VERTICAL FEDERATED LEARNING AS A SOCIAL CHOICE PROBLEM

Anonymous authors
Paper under double-blind review

ABSTRACT

Centralized data aggregation in modern machine learning increasingly conflicts with privacy laws, security risks, and settings where data pooling is impossible. Federated learning addresses this by bringing models to the data. Unlike horizontal federated learning (HFL), where agents share a feature space across distinct individuals, many real-world scenarios require the opposite: agents hold distinct feature sets for overlapping populations. This decentralized setting defines vertical federated learning (VFL). To enable a VFL central server to aggregate multi-agent predictions without accessing raw data or internal models, we propose a voting-theory-based framework satisfying strict privacy and minimal-invasiveness constraints. In this approach, agents submit round-wise predictions (or "votes") based on their local features, which the server aggregates into a final decision. We prove that any stationary aggregation rule incurs linear regret, motivating a dynamic generalization where agent weights evolve as evidence accumulates. Consequently, we introduce Trust-Based Perpetual Voting (TBPV), adapting classical Perpetual Voting for accuracy-driven tasks by concentrating influence on agents the server trusts most. This is primarily a theoretical contribution: our experiments are a proof of concept, intended to check that the mechanism behaves as the theory predicts on controlled vertical partitions, rather than to establish state-of-the-art performance.

1 INTRODUCTION

Federated learning (FL) addresses data privacy constraints by bringing computation to client devices (McMahan et al., 2017). Unlike horizontal FL, vertical federated learning (VFL) arises when agents hold distinct feature sets for overlapping populations (Liu et al., 2024). Existing VFL approaches typically rely on intermediate gradient exchanges (Yang et al., 2019), which mandate compatible model architectures and expose systems to gradient inversion attacks (Zhu et al., 2019). Aggregating predictions without internal model or data access requires overcoming three distinct hurdles: the strategy must be strictly output-only, it must adapt to honest agent heterogeneity without observable ground truth at decision time, and it must optimize cooperative combination rather than simple adversarial filtering (Blanchard et al., 2017).

We formulate output-only VFL aggregation as a social choice problem using agent predictions as "votes." However, standard social choice mechanisms prove insufficient: every stationary aggregation rule incurs linear regret. To solve this, we introduce Trust-Based Perpetual Voting (TBPV). By extending perpetual voting to track performance, TBPV dynamically updates agent weights as evidence accumulates, and, under heterogeneity and a concentration condition, its fixed point provably beats the approval fixed point in collective accuracy—all while preserving agent privacy and architecture independence.

Related Work: Our framework addresses three simultaneous requirements: *output-only* access, adaptation to *heterogeneous* honest agents, and *vertical* feature partitioning. Existing work satisfies at most two. Dynamic FL aggregation methods like FedAWA (Shi et al., 2025) and FedAPA (Sun et al., 2025) adapt client weights, but depend on model parameters or gradients. Output-based trust mechanisms like FLARE (Wang et al., 2022) and EigenTrust (Kamvar et al., 2003) focus on penalizing malicious actors rather than optimizing for honest, heterogeneous agents. FL voting methods—such as FedVote (Yue et al., 2024), FedQV (Chu & Laoutaris, 2024), and majority/Borda

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.

rules (de Camargo et al., 2022)—rely on horizontal partitioning and static, non-adaptive rules. Finally, game-theoretic VFL approaches (Tu et al., 2022) target contribution attribution and carry high computational costs (Wang, 2019). No existing method combines all three requirements.

2 SETUP

What Happens Each Round: There are n agents, indexed $i \in [n]$, and a central server S. Each round t proceeds in three steps. First, a fresh labelled example $(x^t, y^t) \sim \mathcal{D}^t$ is generated, where $x^t = (x_1^t, \ldots, x_n^t)$ is split across the agents: agent i observes only its own slice $x_i^t \in \mathcal{X}_i$, never the slices held by the other agents, and no one observes the label y^t yet. Second, each agent, based only on what it has seen so far, produces a soft prediction $\hat{p}_i^t \in \Delta(\mathcal{Y})$, a probability distribution over the possible labels, without seeing any other agent's prediction. Third, the server collects all n predictions $(\hat{p}_1^t, \ldots, \hat{p}_n^t)$ and combines them into a single output $\hat{y}^t \in \mathcal{Y}$. The true label y^t is never revealed: neither the agents nor the server observes it at any point in the protocol. Each agent updates its model toward the collective outcome $\hat{y}^t$, and the server updates its weights from agreement with that same outcome. Ground-truth labels are used only for the purpose of determining method accuracy.

Measuring Performance: We are primarily concerned with whether the server's final output matches the true label; that is, the *loss* at round t, $\ell^t(\hat{y}^t, y^t) = \mathbf{1}[\hat{y}^t \neq y^t]$, which we can evaluate offline but the server cannot observe. Each agent also has its own *individual loss*, $\ell_i^t = \mathbf{1}[\hat{y}_i^t \neq \hat{y}^t]$, where $\hat{y}_i^t = \arg\max_k \hat{p}_i^t(k)$ is that agent's own best guess; this is measured against the collective outcome rather than the label, since the label is unavailable, and it only affects how that agent updates itself, playing no direct role in the server's objective.

Decision Rules: We call the rule an agent uses to turn what it has observed into a prediction its *strategy*: a map $\sigma_i : \mathcal{H}_i \times \mathcal{X}_i \to \Delta(\mathcal{Y})$, such that given everything agent i has seen up to round t (h_i^t) and its current input x_i^t, it outputs $\hat{p}_i^t = \sigma_i(h_i^t, x_i^t)$. The server has an analogous strategy of its own: a map $\sigma_S : \mathcal{H}_S \times \Delta(\mathcal{Y})^n \to \Delta(\mathcal{Y})$, such that given everything it has seen so far (h_S^t) and the current round's predictions, it outputs $\hat{y}^t \sim \sigma_S(h_S^t, \hat{p}_1^t, \ldots, \hat{p}_n^t)$.

This paper is concerned with σ_S. For this reason we fix each agent's strategy to be an online maximum-likelihood estimator, so that agents naturally output soft, calibrated predictions rather than hard labels; this choice is made purely for analytical convenience, since the server's aggregation problem does not depend on how the agent predictions were produced in the first place.

To compare different choices of σ_S, we ask how much worse a rule's cumulative loss is than that of the best *fixed* weighting $w^* \in \Delta([n])$ of the agents' predictions, chosen with the benefit of hindsight; we call this its *linear pool regret*,

$$R_{\text{pool}}^T(\sigma_S) = \sum_{t=1}^{T} \ell^t\big(\hat{y}^t, y^t\big) - \min_{w \in \Delta([n])} \sum_{t=1}^{T} \ell\big(\textstyle\sum_{i=1}^n w_i \, \hat{p}_i^t,\ y^t\big)$$

A rule whose regret grows linearly in T is, on average, no better than some fixed (but badly-chosen) weighting of the agents; we call a rule *no-regret* if $R_{\text{pool}}^T / T \to 0$. Regret is an evaluation criterion rather than a quantity the server can compute: the comparator requires labels, which the protocol never provides. We use it to show what stationary rules cannot achieve (Proposition 3.3), which motivates history-dependent weights; no label-free rule is no-regret against every distribution (Remark A.1), so we do not claim this for TBPV.

3 VOTING RULES AS SERVER STRATEGIES

An aggregation rule for VFL is a map $f : \Delta(\mathcal{Y})^n \to \mathcal{Y}$ mapping the participating clients' soft predictions to a single output each round. Rather than focusing on methods which inevitably impose structural conditions on the data, we take a different approach and ask: what properties should a good aggregation rule satisfy, independently of the underlying feature distribution? Framing the problem this way reveals a natural analogy. Each round, a set of clients submits a prediction over labels, and the server must produce a single decision. This is precisely the setting of social choice theory, where the question of how to aggregate individual preferences into a collective decision has

been studied axiomatically for centuries. Rather than proposing a specific rule ad hoc, we ask which rules are consistent with a set of desired axiomatic properties.

The three most natural properties we would want any aggregation rule to satisfy are:

- **Anonymity:** the output depends only on the multiset of predictions, not on which client submitted which prediction.
- **Neutrality:** no label is privileged *a priori*; permuting the labels permutes the output accordingly.
- **Population Consistency:** if two disjoint groups of clients independently produce the same output, the combined group produces that output too — ensuring the rule behaves coherently even when not every client participates every round.

These are not strong assumptions; they amount to saying the rule treats clients and labels symmetrically, and behaves coherently when groups are combined. Yet together they are already quite restrictive.

Theorem 3.1 (Young (1975), Theorem 1(i))**.** *A social choice function is symmetric* (*anonymous and neutral*) *and consistent if and only if it is a scoring rule — that is, it can be represented by a sequence of scoring vectors used to rank candidates and break ties.*

A *simple scoring rule* is defined by a single vector $\phi = (\phi_1, \ldots, \phi_m)$ with $\phi_1 \geq \cdots \geq \phi_m$, assigning each label the score $s_\phi(c) = \sum_{i=1}^{n} \phi_{\mathrm{rank}_i(c)}$ and selecting $\arg\max_c s_\phi(c)$; a general scoring rule uses a sequence of such vectors to break ties. We restrict attention to simple scoring rules and identify f with its scoring vector.

Requiring anonymity, neutrality, and consistency has narrowed the class of functions we may consider to simple scoring rules. Which of these should we prefer? Two further properties sharpen the answer:

- **Faithfulness:** when there is a single client, the rule's output ranking coincides exactly with that client's own ranking (so in particular, that client's top-ranked candidate is selected as the winner).
- **Cancellation:** if, for every pair of candidates $a_i \neq a_j$, the number of clients ranking a_i above a_j equals the number ranking a_j above a_i, then the rule ranks all candidates as tied.

Once again, these are natural requirements. Faithfulness ensures the rule respects an individual client's information when it is unambiguous. Cancellation is a symmetry condition: when the evidence for every pair of candidates is exactly balanced across clients, no candidate should be preferred. However, when considered in conjunction with the earlier axioms, they leave only one option.

Theorem 3.2 (Young (1975), Theorem 3)**.** *For any fixed number m of alternatives, Borda Count is the unique social choice function that is neutral, consistent, faithful, and has the cancellation property.*

3.1 Stationary Rules are Insufficient

A server strategy is *stationary* if it ignores the history h_S^t and applies the same aggregation rule at every round. Formally, $\sigma_S^f(h_S^t, \hat{p}^t) = f(\hat{p}^t)$ for some fixed function f independent of h_S^t. All single-round voting rules are stationary in this sense, as they all apply the same aggregation function regardless of what happened in previous rounds.[1]

Proposition 3.3 (Linear regret of stationary voting rules)**.** *Let $f : \Delta(\mathcal{Y})^n \to \mathcal{Y}$ be any stationary voting rule. Then there exists a data-generating distribution $\mathcal{D}$ and a constant $\delta > 0$ such that $\mathbb{E}\big[R_{\mathrm{pool}}^T(\sigma_S^f)\big] \geq \delta T$.*

Corollary 3.4. *No stationary voting rule is no-regret. Convergence to w^* requires a server strategy that uses h_S^t to update its aggregation weights across rounds.*

[1] All proofs are found in the appendix.

Together, Theorems 3.1 and 3.2 and Proposition 3.3 characterise the fundamental issue of single-round voting in VFL: Borda is the axiomatically best available option within the population-consistent class, yet even Borda cannot escape the $\Omega(T)$ regret that afflicts all stationary rules.

3.2 Perpetual Voting Preliminaries

While traditional voting rules consider decisions in isolation, Lackner (2020) introduced *perpetual voting* for sequential elections. We leverage its temporal structure for history-dependent server aggregation: at each round r, agents submit approval ballots $A_r(i)$, and a perpetual rule selects a winner w_r using the full history of past ballots and decisions. Though Young's characterisation (Theorem 3.1) applies to general rankings, our thresholded predictions form dichotomous rankings, placing approval ballots directly within this theoretical framework.

Given a choice sequence w selected by a perpetual voting rule, we describe how well a voter v has been served up to round j via its *satisfaction*, $sat_j(v, w) = |\{I \leq j : w_I \in A_I(v)\}|$ — how often the selected outcome fell inside v's approval set, $A_I(v)$. Voter v's *support* in round j, $supp_j(v) = \frac{1}{n} \max_{c \in A_j(v)} |\{u \in N : c \in A_j(u)\}|$, measures how much company v has for its most broadly held preference, and its *quota* $qu_T(v) = \sum_{k \leq T} supp_k(v)$ accumulates this over time.

Definition 3.5 (Satisfaction guarantee)**.** A perpetual voting rule satisfies the *satisfaction guarantee* with parameter $\beta_{\text{sat}} > 0$ if, for all $v \in N$,

$$\liminf_{T \to \infty} \frac{sat_T(v)}{T} \geq \beta_{\text{sat}} \cdot \liminf_{T \to \infty} \frac{qu_T(v)}{T}.$$

Perpetual voting rules (Lackner, 2020; Lackner & Maly, 2023) aim for long-term fairness by bounding individual voter satisfaction from below. However, this guarantee is inherently weak: it only matches accrued quota *asymptotically* up to a constant factor $\beta_{\text{sat}} < 1$, rather than per round. Crucially, because satisfaction measures whether a voter's *preferred* candidate won rather than the *correct* candidate, high voter satisfaction does not imply high collective decision quality.

4 Trust-Based Perpetual Voting

As we show in Theorem 5.5, the classical satisfaction criterion carries an unavoidable cost in *collective accuracy*, and no rule whose weights track accuracy can meet it in general (Corollary 5.6). Proposition 3.3 established a companion result for the static case: no stationary rule, not even Borda Count, can avoid linear regret. What should an accuracy-oriented rule look like instead? If agent quality were known, the ideal would be *dictatorial*: defer at every round to the single best voter. This is not viable online, where quality is unknown a priori and the ground truth against which it could be measured is never revealed.

We propose *trust-based perpetual voting* (TBPV) as the response. Where the dictatorial mechanism assumes agent quality is known, TBPV *estimates* it from the only evidence available: how often each agent has agreed with the collective's past decisions. Each agent accrues a trust score that rises on agreement and falls on disagreement, so trust becomes a running proxy for the quality a dictator would have used directly. We show in Proposition 5.3 that fixed-point trust is strictly increasing in accuracy, and in Theorem 5.7 that, under stated conditions on how sharply trust concentrates influence, this yields higher collective accuracy than the approval-based fixed point.

4.1 The Abstract Mechanism

A trust-based perpetual voting rule has two components: a trust-weighted scoring rule that selects a candidate at each round, and an update rule that revises trust scores based on each agent's agreement with the collective outcome.

Let $\gamma_k(v)$ denote voter v's trust score in round k. Trust scores are initialised at $\gamma_1(v) = \gamma_0$ (e.g. $\gamma_0 = 0.5$) for all $v \in N$. Given a k-decision history (N, A, C) and a decision instance (N, A_{k+1}, C_{k+1}), the rule selects the candidate $w_{k+1} \in C_{k+1}$ with the highest trust-weighted score with $sc_{k+1}(c) = \sum_{v \in N \,:\, c \in A_{k+1}(v)} \gamma_{k+1}(v)$ and $w_{k+1} = \arg\max_{c \in C_{k+1}} sc_{k+1}(c)$.

The approval sets $A_k(v)$ are the ballots our framework produces when each agent's soft prediction $\hat{p}_v \in \Delta(\mathcal{Y})$ is thresholded into the set of candidates it supports; TBPV thus instantiates the server strategy σ_S of that framework while consuming the agents' predictions in set-valued form. Reducing soft predictions to approval sets discards calibration information, but it is what lets the aggregation be analysed as a voting rule and makes Young's axiomatic guarantees (Theorem 3.1) available.

We characterise the class of trust-based mechanisms abstractly, independently of any specific update rule. A trust-based mechanism updates each voter's trust score via a function $h : [0,1] \times \{0,1\} \to [0,1]$, such that

$$\gamma_{k+1}(v) = h\big(\gamma_k(v),\, z_k(v)\big) \tag{1}$$

where $z_k(v) = \mathbb{I}(w_k \in A_k(v))$ is the consensus-agreement indicator. We require h to satisfy the following axioms.

- **(A1) — Monotonicity:** $h(\gamma, 1) > \gamma$ for all $\gamma \in [0,1)$, and $h(\gamma, 0) < \gamma$ for all $\gamma \in (0,1]$. Agreement with the collective raises trust; disagreement lowers it.
- **(A2) — Boundedness:** $h(\gamma, z) \in [0,1]$ for all $\gamma \in [0,1]$, $z \in \{0,1\}$.
- **(A3) — Lipschitz continuity:** there exists $L_h \in (0,1)$ such that $|h(\gamma, z) - h(\gamma', z)| \leq L_h|\gamma - \gamma'|$ for all $\gamma, \gamma' \in [0,1]$, $z \in \{0,1\}$. A single round cannot move a trust score arbitrarily far.

Axiom (A1) is what distinguishes *trust* from *satisfaction*: where the perpetual-voting literature lowers the weight of agents whose preferred outcome was selected, (A1) *raises* the weight of agents who agreed with the selected outcome. The convergence and axiomatic analyses that follow are proved at the level of (A1)–(A3) and therefore hold for any mechanism in this class.

Two Ways to Instantiate Trust There are two natural ways to calculate the trust score, which give rise to two sub-classes of TBPV rule: *pairwise trust*, where each agent's score is derived from a matrix of pairwise agreement histories (the instantiation we use later is *EigenTrust*), and *global trust*, where a single scalar score per agent is updated directly from its own agreement history (the instantiation we use later is a *Kalman filter*). The global class is the one our analysis covers directly: a scalar update of the form (1) satisfying (A1)–(A3), of which the Kalman filter is an instance.

4.2 Convergence to a Fixed Point

We are working in an *online* setting: TBPV updates agent weights every round, indefinitely. Establishing that TBPV converges to a fixed point rules out indefinite oscillation: after enough rounds, each agent's influence settles down to a value that depends only on its underlying reliability.

Theorem 4.1 (Convergence of trust-based perpetual voting)**.** *Let h satisfy axioms* (A1)–(A3) *with Lipschitz constant L_h, and suppose the contraction condition $\lambda = L_h + 2L_\rho < 1/\sqrt{n}$ of Lemma B.1 holds. Define $V_k = \|\boldsymbol{\gamma}_k - \boldsymbol{\gamma}^*\|^2$, let $\sigma^2 = \sup_{\boldsymbol{\gamma}} \mathrm{Var}(\gamma_{k+1}(v) \mid \boldsymbol{\gamma}_k)$ be the per-round update variance, and set $\mu := n\lambda^2 < 1$. Then for any initialisation $\boldsymbol{\gamma}_1 \in B(\boldsymbol{\gamma}^*, r)$, and for as long as the iterates remain in $B(\boldsymbol{\gamma}^*, r)$, $E[V_k] \;\leq\; \mu^{k-1}V_1 + \frac{n\sigma^2}{1-\mu}$. In particular, the trust weights converge in mean square to a ball around $\boldsymbol{\gamma}^*$ of radius $\sqrt{n\sigma^2/(1-n\lambda^2)}$.*

The bound has two components. The geometric term $\mu^{k-1}V_1$ decays exponentially, so the initialisation is forgotten at a rate set by $\mu = n\lambda^2$; larger updates (smaller L_h) forget faster but raise σ^2. The second term is an irreducible noise floor arising from the stochasticity of $z_k(v)$: even at the fixed point, the weights continue to fluctuate in a ball around $\boldsymbol{\gamma}^*$ rather than settling on it exactly.

4.3 Does TBPV Satisfy Young's Axioms?

Theorem 3.2 implied Borda is the best *static* rule; we now ask which of those same axioms TBPV itself satisfies once weights become dynamic. TBPV is faithful, and a global (not pairwise) rule is consistent within a round for fixed trust. It is not anonymous, neutral, or cancellation-satisfying: these axioms enforce equal treatment irrespective of history, and TBPV's purpose is to weigh agents unequally by reliability. This is not an oversight: Theorem 5.7 shows precisely why that trade improves collective accuracy.

5 FIXED-POINT DOMINANCE OF TRUST OVER APPROVAL

5.1 TRUSTWORTHINESS

When we consider the VFL setting, the appropriate notion of fairness is not that each agent receives its preferred outcome proportionally (the satisfaction guarantee of the perpetual-voting literature) but rather that agents more informative about the true label receive proportionally more influence over time. We formalise this as *trustworthiness*.

Definition 5.1 (Trustworthiness)**.** A perpetual voting rule satisfies *trustworthiness* if, at the long-run fixed point of the trust dynamics, the influence of each agent v, measured by its trust score $\gamma^*(v)$, is strictly increasing in its true accuracy $p_v = P(\hat{y}_v = y)$.

Lemma 5.2 ($\rho^* > 1/K$ at stable fixed points)**.** *Under axioms* (A1)–(A3)*, any stable fixed point $\boldsymbol{\gamma}^*$ of Φ satisfies $\rho(\boldsymbol{\gamma}^*) > \frac{1}{K}$, provided agents have heterogeneous accuracies with $\max_v p_v > \frac{1}{K}$.*

Proposition 5.3 (TBPV satisfies trustworthiness)**.** *Under axioms* (A1)–(A3) *and the conditions of Lemma 5.2, the trust score $\gamma^*(v)$ at the fixed point is strictly increasing in p_v.*

Trustworthiness is incompatible with the satisfaction guarantee of the perpetual-voting literature, and the incompatibility has a precise price. To state it we need a path-level analogue of trustworthiness that applies to any weighted rule, not only to rules with a trust fixed point.

Definition 5.4 (Eventually accuracy-ordered)**.** A weighted approval rule, selecting $w_t \in \arg\max_c \sum_{v:c\in A_t(v)} \omega_t(v)$, is *eventually accuracy-ordered* if for all agents u, v with $p_u > p_v$, the fraction of rounds $t \le T$ with $\omega_t(u) \le \omega_t(v)$ tends to 0 almost surely as $T \to \infty$.

Consider the following two-agent instance. Agents a and b have accuracies $p_a > p_b$; each round y^t is uniform on $K \ge 2$ labels, each agent votes for a single label (so $A_t(v) = \{\hat{y}_v^t\}$), correctly with probability p_v and otherwise uniformly among the wrong labels, independently across agents and rounds. The agents agree with probability $r := p_a p_b + \frac{(1-p_a)(1-p_b)}{K-1}$, a property of the instance alone.

Theorem 5.5 (Accuracy–satisfaction tradeoff)**.** *On this instance, let R be any perpetual voting rule that always selects an approved candidate. If R satisfies the satisfaction guarantee with parameter β_{sat}, then its collective accuracy $\rho_T = \frac{1}{T}\sum_{t\le T} \mathbb{I}(w_t = y^t)$ obeys, almost surely,*

$$\limsup_{T\to\infty} \rho_T \;\le\; p_a - \frac{p_a - p_b}{1 - r}\Big(\beta_{\mathrm{sat}}\tfrac{1+r}{2} - r\Big)_+,$$

whereas deferring to a on every disagreement attains $\rho_T \to p_a$, the best accuracy available to any such rule.

Corollary 5.6 (Impossibility)**.** *For every $\beta_{\mathrm{sat}} \in (0, 1]$ there is an instance of this form, with heterogeneous accuracies, both agents strictly better than chance, and $p_a > \frac{1}{2}$, on which no eventually accuracy-ordered rule satisfies the satisfaction guarantee with parameter β_{sat}.*

TBPV is built to be accuracy-ordered: by Proposition 5.3 its fixed-point trust is strictly increasing in accuracy, and by Theorem 4.1 the weights concentrate around that fixed point. It therefore gives up the satisfaction guarantee on instances like these. Since the server's objective is to minimise prediction error, Theorem 5.5 makes the trade-off explicit: satisfaction is bought with accuracy, at a rate fixed by the instance.

TBPV modifies an existing mechanism rather than building one from scratch: it takes the perpetual voting framework of Lackner (2020) and replaces its satisfaction-based weight update with a trust-based one. The natural comparison is therefore the fixed point of the original *approval-based* class, which up-weights an agent whose preferred candidate lost and down-weights one whose candidate won. Because accurate agents win more often, that update steadily strips influence from the agents the server should be listening to.

We formalise the tension via *accuracy alignment*, $\Lambda(\boldsymbol{w}) = \mathrm{Cov}(w_v, p_v)$: the two fixed points sit on opposite sides of this quantity ($\Lambda(\boldsymbol{\gamma}^*) > 0 > \Lambda(\boldsymbol{\alpha}^*)$ whenever agents are heterogeneous and both fixed points have $\rho > 1/K$). Alignment is not the whole story: accuracy also falls as weight concentrates on fewer agents, so a rule can align perfectly and still lose. We therefore compare the

fixed points on both counts, measuring concentration by *total influence* $\mathcal{C}(\boldsymbol{w})$ (Definition B.5). That trust is no more concentrated than approval is a hypothesis of the theorem rather than a consequence of the ordering; we report the range over which we have verified it in Remark B.16. Appendix B gives the construction of $\boldsymbol{\alpha}^*$ and the supporting lemmas.

For the analysis we assume agents vote for a single label, correctly with probability p_v and otherwise uniformly among the $K-1$ wrong labels, independently across agents given y^t, with ties broken uniformly. Equivalently, an agent of accuracy p_v votes for the truth with probability $\theta_v = \frac{Kp_v-1}{K-1}$ and otherwise votes uniformly over all K labels. We call θ_v the agent's *signal*: zero at chance accuracy, one at perfect accuracy.

Theorem 5.7 (Trust fixed-point dominance)**.** *Assume agents' votes are conditionally independent given y^t, and that accuracies are heterogeneous, $\mathrm{Var}(p_v) > 0$. Suppose both $\boldsymbol{\gamma}^*$ and $\boldsymbol{\alpha}^*$ satisfy $\rho > 1/K$, that $\bar{p} > 1/K$, and that $\mathcal{C}(\boldsymbol{\gamma}^*) \geq \mathcal{C}(\boldsymbol{\alpha}^*)$. Then the trust fixed point dominates the approval fixed point in collective accuracy, $\rho(\boldsymbol{\gamma}^*) \geq \rho(\boldsymbol{\alpha}^*) + R, |R| \leq 2 \cdot 3^n \max_v \theta_v^2$, with the leading-order inequality strict whenever the two fixed points induce pivotalities that are not all equal.*

5.2 Trust Floor, Minimum Influence, and Minimum Satisfaction

Concentrating influence raises concerns about completely marginalizing less accurate agents. We show that influence cannot vanish: under mild conditions, every agent retains a guaranteed weight floor, and the electorate maintains a baseline satisfaction floor.

Definition 5.8 (Income and retention)**.** A trust update h has *income* $\beta \geq 0$ if $h(\gamma, z) \geq \beta$ for all $\gamma \in [0,1]$, $z \in \{0,1\}$; and *retention* $1-\kappa$, with $\kappa \in (0,1)$, if $h(\gamma, z) \geq (1-\kappa)\gamma$ for all $\gamma \in [0,1]$, $z \in \{0,1\}$. Income is the trust an agent receives unconditionally each round; retention is the fraction of its current trust an agent keeps in the worst case.

Theorem 5.9 (Trust floor)**.** *Let h satisfy axioms* (A1)–(A3)*, and let $\gamma_k(v)$ evolve by $\gamma_{k+1}(v) = h(\gamma_k(v), z_k(v))$ from any initialisation $\gamma_1(v) > 0$.* (i) *If h has income $\beta > 0$, then $\gamma_k(v) \geq \beta$ for all $k \geq 2$.* (ii) *If h has retention $1-\kappa$, then $\gamma_k(v) \geq (1-\kappa)^{k-1}\gamma_1(v)$ for all $k \geq 1$; in particular $\gamma_k(v) > 0$ for every finite k.*

This is the mechanism's basic *minimum-influence guarantee*: no agent is permanently silenced by a run of disagreement, whether that guarantee is a constant floor (positive income) or a geometrically decaying one (retention only). We prove in Appendix C that this floor translates into a concrete *minimum satisfaction guarantee*:

Theorem 5.10 (Minimum satisfaction)**.** *In every round $t \leq j$, at least $d\, n\, \gamma_{\min}/(C_{\max}\, \gamma_{\max})$ voters have w_t in their approval set, where d is the approval-set size, $C_{\max}$ the maximum number of candidates in any round, and $\gamma_{\min}, \gamma_{\max}$ bound the trust scores. Consequently $\frac{1}{n}\sum_{v \in N} sat_j(v) \geq \frac{j\, d\, \gamma_{\min}}{C_{\max}\, \gamma_{\max}}$.*

Together, the trust floor and minimum-influence results guarantee that no honest agent is permanently silenced in weight, and Theorem 5.10 adds an absolute floor on satisfaction averaged across the electorate, tied to approval-set size rather than to quota. This is weaker than a per-voter guarantee, and weaker still than the quota-proportional satisfaction guarantee that Corollary 5.6 shows TBPV cannot satisfy; but it holds regardless of how trust is distributed.

6 Results

6.1 Instantiating TBPV: Kalman and EigenTrust

Global trust: Kalman-filter TBPV. The Kalman filter models latent reliability as a scalar state updated via noisy agreement indicator $z_k(v) \in \{0,1\}$: $\gamma_{k+1}(v) = (1 - K_{k+1}(v))\gamma_k(v) + K_{k+1}(v) z_k(v)$. The gain $K_{k+1}(v) \in (0,1)$ decays to a steady-state K^* balancing adaptation speed and noise. This update satisfies (A1)–(A3). Because $h(\gamma, 0) = (1 - K_{k+1}(v))\gamma$ with $K_{k+1}(v) \to K^*$, the mechanism has retention without positive income, yielding a geometrically decaying floor via Theorem 5.9(ii) rather than a constant one.

	Method	Contiguous		Shuffled	
		Letter	Sensorless	Letter	Sensorless
Reference	Non-federated	0.807 ± 0.053	0.977 ± 0.017	0.807 ± 0.053	0.977 ± 0.017
Trust (ours)	Kalman filter	0.649 ± 0.053	0.844 ± 0.041	0.608 ± 0.052	0.954 ± 0.028
	EigenTrust	0.650 ± 0.054	0.840 ± 0.041	0.611 ± 0.052	0.954 ± 0.027
Adaptive	Consensus-Hedge	0.584 ± 0.085	0.863 ± 0.173	0.486 ± 0.082	0.966 ± 0.030
	Canonical Hedge (oracle)	0.621 ± 0.066	0.938 ± 0.036	0.566 ± 0.074	0.976 ± 0.023
Satisfaction	Unit cost	0.474 ± 0.053	0.741 ± 0.050	0.536 ± 0.054	0.933 ± 0.038
	Approval	0.475 ± 0.052	0.740 ± 0.052	0.537 ± 0.053	0.933 ± 0.040
	Equality	0.473 ± 0.052	0.740 ± 0.050	0.536 ± 0.056	0.933 ± 0.039
	Reset	0.472 ± 0.053	0.738 ± 0.052	0.536 ± 0.054	0.934 ± 0.040
	Proportional	0.475 ± 0.054	0.733 ± 0.050	0.536 ± 0.055	0.932 ± 0.039

Table 1: Cumulative accuracy (mean over 100 trials, $\pm$ std) under the contiguous (heterogeneous) and shuffled (homogeneous) partitions. The non-federated baseline has full feature access and is a reference ceiling, not a competitor.

Pairwise trust: EigenTrust TBPV. EigenTrust computes agent scores from a pairwise agreement matrix $c_{uv} \in \{0, 1\}$, row-stochasticized as C and mixed with uniform teleportation $T = \frac{\tau}{n}\mathbf{1}\mathbf{1}^\top + (1-\tau)C$ for $\tau \in (0, 1)$. Agent scores are the principal eigenvector of T, updated per round. Teleportation gives every agent income τ/n, so the normalised stationary vector satisfies $\gamma_k(v) \geq \tau/n$, unlike the Kalman filter's decaying floor.

6.2 Experimental Predictions

The main body of this paper makes three predictions that we now aim to test directly.

1. **Dominance under heterogeneity.** When agents differ in informativeness ($\mathrm{Var}(p_v) > 0$), trust-based aggregation should outperform the satisfaction class at convergence (Theorem 5.7).
2. **Gap tracks heterogeneity.** The trust advantage vanishes in the homogeneous limit (Corollary B.14), and Remark B.15 suggests it should shrink as heterogeneity is reduced.
3. **Robustness to the trust estimator.** Only the Kalman variant falls under our analysis, but if the trust *signal* rather than the estimator is what matters, the Kalman and EigenTrust variants should reach similar long-run accuracy.

Lacking ground truth, we evaluate *Consensus-Hedge*, which uses losses computed against the learner's own predictions (Chuang et al., 2025). Consensus-Hedge serves as our baseline to isolate *history* from *voting structure*: like TBPV, it aggregates across rounds, but it lacks axiomatic guarantees. Oracle-guided *Canonical Hedge* provides an upper bound. Both variants use ranked ballots aggregated by position-weighted support—making their input at least as expressive as TBPV's approval sets (details in Appendix E).

We evaluate on two tabular multiclass datasets: Letter Recognition (Frey & Slate, 1991) (26 classes, 20,000 instances, 16 features) and Sensorless Drive Diagnosis (Bator, 2013) (11 classes, 58,509 instances, 48 features). Both datasets are vertically partitioned across 5 agents using contiguous (heterogeneous) and shuffled (homogeneous) schemes, matching Predictions 1 and 2, respectively. Complete dataset, partitioning, and experimental details are in the appendix.

6.3 Experimental Results

Dominance Under Heterogeneity: Under the contiguous partition (Table 1), the trust rules outperform the entire satisfaction class on both datasets, by 0.175 on Letter and 0.10 on Sensorless: about three and two per-trial standard deviations respectively, and many standard errors over 100 trials, as Theorem 5.7 predicts. The five satisfaction rules are statistically indistinguishable, consistent with a shared anti-aligned update (Lemma B.11) converging near one fixed point.

Dataset	Partition	$\mathrm{Var}(p_v)$	Gap
Letter	Contiguous	0.029	0.175
	Shuffled	0.007	0.074
Sensorless	Contiguous	0.060	0.100
	Shuffled	0.011	0.021

Table 2: Agent heterogeneity and the trust advantage. $\mathrm{Var}(p_v)$ is the across-agent variance of solo accuracy (mean over 100 trials); Gap is the EigenTrust-approval difference in cumulative accuracy.

Robustness to the Trust Estimator: The Kalman and EigenTrust variants yield virtually interchangeable accuracy, differing by only 0.001 on Letter (0.649 vs. 0.650) and 0.004 on Sensorless (0.844 vs. 0.840)—differences well below the trial-to-trial standard deviation. Although only the Kalman variant falls under our analysis, this suggests the trust signal, not the estimator tracking it, drives performance.

The Gap Tracks Heterogeneity: The vanishing-gap corollary and Remark B.15 predict that the trust–satisfaction gap shrinks as heterogeneity decreases ($\mathrm{Var}(p_v) \to 0$). Testing this via shuffled feature partitioning—which balances agent informativeness—confirms the prediction (Table 2). On Letter, shuffling reduces $\mathrm{Var}(p_v)$ from 0.029 to 0.007 and the gap from 0.175 to 0.074. On Sensorless, $\mathrm{Var}(p_v)$ drops from 0.060 to 0.011 and the gap from 0.10 to 0.021. The collapse is more pronounced on Sensorless due to its larger feature set (48 vs. 16 features on Letter), which leaves less residual variation after shuffling.

Comparison to Hedge: The comparison with consensus-Hedge is nuanced (Table 1). Trust rules outperform Hedge on Letter (0.584 contiguous, 0.486 shuffled), whereas Hedge achieves higher mean accuracy on Sensorless (0.863 contiguous, 0.966 shuffled). Crucially, Theorem 5.7 targets satisfaction-based rules, not adaptive schemes like Hedge. Under Sensorless contiguous partitions, Hedge's higher mean entails four times TBPV's standard deviation (0.173 vs. 0.041), reflecting high sensitivity to early errors; this variance gap vanishes under shuffled partitions (0.030 vs. 0.028). Because Sensorless is easier (non-federated baseline 0.977 vs. 0.807 on Letter), all methods operate in a narrow band. TBPV's structural advantage over Hedge is clearest on harder tasks with greater capacity for separation.

7 Conclusion

We introduce Trust-Based Perpetual Voting (TBPV), an output-only server strategy σ_S that shifts focus from representation to collective accuracy by weighting ballots using dynamic trust signals. We prove TBPV's trust weights converge in mean square to a neighbourhood of a fixed point (Theorem 4.1) that, under heterogeneity and a concentration condition, yields higher collective accuracy (Theorem 5.7). TBPV ranks agent influence by reliability (Proposition 5.3), though satisfying traditional fairness guarantees strictly compromises accuracy (Theorem 5.5, Corollary 5.6). In ground-truth-free streams prone to distribution shifts, empirical metrics offer limited assurance during regime changes. TBPV addresses this with durable axiomatic guarantees: a trust floor keeps every agent's trust positive, trustworthiness aligns influence with reliability, and convergence bounds post-settlement weight drift. Ultimately, TBPV provides a principled, mathematically predictable foundation for vertical federated learning beyond purely performance-tuned aggregation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with editing and formatting prose, and to check proofs. All mathematical statements, proofs and experimental results were verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses two public benchmark datasets (Letter Recognition and Sensorless Drive Diagnosis), neither of which contains personal or identifiable information, and involves no human subjects.

REPRODUCIBILITY STATEMENT

The protocol, datasets, partitioning scheme, model architecture, and hyperparameters required to reproduce reported numbers are given in Appendix F, with per-method hyperparameters in Appendix F.4. All reported results are means over 100 independent trials with seeds stated in the appendix. Proofs of the theoretical claims, together with the assumptions each requires, are given in Appendices A–D; the stability step of Lemma 5.2 is sketched rather than proved in full; the numerical checks reported in Remark B.16 are exact enumerations over the stated ranges.

REFERENCES

Martyna Bator. Dataset for Sensorless Drive Diagnosis. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C5VP5F.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Tianyue Chu and Nikolaos Laoutaris. Fedqv: Leveraging quadratic voting in federated learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 8(2):1–36, 2024.

Yun-Shiuan Chuang, Xiaojin Zhu, and Timothy T Rogers. The delusional hedge algorithm as a model of human learning from diverse opinions. *Topics in Cognitive Science*, 17(1):73–87, 2025.

Igor Felipe de Camargo, Rodolfo Stoffel Antunes, and Gabriel de O. Ramos. On social consensus mechanisms for federated learning aggregation. In *Brazilian Conference on Intelligent Systems*, pp. 236–250. Springer, 2022.

Peter W. Frey and David J. Slate. Letter recognition using holland-style adaptive classifiers. *Machine Learning*, 6(2):161–182, 1991.

Sepandar D Kamvar, Mario T Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th International Conference on World Wide Web*, pp. 640–651, 2003.

Martin Lackner. Perpetual voting: Fairness in long-term decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2103–2110, 2020.

Martin Lackner and Jan Maly. Proportional decisions in perpetual voting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 5722–5729, 2023.

Yang Liu, Yan Kang, Tianyuan Zou, Yanhong Pu, Yuanqin He, Xiaozhou Ye, Ye Ouyang, Ya-Qin Zhang, and Qiang Yang. Vertical federated learning: Concepts, advances, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3615–3634, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL `https://api.semanticscholar.org/CorpusID:53592270`.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. Pmlr, 2017.

Changlong Shi, He Zhao, Bingjie Zhang, Mingyuan Zhou, Dandan Guo, and Yi Chang. Fedawa: Adaptive optimization of aggregation weights in federated learning using client vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30651–30660, 2025.

Yuxia Sun, Aoxiang Sun, Siyi Pan, Zhixiao Fu, and Jingcai Guo. Fedapa: Server-side gradient-based adaptive personalized aggregation for federated learning on heterogeneous data. *arXiv preprint arXiv:2502.07456*, 2025.

Xuezhen Tu, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Yang Zhang, and Juan Li. Incentive mechanisms for federated learning: From economic and game theoretic perspective. *IEEE Transactions on Cognitive Communications and Networking*, 8(3):1566–1593, 2022.

Guan Wang. Interpret federated learning with shapley values. *arXiv preprint arXiv:1905.04519*, 2019.

Ning Wang, Yang Xiao, Yimin Chen, Yang Hu, Wenjing Lou, and Y Thomas Hou. Flare: Defending federated learning against model poisoning attacks via latent space representations. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, pp. 946–958, 2022.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019.

H Peyton Young. Social choice scoring functions. *SIAM Journal on Applied Mathematics*, 28(4): 824–838, 1975.

Kai Yue, Richeng Jin, Chau-Wai Wong, and Huaiyu Dai. Federated learning via plurality vote. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):8215–8228, 2024. doi: 10.1109/TNNLS.2022.3225715.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in Neural Information Processing Systems*, 32, 2019.

A Proof of Proposition 3.3

Proof. Let $n = 2$ and let y^t be uniform on $\{0, 1\}$. Define two candidate distributions, identical except for which client is informative:

$$\mathcal{D}_1 : \; p_1 = 0.9, \; p_2 = 0.1, \qquad \mathcal{D}_2 : \; p_1 = 0.1, \; p_2 = 0.9,$$

with clients conditionally independent given y^t. In this construction agents report hard votes, i.e. point masses on a single label, so that a profile carries no information beyond the labels voted for and is therefore identical under the two distributions. Under $\mathcal{D}_1$ the best fixed pool is $w^* = e_1$ and under $\mathcal{D}_2$ it is $w^* = e_2$; in both cases the comparator returns the informative client's prediction and incurs per-round error 0.1.

Consider the profile $P = (0, 1)$, in which client 1 votes 0 and client 2 votes 1. It arises with probability $\frac{1}{2}[p_1(1 - p_2) + (1 - p_1)p_2] = 0.41$ under both distributions, but

$$\Pr_{\mathcal{D}_1}(y^t = 0 \mid P) = \frac{0.405}{0.41}, \qquad \Pr_{\mathcal{D}_2}(y^t = 0 \mid P) = \frac{0.005}{0.41},$$

so the optimal response to P is label 0 under $\mathcal{D}_1$ and label 1 under $\mathcal{D}_2$. Since f is stationary, $f(P)$ is a single fixed label, and is therefore wrong under at least one of the two distributions.

Suppose $f(P) = 0$ and take $\mathcal{D} = \mathcal{D}_2$. Then f is incorrect on every occurrence of P except when $y^t = 0$, contributing $0.41 \cdot \left(1 - \frac{0.005}{0.41}\right) = 0.405$ to its expected per-round error, while $w^* = e_2$

incurs 0.1 in total. Hence $\delta \geq 0.405 - 0.1 = 0.305 > 0$. If instead $f(P) = 1$, the same computation with $\mathcal{D} = \mathcal{D}_1$ and $w^* = e_1$ gives the same bound. Summing the per-round excess over T rounds, and using $\mathbb{E}[\min_w \sum_t \ell_w^t] \leq \min_w \mathbb{E}[\sum_t \ell_w^t]$, yields $\mathbb{E}[R_{\text{pool}}^T(\sigma_S^f)] \geq \delta T$. □

Remark A.1 (Label-free rules). $\mathcal{D}_1$ and $\mathcal{D}_2$ induce the same distribution over vote profiles: $(0,0)$ and $(1,1)$ each occur with probability 0.09, and $(0,1)$ and $(1,0)$ each with probability 0.41. Since labels are never revealed, any server strategy, however it uses h_S^t, produces the same distribution of outputs under both, and the argument above gives linear expected regret under one of them. Using history is therefore necessary for no-regret (Corollary 3.4) but not sufficient without labels. The construction has $\bar{p} = \frac{1}{2} = \frac{1}{K}$ and one agent worse than chance, so it lies outside the conditions of Theorem 5.7.

B Proofs for Trust-Based Perpetual Voting

B.1 Existence and Convergence

Lemma B.1 (Contraction of Φ)**.** *Assume $\rho(\boldsymbol{\gamma}) > \frac{1}{K}$ for all $\boldsymbol{\gamma}$ in a neighbourhood of $\boldsymbol{\gamma}^*$, and assume $\rho(\cdot)$ is Lipschitz on that neighbourhood with constant L_ρ (see Remark B.2). Under axioms* (A1)–(A3)*, if*

$$\lambda := L_h + 2L_\rho < \frac{1}{\sqrt{n}}, \tag{2}$$

then Φ satisfies the operator-level bound

$$\|\Phi(\boldsymbol{\gamma}) - \boldsymbol{\gamma}^*\| \leq \sqrt{n}\,\lambda \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|, \tag{3}$$

which is a contraction since $\sqrt{n}\,\lambda < 1$.

Proof. Fix any $\boldsymbol{\gamma} \in [0,1]^n$. For each agent v, decompose the deviation from the fixed point:

$$\Phi(\boldsymbol{\gamma})_v - \gamma_v^* = \big[q_v(\boldsymbol{\gamma}) - q_v(\boldsymbol{\gamma}^*)\big]\big[h(\gamma_v, 1) - h(\gamma_v, 0)\big] \tag{4}$$

$$+ q_v(\boldsymbol{\gamma}^*)\big[h(\gamma_v, 1) - h(\gamma_v^*, 1)\big] \tag{5}$$

$$+ (1 - q_v(\boldsymbol{\gamma}^*))\big[h(\gamma_v, 0) - h(\gamma_v^*, 0)\big] \tag{6}$$

Bounding term (4). Since $q_v(\boldsymbol{\gamma}) = \rho(\boldsymbol{\gamma})\, p_v + (1 - \rho(\boldsymbol{\gamma}))\dfrac{1 - p_v}{K - 1}$,

$$|q_v(\boldsymbol{\gamma}) - q_v(\boldsymbol{\gamma}^*)| = |\rho(\boldsymbol{\gamma}) - \rho(\boldsymbol{\gamma}^*)| \cdot \left| p_v - \frac{1 - p_v}{K - 1} \right| \leq L_\rho \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|, \tag{7}$$

since $\left|p_v - \frac{1-p_v}{K-1}\right| \leq 1$. By (A2), $|h(\gamma_v, 1) - h(\gamma_v, 0)| \leq 1$, so this term contributes at most $L_\rho \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|$. **Bounding terms (5) and (6).** By (A3),

$$|h(\gamma_v, z) - h(\gamma_v^*, z)| \leq L_h |\gamma_v - \gamma_v^*| \leq L_h \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|. \tag{8}$$

Since $q_v(\boldsymbol{\gamma}^*) \in [0,1]$, terms (5) and (6) together contribute at most $L_h \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|$. **Combining.**

$$|\Phi(\boldsymbol{\gamma})_v - \gamma_v^*| \leq (L_h + L_\rho)\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\| \leq \lambda \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|, \tag{9}$$

where the second inequality uses $\lambda = L_h + 2L_\rho \geq L_h + L_\rho$; we keep the more conservative constant λ so that the same condition (2) is quoted throughout. Squaring and summing over all n agents,

$$\|\Phi(\boldsymbol{\gamma}) - \boldsymbol{\gamma}^*\|^2 = \sum_v |\Phi(\boldsymbol{\gamma})_v - \gamma_v^*|^2 \leq n\lambda^2 \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|^2. \tag{10}$$

Taking square roots gives $\|\Phi(\boldsymbol{\gamma}) - \boldsymbol{\gamma}^*\| \leq \sqrt{n}\,\lambda \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^*\|$, a contraction since $\lambda < 1/\sqrt{n}$. □

Remark B.2 (Smoothness of ρ). With finitely many agents and labels, $\rho(\boldsymbol{\gamma})$ is piecewise constant in the weights, jumping where a perturbation changes which candidate attains the maximum score, so it is not Lipschitz in the exact finite model. We therefore work with the smoothed dynamics obtained by breaking ties at random and averaging over the tie-breaking randomness, under which ρ is Lipschitz with a constant L_ρ determined by the density of score margins near zero; the mean-field operator Φ is defined on these smoothed quantities throughout. The condition (2) should be read as a statement about that smoothed model.

Proof of Theorem 4.1. For each agent v, decompose the one-step deviation using the law of total variance:

$$E\big[(\gamma_{k+1}(v) - \gamma_v^*)^2 \mid \boldsymbol{\gamma}_k\big] = E\big[(\gamma_{k+1}(v) - \Phi(\boldsymbol{\gamma}_k)_v)^2 \mid \boldsymbol{\gamma}_k\big] \tag{11}$$

$$+ \big(\Phi(\boldsymbol{\gamma}_k)_v - \gamma_v^*\big)^2, \tag{12}$$

where the cross term vanishes because $E[\gamma_{k+1}(v) \mid \boldsymbol{\gamma}_k] = \Phi(\boldsymbol{\gamma}_k)_v$ by definition. The first term is the conditional variance of the update, bounded by σ^2. The second is the squared bias, bounded by Lemma B.1:

$$\big(\Phi(\boldsymbol{\gamma}_k)_v - \gamma_v^*\big)^2 \le \lambda^2 \|\boldsymbol{\gamma}_k - \boldsymbol{\gamma}^*\|^2 = \lambda^2 V_k. \tag{13}$$

Summing over all n agents and taking expectations,

$$E[V_{k+1} \mid \boldsymbol{\gamma}_k] \le n\lambda^2 V_k + n\sigma^2 = \mu V_k + n\sigma^2. \tag{14}$$

Taking expectations unconditionally and unrolling the recursion for $k-1$ steps,

$$E[V_k] \le \mu\, E[V_{k-1}] + n\sigma^2 \tag{15}$$

$$\le \mu^{k-1} V_1 + n\sigma^2 \sum_{j=0}^{k-2} \mu^j \tag{16}$$

$$\le \mu^{k-1} V_1 + \frac{n\sigma^2}{1-\mu}, \tag{17}$$

which is the stated bound. □

B.2 Trustworthiness

Proof of Lemma 5.2. Suppose for contradiction $\rho(\boldsymbol{\gamma}^*) \le \frac{1}{K}$. The agreement probability $q_v(\boldsymbol{\gamma}^*) = \rho^* p_v + (1-\rho^*)\frac{1-p_v}{K-1}$ then has $\rho^* - \frac{1-\rho^*}{K-1} \le 0$, so q_v is non-increasing in p_v: less accurate agents agree with the collective at least as often as more accurate ones. By (A1), agreeing agents gain trust, so inaccurate agents accumulate trust at least as fast as accurate ones, driving ρ further below $\frac{1}{K}$. We sketch the stability step: writing $V_k = \|\boldsymbol{\gamma}_k - \boldsymbol{\gamma}^*\|^2$ satisfies $\mathbb{E}[V_{k+1} \mid \boldsymbol{\gamma}_k] > V_k$ in a neighbourhood of $\boldsymbol{\gamma}^*$, contradicting stability. Hence any stable fixed point has $\rho > \frac{1}{K}$. □

Proof of Proposition 5.3. By the implicit function theorem, $d\gamma_v^*/dq_v > 0$ (numerator positive by (A1), denominator positive by (A3)), and q_v is strictly increasing in p_v because $\rho^* > \frac{1}{K}$ by Lemma 5.2. By transitivity, we have that $\gamma^*(v)$ is strictly increasing in p_v. □

Remark B.3 (Why two agents). An impossibility statement requires only one instance, and the two-agent instance is the sharpest available. With two agents, ordering alone pins down the outcome: the more trusted agent wins every disagreement, so its counterpart's satisfaction is exactly the agreement rate r while its quota remains $(1+r)/2$.

Proof of Theorem 5.5. Let $G_t = \mathbb{I}(\hat{y}_a^t = \hat{y}_b^t)$ indicate agreement and let $\psi_T = \frac{1}{T}\#\{t \le T : G_t = 0,\ w_t = \hat{y}_b^t\}$ be the rate of rounds decided against a. Rounds are i.i.d., so $\frac{1}{T}\sum_{t\le T} G_t \to r$ almost surely.

Satisfaction and quota of b. When the agents agree, R selects the common label; when they disagree, it selects one of the two. Hence $sat_T(b)/T = \frac{1}{T}\sum_{t\le T} G_t + \psi_T$. With singleton ballots, $supp_t(b) = \frac{1}{2}(1+G_t)$, so $qu_T(b)/T \to \frac{1+r}{2}$ almost surely, whatever R does. The satisfaction guarantee therefore requires

$$\liminf_{T\to\infty} \psi_T \ \ge\ \beta_{\mathrm{sat}}\tfrac{1+r}{2} - r \quad \text{a.s.} \tag{18}$$

Accuracy. For labels $\ell_a \ne \ell_b$, a direct computation gives $\Pr(y^t = \ell_a \mid \hat{y}_a^t = \ell_a, \hat{y}_b^t = \ell_b) = \pi_a := p_a(1-p_b)/(1-r)$ and $\Pr(y^t = \ell_b \mid \cdot) = \pi_b := p_b(1-p_a)/(1-r)$; on agreement, $\Pr(y^t = \hat{y}_a^t \mid G_t = 1) = \pi_0 := p_a p_b / r$. Because y^t is independent of the past, and R's decision depends only on the past, the current ballots and internal randomness, these are also the conditional

probabilities given everything R sees. The differences between $\mathbb{I}(w_t = y^t)$ and these conditional probabilities are bounded martingale differences, so by the martingale strong law,

$$\rho_T = \pi_0 \tfrac{1}{T}\sum_t G_t + \pi_a\Big(\tfrac{1}{T}\sum_t(1-G_t) - \psi_T\Big) + \pi_b\,\psi_T + o(1) = p_a - \tfrac{p_a - p_b}{1-r}\,\psi_T + o(1) \quad \text{a.s.},$$

using $\pi_0 r + \pi_a(1-r) = p_a$ and $(\pi_a - \pi_b)(1-r) = p_a - p_b$. Taking $\limsup$ and applying (18) gives the bound. Since $p_a > p_b$ implies $\pi_a > \pi_b$, accuracy is maximised by $\psi_T \equiv 0$, i.e. by deferring to a, which attains p_a. □

Proof of Corollary 5.6. With two agents and singleton ballots, a weighted approval rule selects $\hat{y}_a^t$ on every disagreement round in which $\omega_t(a) > \omega_t(b)$. If it is eventually accuracy-ordered, ψ_T is at most the fraction of rounds with $\omega_t(a) \le \omega_t(b)$, which tends to 0. By (18), the satisfaction guarantee then requires $r \ge \beta_{\mathrm{sat}}\frac{1+r}{2}$, i.e. $r \ge \beta_{\mathrm{sat}}/(2-\beta_{\mathrm{sat}})$.

Choose $K \ge 4$ with $3/(K-1) < \beta_{\mathrm{sat}}/(2-\beta_{\mathrm{sat}})$, set $p_b = 2/K$, which is above chance $1/K$, and any $p_a \in (\frac{1}{2}, 1]$, so that $p_a > \frac{1}{2} \ge p_b$. Then $r \le p_b + \frac{1}{K-1} \le \frac{2}{K} + \frac{1}{K-1} < \frac{3}{K-1} < \beta_{\mathrm{sat}}/(2-\beta_{\mathrm{sat}})$, and the guarantee fails. □

B.3 Fixed-Point Dominance

B.3.1 Characterising the Approval Fixed Point

An approval-based perpetual voting mechanism updates weights via

$$\alpha_{k+1}(v) = \begin{cases} f(\alpha_k(v)) & \text{if } w_k \notin A_k(v), \\ g(\alpha_k(v)) & \text{if } w_k \in A_k(v), \end{cases} \tag{19}$$

where f and g are increasing functions with $f(\alpha) > \alpha$ and $g(\alpha) \le \alpha$: losers are up-weighted, winners are down-weighted or unchanged. We additionally require f and g to be **Lipschitz regular**: there exists $L_f \in (0,1)$ such that $|f'(\alpha)|, |g'(\alpha)| \le L_f < 1$ for all $\alpha \in [0,1]$.

We abstract the approval class via a mean-field operator $\Psi : \Delta^{n-1} \to \Delta^{n-1}$ on the normalised weight simplex:

$$\Psi(\boldsymbol{\alpha})_v = q_v(\boldsymbol{\alpha}) \cdot g(\alpha_v) + (1 - q_v(\boldsymbol{\alpha})) \cdot f(\alpha_v), \tag{20}$$

normalised so that $\sum_v \Psi(\boldsymbol{\alpha})_v = 1$. As with the trust fixed point, the domain Δ^{n-1} is a compact convex subset of $\mathbb{R}^n$, and Ψ maps Δ^{n-1} into itself continuously (by the Lipschitz regularity of f and g and continuity of $q_v(\cdot)$); by the Brouwer Fixed Point Theorem, Ψ admits a fixed point $\boldsymbol{\alpha}^* \in \Delta^{n-1}$.

B.3.2 Weight–Accuracy Alignment

Definition B.4 (Accuracy alignment)**.** For a weight vector $\boldsymbol{w} \in \Delta^{n-1}$ and accuracy vector $\mathbf{p} = (p_1, \ldots, p_n)$, the accuracy alignment is

$$\Lambda(\boldsymbol{w}) = \mathrm{Cov}(w_v, p_v) = \frac{1}{n}\sum_{v=1}^{n} w_v p_v - \bar{w}\,\bar{p}, \tag{21}$$

where $\bar{w} = \frac{1}{n}\sum_v w_v$ and $\bar{p} = \frac{1}{n}\sum_v p_v$.

Alignment alone does not determine collective accuracy: concentrating weight on the best-aligned agents raises Λ but reduces the number of agents with any effective say. The following decomposition separates the two effects.

Work in the signal parametrisation $\theta_v = (Kp_v - 1)/(K-1)$ of Section 5, and write $\bar{\theta} = \frac{1}{n}\sum_v \theta_v$. For $S \subseteq N$, let $\rho_S(\boldsymbol{w})$ denote the probability that the true label wins when the agents in S vote for the truth and all others vote uniformly at random. By label symmetry $\rho_\emptyset = 1/K$ for every $\boldsymbol{w}$: with no signal in the electorate, no weighting does better than chance.

Definition B.5 (Pivotality and total influence)**.** The *pivotality* of agent v under weights $\boldsymbol{w}$ is

$$\mathrm{Inf}_v(\boldsymbol{w}) \;:=\; \rho_{\{v\}}(\boldsymbol{w}) - \rho_\emptyset(\boldsymbol{w}),$$

the gain in the probability that the truth wins when v alone votes truthfully against a background of random votes. The *total influence* of $\boldsymbol{w}$ is $\mathcal{C}(\boldsymbol{w}) := \sum_v \mathrm{Inf}_v(\boldsymbol{w})$. Both depend only on the weights, not on the agents' accuracies.

Lemma B.6 (Heavier agents are more pivotal)**.** *If* $w_i \geq w_j$ *then* $\mathrm{Inf}_i(\boldsymbol{w}) \geq \mathrm{Inf}_j(\boldsymbol{w})$.

Proof. Both pivotalities share the term $\rho_\emptyset$, so it suffices to compare $\rho_{\{i\}}$ with $\rho_{\{j\}}$. Couple the two experiments: draw the votes of all agents other than i and j once, giving a score vector s over labels, and draw a single uniform label U for whichever of i, j is not forced to vote truthfully. Under $\rho_{\{i\}}$ the score vector is $s + w_i e_0 + w_j e_U$ and under $\rho_{\{j\}}$ it is $s + w_j e_0 + w_i e_U$, where e_c is the indicator vector of label c and 0 denotes the true label.

If $U = 0$ the two vectors coincide. If $U \neq 0$, the first exceeds the second by $w_i - w_j \geq 0$ in coordinate 0 and falls short of it by the same amount in coordinate U, with all other coordinates equal. The probability that label 0 wins under uniform tie-breaking is non-decreasing in s_0 and non-increasing in s_c for every $c \neq 0$, so it is weakly larger in the first configuration. The inequality holds for every realisation of (s, U); taking expectations gives the claim. □

Lemma B.7 (Equal weights maximise total influence)**.** $\mathcal{C}(\boldsymbol{w}) \leq \mathcal{C}(\mathbf{1})$ *for every* $\boldsymbol{w}$.

Proof. Suppose every agent has the same signal θ, and let n_c be the number of votes cast for label c. Under the signal parametrisation,

$$\Pr(\text{votes} \mid y = c) = \prod_v \Big[\theta\, \mathbb{I}(v \text{ votes } c) + \frac{1-\theta}{K}\Big] = \Big(\frac{1-\theta}{K}\Big)^n \Big(1 + \frac{\theta K}{1-\theta}\Big)^{n_c},$$

which is strictly increasing in n_c. With a uniform prior over labels, the Bayes-optimal decision is therefore unweighted plurality, with any tie-break among maximisers. Hence $\rho_{\boldsymbol{w}}(\theta) \leq \rho_{\mathbf{1}}(\theta)$ for every $\theta \in [0, 1]$. Both sides equal $1/K$ at $\theta = 0$, and by multilinearity $\rho_{\boldsymbol{w}}(\theta) = 1/K + \theta\, \mathcal{C}(\boldsymbol{w}) + O(\theta^2)$. Dividing by θ and letting $\theta \downarrow 0$ gives $\mathcal{C}(\boldsymbol{w}) \leq \mathcal{C}(\mathbf{1})$. □

Lemma B.8 (Alignment and concentration determine collective accuracy)**.** *For any weights* $\boldsymbol{w}$*,*

$$\rho(\boldsymbol{w}) = \frac{1}{K} + n\,\mathrm{Cov}\big(\mathrm{Inf}_v(\boldsymbol{w}),\, \theta_v\big) + \mathcal{C}(\boldsymbol{w})\,\bar{\theta} + R(\boldsymbol{w}), \qquad |R| \leq 3^n \max_v \theta_v^2. \tag{22}$$

Collective accuracy is therefore increasing, to leading order in the signal, both in the alignment of influence *with signal and in total influence.*

Proof. Because the agents' informed indicators are independent, $\rho(\boldsymbol{w}) = \sum_{S \subseteq N} \big(\prod_{v \in S} \theta_v\big)\big(\prod_{v \notin S}(1 - \theta_v)\big)\rho_S(\boldsymbol{w})$, a multilinear polynomial in $\boldsymbol{\theta}$. Its constant term is $\rho_\emptyset = 1/K$ and its linear terms are $\sum_v \theta_v(\rho_{\{v\}} - \rho_\emptyset) = \sum_v \theta_v \mathrm{Inf}_v(\boldsymbol{w})$. Every remaining monomial has degree at least two. Collecting terms, the coefficient of $\prod_{v \in S} \theta_v$ is $c_S = \sum_{T \subseteq S} (-1)^{|S \setminus T|} \rho_T$, so $|c_S| \leq 2^{|S|}$ since each $\rho_T \in [0, 1]$, and the degree-≥ 2 terms are bounded by $\max_v \theta_v^2 \sum_{k \geq 2} \binom{n}{k} 2^k \leq 3^n \max_v \theta_v^2$. Finally $\sum_v \theta_v \mathrm{Inf}_v = n\,\mathrm{Cov}(\mathrm{Inf}_v, \theta_v) + \mathcal{C}(\boldsymbol{w})\bar{\theta}$ by the definition of covariance. □

Remark B.9. Equation (22) is the precise form of the intuition behind Λ. The first term is alignment, measured on pivotality rather than on raw weights; the second is the cost of concentration, which no purely alignment-based statistic can see. Concentrating all weight on the single most accurate agent maximises alignment while driving every other agent's pivotality to zero, which is why a weighting can be perfectly aligned and still lose. This is why Theorem 5.7 carries hypothesis $\mathcal{C}(\boldsymbol{\gamma}^*) \geq \mathcal{C}(\boldsymbol{\alpha}^*)$. Since θ_v is an increasing affine function of p_v, with $\mathrm{Cov}(\mathrm{Inf}_v, \theta_v) = \frac{K}{K-1}\mathrm{Cov}(\mathrm{Inf}_v, p_v)$, the covariance term may equally be read in terms of accuracy.

B.3.3 Sign of Alignment at Each Fixed Point

The two lemmas below differentiate each agent's fixed-point weight with respect to its own agreement probability q_v, holding the collective accuracy ρ fixed. This is the mean-field simplification used throughout: q_v depends on ρ, which in turn depends on the whole weight vector, and we treat that dependence as exogenous when computing $d\gamma_v^*/dq_v$.

We now show the two fixed points carry alignment of opposite sign, and order the agents in opposite directions. Combined with Lemma B.8, this orders their collective accuracies once the concentration condition $\mathcal{C}(\boldsymbol{\gamma}^*) \geq \mathcal{C}(\boldsymbol{\alpha}^*)$ holds.

Lemma B.10 (Alignment at the trust fixed point)**.** *At the trust fixed point $\boldsymbol{\gamma}^*$, provided $\rho(\boldsymbol{\gamma}^*) > \frac{1}{K}$ and $\mathrm{Var}(p_v) > 0$,*

$$\Lambda(\boldsymbol{\gamma}^*) > 0. \tag{23}$$

Moreover γ_v^* is strictly increasing in p_v.

Proof. Define $F(\gamma, q) := q \cdot h(\gamma, 1) + (1-q) \cdot h(\gamma, 0) - \gamma$. The fixed-point condition for agent v is $F(\gamma_v^*, q_v(\boldsymbol{\gamma}^*)) = 0$. The partial derivatives are

$$\frac{\partial F}{\partial q} = h(\gamma, 1) - h(\gamma, 0) > 0 \tag{24}$$

by (A1), and

$$\frac{\partial F}{\partial \gamma} = q \cdot h_\gamma(\gamma, 1) + (1-q) \cdot h_\gamma(\gamma, 0) - 1 \leq L_h - 1 < 0 \tag{25}$$

by (A3). By the implicit function theorem,

$$\frac{d\gamma_v^*}{dq_v} = -\frac{\partial F/\partial q}{\partial F/\partial \gamma} \tag{26}$$

$$= \frac{h(\gamma_v^*, 1) - h(\gamma_v^*, 0)}{1 - q_v h_\gamma(\gamma_v^*, 1) - (1-q_v) h_\gamma(\gamma_v^*, 0)} \tag{27}$$

$$> 0, \tag{28}$$

so γ_v^* is strictly increasing in $q_v(\boldsymbol{\gamma}^*)$. Since $q_v(\boldsymbol{\gamma}^*) = \rho^* p_v + (1-\rho^*)\frac{1-p_v}{K-1}$ is strictly increasing in p_v when $\rho^* > \frac{1}{K}$, γ_v^* is strictly increasing in p_v. Hence $\Lambda(\boldsymbol{\gamma}^*) = \mathrm{Cov}(\gamma_v^*, p_v) > 0$ whenever $\mathrm{Var}(p_v) > 0$. □

Lemma B.11 (Alignment at the approval fixed point)**.** *At the approval fixed point $\boldsymbol{\alpha}^*$, provided $\rho(\boldsymbol{\alpha}^*) > \frac{1}{K}$ and $\mathrm{Var}(p_v) > 0$,*

$$\Lambda(\boldsymbol{\alpha}^*) < 0. \tag{29}$$

Moreover α_v^* is strictly decreasing in p_v.

Proof. At the approval fixed point $\Psi(\boldsymbol{\alpha}^*) = \boldsymbol{\alpha}^*$. Define $G(\alpha, q) := q \cdot g(\alpha) + (1-q) \cdot f(\alpha) - \alpha$, with fixed-point condition $G(\alpha_v^*, q_v(\boldsymbol{\alpha}^*)) = 0$. The partials are

$$\frac{\partial G}{\partial q} = g(\alpha) - f(\alpha) < 0 \tag{30}$$

since $f(\alpha) > \alpha \geq g(\alpha)$ for interior α, and

$$\frac{\partial G}{\partial \alpha} = q \cdot g'(\alpha) + (1-q) \cdot f'(\alpha) - 1 \leq L_f - 1 < 0 \tag{31}$$

by the Lipschitz regularity of f and g. By the implicit function theorem,

$$\frac{d\alpha_v^*}{dq_v} = \frac{g(\alpha_v^*) - f(\alpha_v^*)}{1 - q_v g'(\alpha_v^*) - (1-q_v) f'(\alpha_v^*)} < 0, \tag{32}$$

so α_v^* is strictly decreasing in $q_v(\boldsymbol{\alpha}^*)$, hence in p_v (since q_v increases in p_v when $\rho^* > \frac{1}{K}$). Therefore $\Lambda(\boldsymbol{\alpha}^*) = \mathrm{Cov}(\alpha_v^*, p_v) < 0$ whenever $\mathrm{Var}(p_v) > 0$; the covariance vanishes only in the degenerate case $\mathrm{Var}(p_v) = 0$, which is excluded by assumption. □

Remark B.12. Lemmas B.10 and B.11 expose the structural difference between the two classes. Trust mechanisms assign weight *proportional* to consensus agreement q_v, which increases in p_v when $\rho > \frac{1}{K}$. Approval mechanisms assign weight *inversely* to q_v by design, actively counteracting accuracy concentration. The two fixed points therefore sit on opposite sides of the alignment spectrum.

Proof of Theorem 5.7. By Lemma B.10, γ_v^* is strictly increasing in p_v, hence in θ_v; by Lemma B.6, so is $\mathrm{Inf}_v(\boldsymbol{\gamma}^*)$. Two similarly ordered sequences have non-negative covariance, so $\mathrm{Cov}(\mathrm{Inf}_v(\boldsymbol{\gamma}^*), \theta_v) \geq 0$, with strict inequality unless the pivotalities are all equal — which can happen even for strictly ordered weights, as when every pair of agents outweighs any third and the rule reduces to unweighted plurality. By Lemma B.11, α_v^* is strictly decreasing in p_v, so the same argument applied to oppositely ordered sequences gives $\mathrm{Cov}(\mathrm{Inf}_v(\boldsymbol{\alpha}^*), \theta_v) \leq 0$. Subtracting the two instances of (22),

$$\rho(\boldsymbol{\gamma}^*) - \rho(\boldsymbol{\alpha}^*) = n\Big[\mathrm{Cov}\big(\mathrm{Inf}_v(\boldsymbol{\gamma}^*), \theta_v\big) - \mathrm{Cov}\big(\mathrm{Inf}_v(\boldsymbol{\alpha}^*), \theta_v\big)\Big] + \bar{\theta}\Big[\mathcal{C}(\boldsymbol{\gamma}^*) - \mathcal{C}(\boldsymbol{\alpha}^*)\Big] + R.$$

The first bracket is non-negative, and strictly positive whenever the pivotalities induced by $\boldsymbol{\gamma}^*$ are not all equal. The second is non-negative by $\mathcal{C}(\boldsymbol{\gamma}^*) \geq \mathcal{C}(\boldsymbol{\alpha}^*)$, and $\bar{\theta} > 0$ because $\bar{p} > 1/K$. The remainder is bounded as in Lemma B.8. □

Remark B.13 (Relation between the two fixed points). The two fixed points are closely related in shape. Let $\gamma_v^* = q_v$ (the fixed point of the Kalman instantiation, with steady-state gain) and $\alpha_v^* \propto 1 - q_v$, both normalised to the simplex, and let $\bar{q} = \frac{1}{n}\sum_v q_v$. A direct computation gives $\hat{\gamma}_v^* - \frac{1}{n} = (q_v - \bar{q})/(n\bar{q})$ and $\hat{\alpha}_v^* - \frac{1}{n} = -(q_v - \bar{q})/(n(1-\bar{q}))$, so

$$\hat{\alpha}_v^* - \tfrac{1}{n} = -\frac{\bar{q}}{1-\bar{q}}\Big(\hat{\gamma}_v^* - \tfrac{1}{n}\Big) :$$

the approval fixed point carries the same deviations from uniform as the trust fixed point, reflected, and scaled by $\bar{q}/(1-\bar{q})$, which exceeds 1 exactly when $\bar{q} > \frac{1}{2}$. The approval class therefore does not merely invert the ordering of the agents; when agents agree with the collective more often than not, it also places the weights further from uniform.

This does not by itself establish $\mathcal{C}(\boldsymbol{\gamma}^*) \geq \mathcal{C}(\boldsymbol{\alpha}^*)$. Total influence is not a function of distance from uniform: with $n = 3$ and deviations $d = (0.2, -0.1, -0.1)$, the weights $\mathbf{u} + d = (0.533, 0.233, 0.233)$ give $\mathcal{C} = 0.500$ at $K = 2$, while their reflection $\mathbf{u} - d = (0.133, 0.433, 0.433)$, equally far from uniform, gives $\mathcal{C} = 0.750$ — concentrating decisive power on one agent costs more total influence than sidelining one agent among three. Nor are reflected pairs ordered by majorization, so no Schur-concavity property of $\mathcal{C}$ would bridge the gap. We therefore treat $\mathcal{C}(\boldsymbol{\gamma}^*) \geq \mathcal{C}(\boldsymbol{\alpha}^*)$ as a hypothesis to be checked for the instantiation at hand, and report in Remark B.16 how often it holds.

B.3.4 COROLLARIES

Corollary B.14 (Vanishing gap under homogeneity)**.** *If all agents have identical accuracy ($p_v = \bar{p}$), then $\mathrm{Var}(p_v) = 0$ and $\rho(\boldsymbol{\gamma}^*) = \rho(\boldsymbol{\alpha}^*)$. Trust weighting provides no benefit over approval weighting when agents are homogeneous; its advantage is entirely due to heterogeneity.*

Remark B.15 (Heterogeneity and the size of the gap). Holding total influence fixed, the covariance term in (22) grows with the spread of $\{p_v\}$, so in VFL settings where feature partitions induce greater variation in agent informativeness, the trust advantage is larger. We do not state this as a corollary, since varying $\mathrm{Var}(p_v)$ also moves $\mathcal{C}(\boldsymbol{\gamma}^*)$ and $\mathcal{C}(\boldsymbol{\alpha}^*)$; the empirical trend is reported in Table 2.

Remark B.16 (Numerical support). We computed both fixed points exactly by enumeration for $n = 3, \ldots, 7$ agents, using $\gamma_v^* = q_v$ and $\alpha_v^* \propto 1 - q_v$. For $K = 2$, with accuracies drawn uniformly from $(\frac{1}{2}, 0.95)$, dominance held in all 150 instances and condition $\mathcal{C}(\boldsymbol{\gamma}^*) \geq \mathcal{C}(\boldsymbol{\alpha}^*)$ in all 150. Replacing the trust map by the sharper $\gamma_v^* = q_v^k$ shows where the trade in (22) turns: dominance still held in every instance for $k \leq 4$, and failed in $10/150$ instances at $k = 8$ and $26/150$ at $k = 16$, where concentration outweighs alignment. In all 750 binary instances there was no case in which $\mathcal{C}(\boldsymbol{\gamma}^*) \geq \mathcal{C}(\boldsymbol{\alpha}^*)$ held and dominance failed, including well outside the low-signal regime. For $K \in \{3, 4, 5\}$, with accuracies drawn uniformly from $(1/K + 0.05,\ 0.95)$, dominance held in 179 of 180 instances, while $\mathcal{C}(\boldsymbol{\gamma}^*) \geq \mathcal{C}(\boldsymbol{\alpha}^*)$ held in $59/60$, $54/60$ and $46/60$ instances respectively: the condition is sufficient but increasingly conservative as K grows, consistent with $\bar{q}$ falling below $\frac{1}{2}$ as the number of labels grows (Remark B.13).

Corollary B.17 (Long-run accuracy dominance)**.** *Under the conditions of Theorem 5.7, and by Theorem 4.1, the trust dynamics converge to $\boldsymbol{\gamma}^*$; assume the approval dynamics likewise converge to $\boldsymbol{\alpha}^*$. Since $\rho(\boldsymbol{\gamma}^*) > \rho(\boldsymbol{\alpha}^*)$, there exists $k_0 < \infty$ such that $E[\rho_k^T] > E[\rho_k^A]$ for all $k \geq k_0$, where*

ρ_k^T and ρ_k^A denote the collective accuracy of a trust-based and an approval-based mechanism at round k.

Remark B.18 (Role of $\rho > \frac{1}{K}$). Both alignment lemmas require $\rho > \frac{1}{K}$ at the respective fixed points. At the trust fixed point this is guaranteed by the stability argument of Lemma 5.2: a fixed point with $\rho \leq \frac{1}{K}$ is unstable. For the approval fixed point we take $\rho(\boldsymbol{\alpha}^*) > \frac{1}{K}$ as a mild precondition, satisfied whenever the agents are jointly informative enough that an unweighted vote beats random. This is the regime of interest: if no weighting scheme could exceed random accuracy the aggregation problem would be vacuous, and it is the condition under which approval-based perpetual voting is itself designed to operate.

C Proof of Minimum Satisfaction Guarantee

C.1 Trust Floor

Proof of Theorem 5.9. (i) For $k \geq 2$, $\gamma_k(v) = h(\gamma_{k-1}(v), z_{k-1}(v)) \geq \beta$ by the definition of income. (ii) By induction on k: the case $k = 1$ is immediate, and if $\gamma_k(v) \geq (1-\kappa)^{k-1}\gamma_1(v)$ then $\gamma_{k+1}(v) = h(\gamma_k(v), z_k(v)) \geq (1-\kappa)\gamma_k(v) \geq (1-\kappa)^k\gamma_1(v)$ by the definition of retention. Positivity follows since $\kappa < 1$ and $\gamma_1(v) > 0$. □

C.2 Setup

- N is the set of voters and C_t the candidates in round t.
- $A_t(v) \subseteq C_t$ is v's approval set in round t, with $|A_t(v)| = d$ for all v and $t \leq j$.
- $\gamma_t(v)$ is the trust score of v in round t, and the winner is $w_t = \arg\max_{c \in C_t} sc_t(c)$ with $sc_t(c) = \sum_{v : c \in A_t(v)} \gamma_t(v)$.
- Satisfaction up to round j is $sat_j(v) = |\{t \leq j : w_t \in A_t(v)\}|$.
- Trust scores are bounded below: $\gamma_t(v) \geq \gamma_{\min}$ for all v and $t \leq j$. The maximum candidate count is $C_{\max} = \max_{t \leq j} |C_t|$.
- Trust scores are also bounded above: $\gamma_t(v) \leq \gamma_{\max}$ for all v and $t \leq j$.

C.3 Result

Proof of Theorem 5.10. In round t, the total trust-weighted score across candidates is

$$\sum_{c \in C_t} sc_t(c) = \sum_{c \in C_t} \sum_{v : c \in A_t(v)} \gamma_t(v) = \sum_{v \in N} \gamma_t(v)\,|A_t(v)| = d \sum_{v \in N} \gamma_t(v) \;\geq\; d\,n\,\gamma_{\min}.$$

The winner takes at least the average score, so $sc_t(w_t) \geq \frac{1}{|C_t|} \sum_{c \in C_t} sc_t(c) \geq d\,n\,\gamma_{\min}/|C_t|$. Because $sc_t(w_t) = \sum_{v : w_t \in A_t(v)} \gamma_t(v) \leq \gamma_{\max}\,|\{v : w_t \in A_t(v)\}|$,

$$|\{v : w_t \in A_t(v)\}| \;\geq\; \frac{d\,n\,\gamma_{\min}}{|C_t|\,\gamma_{\max}} \;\geq\; \frac{d\,n\,\gamma_{\min}}{C_{\max}\,\gamma_{\max}}.$$

Summing over $t \leq j$ and exchanging the order of summation gives $\sum_{v \in N} sat_j(v) \geq j\,d\,n\,\gamma_{\min}/(C_{\max}\gamma_{\max})$; dividing by n gives the stated bound. □

Remark C.1. The bound is an average across the electorate, not a floor for each individual voter: a lower bound on the number of satisfied voters per round does not preclude one voter being unsatisfied in every round. For the EigenTrust instantiation the constants are explicit, with $\gamma_{\min} = \tau/n$ from the teleportation income and $\gamma_{\max} \leq 1 - (n-1)\tau/n$; for the Kalman instantiation Theorem 5.9(ii) gives only a geometrically decaying $\gamma_{\min}$, and the guarantee decays with it.

D Proof of Axiomatic Standing of TBPV

Proposition D.1 (Anonymity)**.** *All TBPV rules are anonymous.*

Proof. $sc_k(c) = \sum_{v:c\in A_k(v)} \gamma_k(v)$ is a function of the multiset of trust scores of approving agents, not of identities. Permuting the identities of the agents while preserving their trust scores and ballots leaves every $sc_k(c)$ unchanged. □

Proposition D.2 (Failure of neutrality)**.** *TBPV rules are not neutral.*

Proof. Trust scores encode historical information about which candidates each agent supported. Let $N = \{v_1, v_2\}$, $C = \{c_1, c_2\}$, with $\gamma_k(v_1) = 0.8$, $\gamma_k(v_2) = 0.2$, $A_k(v_1) = \{c_1\}$, $A_k(v_2) = \{c_2\}$. Then $sc_k(c_1) = 0.8 > sc_k(c_2) = 0.2$, so $w_k = c_1$; v_1 agrees and gains trust, v_2 disagrees and loses trust. Now permute $c_1 \leftrightarrow c_2$ in the ballots. The scores swap and $w_k = c_2$—the outcome permutes correctly in this round. But at round $k+1$ the original problem rewards v_1 for agreeing with c_1 while the permuted problem rewards v_1 for agreeing with c_2; the accumulated trust is not relabelled by the permutation, creating asymmetry in subsequent rounds. Neutrality fails. □

Proposition D.3 (Consistency)**.** *A global TBPV rule satisfies consistency; a pairwise TBPV rule does not in general.*

Proof. Recall $sc_k(c) = \sum_{v\in N:c\in A_k(v)} \gamma_k(v)$ and $w_k = \arg\max_{c\in C_k} sc_k(c)$. Let $N_1, N_2 \subset N$ be disjoint with $N_1 \cup N_2 = N$, and suppose every agent's trust score is the same whether evaluated within its sub-group or within N. If TBPV on N_1 alone selects c and TBPV on N_2 alone selects c, then for all $c' \neq c$,

$$\sum_{v\in N_1:c\in A_k(v)} \gamma_k(v) > \sum_{v\in N_1:c'\in A_k(v)} \gamma_k(v),$$

$$\sum_{v\in N_2:c\in A_k(v)} \gamma_k(v) > \sum_{v\in N_2:c'\in A_k(v)} \gamma_k(v).$$

Adding,

$$\sum_{v\in N:c\in A_k(v)} \gamma_k(v) > \sum_{v\in N:c'\in A_k(v)} \gamma_k(v),$$

i.e. $sc_k(c) > sc_k(c')$ for all $c' \neq c$, so TBPV on N selects c. The argument requires trust scores to be *group-composition independent*. This holds for any *global* trust rule, where $\gamma_k(v)$ depends only on v's own history and the selected outcomes, not on which other agents participate. It fails for *pairwise* trust rules, where each $\gamma_k(v)$ is derived from a peer-trust matrix that changes when the agent set changes. □

Proposition D.4 (Faithfulness)**.** *All TBPV rules are faithful.*

Proof. For a single-agent electorate $\{v\}$ at any round k, $sc_k(c) = \gamma_k(v)$ for $c \in A_k(v)$ and $sc_k(c) = 0$ otherwise. By the trust floor (Theorem 5.9), $\gamma_k(v) > 0$, so the maximisers of sc_k are exactly the candidates in $A_k(v)$. The rule therefore selects v's approved set, which is faithfulness. □

Proposition D.5 (Failure of cancellation)**.** *TBPV rules do not satisfy cancellation whenever trust scores are non-uniform.*

Proof. Cancellation requires that a balanced profile produce a complete tie. Let $N = \{v_1, v_2\}$, $C = \{c_1, c_2\}$, $A_k(v_1) = \{c_1\}$, $A_k(v_2) = \{c_2\}$: each candidate is approved by exactly one agent, so the profile is balanced. If $\gamma_k(v_1) \neq \gamma_k(v_2)$ (e.g. 0.8 and 0.2), then $sc_k(c_1) = 0.8 \neq 0.2 = sc_k(c_2)$, and R selects c_1 alone rather than tying. Cancellation fails. Note that this is by design; cancellation is the axiom which enforces equal treatment of voters, and TBPV's purpose is to weigh agents unequally by reliability. A TBPV rule satisfies cancellation only in the degenerate case of uniform trust $\gamma_k(v) \equiv \bar{\gamma}$, where it reduces to unweighted approval voting. □

E HEDGE COMPARATORS

Both Hedge variants share a weight $\gamma^t(v)$ per agent, initialised uniformly. Each agent submits a full ranked ballot over the candidates, obtained by sorting its soft prediction $\hat{p}_v^t$, and the server aggregates these by trust-weighted, position-weighted (Borda-like) support:

$$\hat{y}^t = \arg\max_{c \in \mathcal{Y}} \sum_v \gamma^t(v)\, b\big(\mathrm{rank}_v^t(c)\big),$$

$$b(j) = K - j, \qquad \hat{y}_v^t = \arg\max_k \hat{p}_v^t(k),$$

updating multiplicatively, $\gamma^{t+1}(v) = \gamma^t(v) \cdot e^{-\eta\, \ell^t(v)}$ for learning rate $\eta > 0$, renormalised to sum to one. The Hedge ballot is therefore at least as expressive as the approval ballot TBPV consumes: it carries the agent's full ordering, where TBPV sees only a thresholded set. Any advantage TBPV shows over Hedge is thus not an artefact of richer ballots. *Canonical Hedge* sets $\ell^t(v) = \mathbb{I}(\hat{y}_v^t \neq y^t)$, requiring oracle access to the true label; *consensus-Hedge* instead substitutes the server's own output for the label, $\ell^t(v) = \mathbb{I}(\hat{y}_v^t \neq \hat{y}^t)$.

Consensus-Hedge is our comparator because it isolates *history* from *voting structure*: like TBPV it carries information across rounds, but it is not a voting rule and has none of the axiomatic guarantees discussed above. Canonical Hedge, which requires oracle label access, serves as a reference ceiling.

F EXPERIMENTAL SETUP

F.1 DATASETS

We use two tabular multiclass classification datasets. **Letter Recognition** (Frey & Slate, 1991) comprises 20,000 instances over 16 geometric features, with a 26-class target. **Sensorless Drive Diagnosis** (Bator, 2013) comprises 58,509 instances over 48 motor-current features, with an 11-class target. For both datasets, features are scaled with StandardScaler and the data is split $40/60$ train/test, with a fresh draw per trial.

F.2 VERTICAL PARTITIONING

The vertical federated setting distributes overlapping slices of the feature axis across 5 agents: each receives a contiguous band of width $\approx d/5$ with an overlap of $\approx d/(2 \cdot 5)$ on each side. **Contiguous partition (heterogeneous):** features are sliced in dataset order, inducing $\mathrm{Var}(p_v) > 0$. **Shuffled partition (homogeneous):** the feature axis is randomly permuted before slicing (fresh permutation per trial), balancing agent informativeness.

F.3 PROTOCOL AND MODELS

Each configuration is run for 100 independent trials of 100 rounds. We report each method's cumulative accuracy (mean per-round accuracy over the full run) averaged across trials. Each agent's model is a 3-layer feed-forward network with ReLU activations and 20% dropout, optimised with AdamW (Loshchilov & Hutter, 2017) under L_2 regularisation. We compare TBPV (Kalman and EigenTrust variants) against the satisfaction/approval class (approval, proportional, equality, reset, unit-cost), an adaptive no-regret comparator (consensus-Hedge), and a non-federated model with full feature access as a reference ceiling.

F.4 HYPERPARAMETERS

F.4.1 LOCAL MODELS

Each agent uses a 3-layer feed-forward network: $\mathrm{Linear}(d_i, 64) \to \mathrm{ReLU} \to \mathrm{Dropout}(0.2) \to \mathrm{Linear}(64, 32) \to \mathrm{ReLU} \to \mathrm{Dropout}(0.2) \to \mathrm{Linear}(32, C)$, trained with AdamW (lr $= 10^{-3}$, weight decay 10^{-4}). Each agent pretrains locally and independently on its own feature slice of the training split for 200 epochs with batch size 64; no gradients, activations or parameters are exchanged at any point. In the online protocol, each voting round triggers one AdamW step on that round's feature vector, with cross-entropy loss toward the consensus outcome w_k.

F.4.2 SEEDING

Each trial t uses seed $42 + t$ for the global RNG state (`random`, NumPy, and PyTorch) and for the train/test split (`sklearn`'s `random_state`). The randomised feature partition (`--random-feature-partition`) uses seed $500{,}000 + t$.

F.4.3 CONSENSUS-HEDGE

Multiplicative-weights update $w_i \leftarrow w_i \exp(-\eta\, \ell_i)$ with $\eta = 0.5$, renormalised after each round. Ballots are full rankings of the candidates, obtained by sorting each agent's soft prediction, and are aggregated by weight- and position-weighted (Borda-like) support; the loss ℓ_i compares agent i's top-ranked candidate with the collective outcome.

F.4.4 KALMAN TRUST

Scalar filters use process noise $Q = 0.01$, observation noise $R = 0.1$, initial confidence $\gamma_0 = 0.5$, and initial uncertainty $P_0 = 1$.

F.4.5 EIGENTRUST

Power iteration on the row-stochastic trust matrix, with teleportation parameter $\tau = 0.01$, convergence tolerance 10^{-6}, and a cap of 100 iterations. Agreement history uses decay 0.9; the trust-matrix EMA step size is 0.1; agents falling below trust 0.01 are excluded from the weighted-support computation. Since $0.01 > \tau/n = 0.002$, this exclusion can override the trust floor of Section 6.1.