risk under f_0, not the accuracy, so we treat accuracy as corroborating evidence only.

The guard on these federations. Both give $\widehat{\kappa}$ on the correct side of the threshold, 0.52 against $\kappa = 0.48$ on the sensor federation and 0.53 against 0.48 on the convention one. The drift check of Remark C.3 raises no alarm on either probe.

Validation-tuned weights. The only validation data a federation has are those of its sites. In that baseline each site scores the consensus model on its own held-out split, weights are taken proportional to a power of the inverse score, and the exponent is chosen by the pooled held-out loss. In the convention federation it selects the uniform exponent and returns exactly uniform weighting, 1.000; in the sensor federation it reaches 0.918, against 0.787 for CROWD. This follows from Proposition 3.1(i): held-out data drawn from a site's own tilted distribution measure how well the model fits *that site*, and no sample from f_0 exists to measure anything else. Validation tuning therefore detects sites that the model fits poorly, which need not be the sites with large tilts.

Beyond $\kappa < 1$. Section H.4 sweeps the common detector severity and finds CROWD below uniform weighting on both sides of $\kappa = 1$, at 0.851 when $\kappa = 1.28$ as well as at 0.832 when $\kappa = 0.37$. Incoherence is what makes Proposition 4.7 non-vacuous; the estimator itself needs less, and the floor $3\kappa\bar{q}$ of Lemma 4.2(iii) is worst case over sign patterns. Closing the gap between the condition the theory uses and the condition the estimator needs is left open.

How far the crowd can grow. Panel (c) of Figure 11 grows the sensor federation on the same data. The excess risk relative to uniform weighting is 0.57, 0.70, 0.52, 0.68 and 0.75 at $n = 6, 8, 12, 16, 22$, while the measured κ rises through 0.32, 0.45, 0.61, 0.86 and 1.21. The gain persists over the whole range, including $n = 22$ where $\kappa > 1$, and with four seeds per point the series is too noisy to locate a threshold. The evidence that the window closes is the incoherence measurement itself, here and in

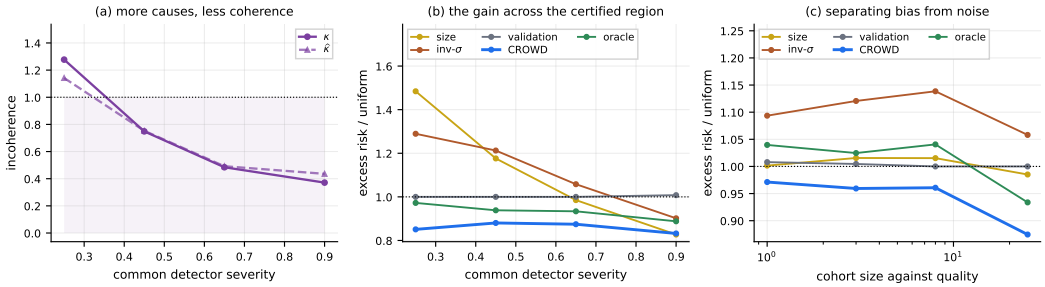


Figure 12: Sweeping the two knobs of the convention federation ($n = 8$, 6 seeds, medians). (a) The common detector severity controls incoherence, the shaded band marking the region $\kappa < 1$. (b) Excess risk relative to uniform weighting along that axis. (c) The same against the cohort-size contrast, which controls how far bias and sampling noise are pulled apart. Below one is a gain.

Section H.2(a). The intended scale is a federation of a few to a dozen sites with physically distinct instruments; device-scale federated learning lies outside it.

H.4 The gain over a region of settings

The federation of Section H.3 is built rather than found, so what needs testing is whether the gain holds over a region of settings. We sweep the two knobs the theory says should matter, keep everything else fixed, and fix the predictions in advance. The *common detector severity* is the same at every site, so it creates no quality spread; what it supplies is independent causes of heterogeneity, which should push κ down. The *cohort-size contrast* is how strongly the cohort sizes run against quality, which should decide whether a precision rule can see the bias at all.

Along the incoherence axis. Raising the common severity from 0.25 to 0.90 takes κ from 1.28 to 0.37 and the estimate $\widehat{\kappa}$ from 1.14 to 0.44, so the estimate tracks the crossing. CROWD reaches $0.851, 0.880, 0.875, 0.832$ of the uniform excess risk along that axis: a gain at every point, on both sides of $\kappa = 1$, and below the oracle diagonal weights $(0.972, 0.938, 0.934, 0.888)$ at every point. Precision-weighted inverse-noise reaches $1.289, 1.213, 1.058, 0.902$ and size-proportional weighting $1.484, 1.176, 0.985, 0.827$. Both lose at the smaller severities and gain at the largest, where the detector distortion, of the same scale at every site, outweighs the conventions, so that large cohorts are no longer the badly biased ones.

Along the decoupling axis. With the severity fixed at 0.65, so that κ stays between 0.46 and 0.52 throughout, and the cohort contrast swept from 1 to 25, CROWD gives $0.971, 0.959, 0.961, 0.875$ and inverse-noise $1.094, 1.121, 1.139, 1.058$. The two sweeps share one setting, so they cover seven distinct ones. CROWD improves on uniform weighting at all seven and inverse-noise weighting is worse than uniform at six. The separation of contribution (1) therefore holds over a region of the mechanism's parameters.

Validation-tuned weights. Along both axes validation tuning returns essentially uniform weights, at 1.000 to 1.008 of the uniform excess risk. A site's own held-out loss measures how well the consensus model fits that site, and by Proposition 3.1(i) no held-out sample from f_0 exists to measure anything else.

Where the oracle weights lose. The oracle diagonal weights do worse than CROWD at every setting, and worse than uniform weighting at the three smaller contrasts (1.025 to 1.041). They minimise the realised bias energy, which is a first-order proxy for the excess risk: it ignores the curvature of f_0, the curvature differences between sites, and the error in the sampling term σ_i^2/n_i. On a real instance the error of this proxy can exceed the error of the estimator.

H.5 FEMNIST: writers as a natural federation

The federations of Sections H.3 and H.4 are constructed. Federated EMNIST supplies a crowd without construction: each client is one real writer, and nothing is distorted or relabelled.

Setup. We use the digits of 1000 randomly chosen writers of the release, about 100 training digits each, one frozen random ReLU layer $784 \to 128$ shared by all writers, and ℓ_2-regularised logistic regression with coefficient 10^{-3}. In each seed 400 writers form a pool whose writer-averaged risk is f_0, and n other writers are the clients, each with the empirical risk of its own training digits. Probing takes $K_0 = 2000$ rounds with batch 16 after a burn-in of 600, and training takes $K = 3000$ rounds under each rule with identical seeds. The excess risk is evaluated exactly on f_0. CROWD and the oracle add the noise term $\sigma_i^2/(bK)$ of Algorithm 2. We run $n \in \{10, 20, 40, 80\}$ with 12 seeds each.

Why the gain of natural writers fades with n. The first four rows of Table 3 and Figure 2c show the natural federation. With ten writers $\kappa = 0.52$, and CROWD reaches 0.897 of the uniform excess risk, as does the oracle; it beats uniform weighting in all twelve seeds. The gain then fades: 0.922 at $n = 20$, 0.991 at $n = 40$, and at $n = 80$ every weighting rule, the oracle included, is about 10% worse than uniform weighting. As n grows, two effects combine. The excess risk of uniform weighting halves with each doubling of n (the "uniform excess" column), as the bias energy AM/n of the uniform crowd does, so the weights act on a shrinking part of the risk. And κ roughly doubles, from 0.52 to 4.08, so the estimate of the energies degrades. The first effect is why even the oracle loses its gain, as the amplified series below confirms.

Keeping the bias from vanishing. To separate the two effects we amplify each writer's style. Every digit is moved by $(\alpha - 1)$ times the difference between its writer's mean image for that class and the pooled class mean. The writers' mean deviation stays zero, so the crowd remains centred, while every tilt grows with α. The incoherence would be unchanged by a pure rescaling of the tilts; here it changes little up to $n = 40$ (1.10 against 1.09, 1.97 against 2.07). With $\alpha = \sqrt{n/10}$ the uniform crowd's excess risk decays much more slowly, from 0.70 at $n = 10$ to 0.28 at $n = 80$ instead of 0.10. The gain persists (Figure 2c): CROWD reaches 0.922 at $n = 20$ and 0.929 at $n = 40$, beating uniform weighting in eleven and ten of twelve seeds, although $\kappa \approx 2$ at $n = 40$. At $n = 80$, $\kappa = 3.39$ and the estimate stops working: the oracle still reaches 0.971, but CROWD reaches 1.027 and beats uniform weighting in five of twelve seeds.

Writers	α	κ	$\widehat{\kappa}$	uniform excess	inv-σ	CROWD	oracle	CROWD $<$ unif.
10	1	0.52	0.50	0.70	0.904	**0.897**	0.897	12/12
20	1	1.09	0.92	0.38	0.946	**0.922**	0.909	10/12
40	1	2.07	1.87	0.19	0.973	0.991	0.980	6/12
80	1	4.08	3.54	0.10	1.106	1.112	1.102	0/12
20	1.41	1.10	0.93	0.51	0.925	**0.922**	0.891	11/12
40	2	1.97	1.75	0.38	0.930	**0.929**	0.920	10/12
80	2.83	3.39	3.03	0.28	1.010	1.027	0.971	5/12

Table 3: FEMNIST, medians over 12 seeds. The last four columns are the excess risk on f_0 relative to uniform weighting and the number of seeds in which CROWD beats uniform weighting; "uniform excess" is the absolute excess risk of uniform weighting. Size-proportional weighting, validation tuning and the common-ζ^2 fit stay within 0.01 of uniform weighting in every row.

The trade-off. The weights help when three conditions hold together. First, the crowd's summed tilt must vanish slowly as n grows, so that the bias term is a sizeable part of the risk. Second, the clients must differ in quality, here only mildly ($\mathrm{AM}/\mathrm{HM} \approx 1.7$ to 2.3, against 86.9 in the sensor federation). Third, the crowd must not be much larger than the number of independent causes of heterogeneity, or the estimate fails. On FEMNIST the first condition limits natural writers to about twenty, and amplification extends the window to forty, where the third condition takes over. The theory's sufficient condition $\kappa < 1$ is conservative here: CROWD keeps its gain up to $\kappa \approx 2$.

Noise and accuracy. Inverse-noise weighting tracks CROWD throughout. A writer whose digits are hard to fit has both a larger gradient variance and a larger tilt, as with the miscalibrated sensors of Section H.3, so FEMNIST does not separate bias from noise; only the convention federation does. At the limit points of the weighted objectives every ratio changes by less than 0.02. Accuracy on the pool writers' test digits moves with the risk, from 0.729 under uniform weighting to 0.734 under CROWD at $n = 10$ (oracle 0.741). The guard is wrong at $n = 20$, where $\widehat{\kappa} \approx 0.92$ against $\kappa \approx 1.1$, the under-reporting of Section H.2(c); in these runs CROWD still gains there.

H.6 NEURAL NETWORKS OUTSIDE THE THEORY

Every guarantee of the paper assumes a strongly convex truth. To see what happens without it, we train a two-layer network, $784 \to 64 \to 10$ with ReLU and ℓ_2-coefficient 10^{-4}, end to end on the raw FEMNIST pixels. The clients are the writers of Section H.5 (batch 16, step 0.3, burn-in 600, $K_0 = 2000$ probing rounds, $K = 3000$ training rounds). The truth f_0 is again the writer-averaged risk of 400 pool writers. It is non-convex, so we report $f_0(x) - f_0(x_{\mathrm{ref}})$ relative to uniform weighting, with x_{ref} a network trained on the pool by minibatch SGD from the common initialisation. The oracle weights by the energies of $\nabla f_i - \nabla f_0$ at x_{ref}, which is a local reference and not a target the theory defines.

The risk ratios, incoherences and win rates are those of Table 1 in the main text (8, 8 and 6 seeds for the three rows). CROWD lowers the risk in every setting, by 5 to 11% in the median, although the incoherence of the tilts at the reference point exceeds one throughout. Inverse-noise weighting and the oracle do well with ten writers and poorly with forty amplified ones, where they are 43% and 33% worse than uniform weighting. CROWD measures the disagreement along its own trajectory rather than at a fixed point, and it is the only rule that improves on uniform weighting in all three settings. Accuracy on the pool writers' test digits moves by less than one point in either direction ($0.807 \to 0.815$, $0.862 \to 0.859$ and $0.852 \to 0.846$ under CROWD). We read this as a stress test: the mechanism carries over to a non-convex model, and the size of the effect is modest.

A convolutional network on the digit scans, and the edge of the window. To move further from the theory we replace the frozen feature map of the sensor federation (Section H.3) by a network trained end to end on the raw scans. It has a conv 3×3 layer with eight channels, ReLU, 2×2 max-pooling and a linear layer, with ℓ_2-coefficient 10^{-3}, and is trained on each site's distorted 8×8 images ($n = 8$, batch 32, step 0.15, burn-in 600, $K_0 = 2000$, $K = 2500$, 6 seeds). The step was fixed on a pilot seed outside the evaluation set. The truth is defined as in Section H.3, on 40 population draws, and the protocol is that of the two-layer network above, with the reference θ_{ref} trained on the population draws. The energies at the reference spread over three decades ($\mathrm{AM}/\mathrm{HM} \approx 2 \cdot 10^3$), and the realised incoherence is $\kappa = 1.63$ in the median. The cross-fitted statistic is uncorrelated with the realised energies (Spearman -0.15). No diagonal rule improves on uniform weighting (Table 1). The no-inversion and common-variance rules reach 1.06, as CROWD does, the fixed-point oracle is again the worst rule, and accuracy changes by less than one point. The guard of Remark C.2 declines in five of six seeds and would return uniform weighting, which is the better choice on this instance. On the two-layer network it declines as well, in 20 of 22 seeds, although unguarded CROWD improves on uniform weighting there. On these models the guard is therefore conservative: it does not separate the case where weighting helps from the case where it does not.