LOCAL UPDATES MEET ERROR FEEDBACK: COMMUNICATION-EFFICIENT VERTICAL MINIMAX FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vertical federated learning (VFL) couples model blocks held by different clients, making block-coordinate methods a natural approach to optimization. In minimax learning, local updates and compressed exchanges introduce errors that interact with the evolving maximization variables. We develop two local block methods: FedBCD-Max combines approximate server-side maximization with client-local descent, whereas FedBCDA performs client-local descent-ascent when the maximization variables are partitioned across clients. We further equip both methods with bidirectional error feedback to correct uplink and downlink compression errors. We establish the linear convergence guarantee for strongly convex-strongly concave problems, and an $O(T^{-1})$ convergence rate for nonconvex-strongly concave problems. The analysis jointly controls local trajectories, maximization error, and two coupled compression memories. It characterizes admissible effective step sizes and preserves the uncompressed accuracy dependence with an explicit ω^{-2} factor in the leading communication-round bounds, where ω is the compressor contraction parameter. Experiments on fair vertical classification and vertical adversarial learning illustrate the outperformance of our proposed algorithms, and further show that error feedback mitigates the degradation caused by direct compression.

1 INTRODUCTION

Vertical federated learning (VFL) (Liu et al., 2024) enables clients holding different feature blocks of the same samples to train a joint model without centralizing their raw features. This induces a natural block-structured optimization problem, which naturally leads to block-coordinate methods, where clients update their local parameter blocks using information exchanged with the other participants (Liu et al., 2022b; Valdeira et al., 2025). Beyond standard minimization, several VFL tasks (Liu et al., 2022a; Man et al., 2025) are naturally formulated as a minimax problems of the form

$$\min_{x=(x_1,\dots,x_K)} \max_{y\in\mathcal{Y}} F(x_1,\dots,x_K,y), \tag{1}$$

where x retains the vertically partitioned client-side structure, while y represents an auxiliary maximization variable. Depending on the application, the maximization variable may be shared globally or partitioned across clients, giving rise to the global-y and local-y settings studied below.

Two settings. We study coupled block minimax problems as in Equation (1), where client i owns the model block x_i. We distinguish two settings according to the ownership of the maximization variables, as illustrated in Figure 1:

1. In the *global-y* setting, a shared variable $y \in \mathcal{Y}_g$ is maintained at the server. A motivating example is fair vertical classification, where the dual variable y represents Lagrange multipliers associated with fairness constraints defined on the joint prediction model (Liu et al., 2022a).
2. In the *local-y* setting, the maximization variable is block-partitioned across clients: client i maintains $y_i \in \mathcal{Y}_i$, with $y = (y_1, \dots, y_K)$ and $\mathcal{Y} = \prod_{i=1}^{K} \mathcal{Y}_i$. This structure arises naturally in vertical adversarial learning, where the adversarial perturbation is partitioned according to the vertically distributed feature blocks (Man et al., 2025; 2026).

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.

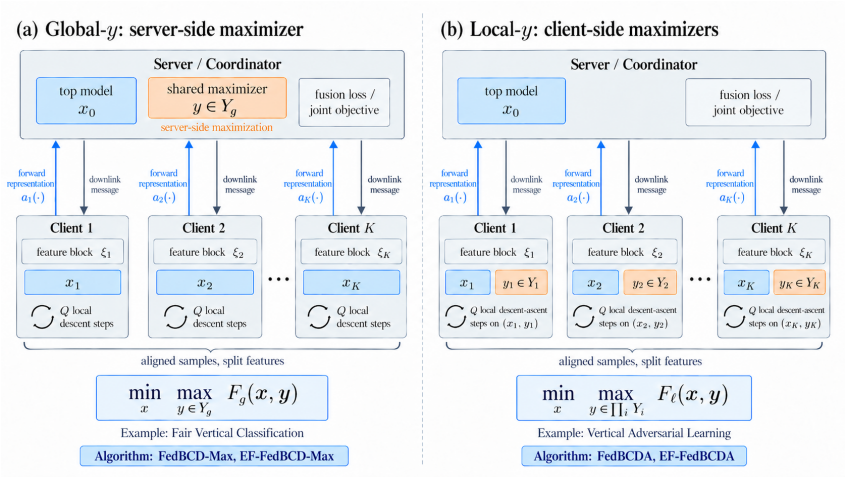


Figure 1: Problem setup and algorithm schemes

Table 1: Summary of Results

Method	**Setting**	**Communication rounds T**	
		Exact exchange	Bidirectional EF
FedBCD-Max	SC–SC	$\widetilde{O}\big(\kappa^2 \log(1/\epsilon)\big)$	$\widetilde{O}\big(\kappa^2 \omega^{-2} \log(1/\epsilon)\big)$
(Global-y)	NC–SC	$O\big(L(1+\kappa)\Delta_\Phi \epsilon^{-1}\big)$	$O\big(L(1+\kappa)\omega^{-2}\Delta_\Phi \epsilon^{-1}\big)$
FedBCDA	SC–SC	$\widetilde{O}\big(\kappa^2 \log(1/\epsilon)\big)$	$\widetilde{O}\big(\kappa^2 \omega^{-2} \log(1/\epsilon)\big)$
(Local-y)	NC–SC	$O\big(L\kappa^2 \mathcal{W}_0 \epsilon^{-1}\big)$	$O\big(L\kappa^2 \omega^{-2} \mathcal{W}_0 \epsilon^{-1}\big)$

Note: Here $\kappa = L/\mu$, and $\omega \in (0, 1]$ is the compressor contraction parameter, $\Delta_\Phi = \Phi(x^0) - \Phi^\star$, and $\mathcal{W}_0 = \Delta_\Phi + \lambda[\Phi(x^0) - F_\ell(x^0, y^0)]$ with $1 \le \lambda \le 2$. The bounds hold for any $Q \ge 1$ with the admissible effective steps in the theorems. $\widetilde{O}$ suppresses additional logarithmic factors in parameters. See Appendix A.3 for the finite-time bounds.

Local updates and compressed communication. Local block updates and communication compression are well studied for vertical minimization (Liu et al., 2022b; Castiglia et al., 2022; Valdeira et al., 2025), while minimax formulations increasingly arise in fair and adversarial VFL (Liu et al., 2022a; Man et al., 2026; Beznosikov et al., 2026). Unlike horizontal federated minimax learning, where clients typically optimize replicated models, VFL clients update distinct blocks of a coupled objective using stale cross-client information (Deng & Mahdavi, 2021; Sharma et al., 2022; Zhang et al., 2024). Maximization and compression introduce additional errors through approximate or evolving dual variables and tracked communication states. We therefore ask: *Can local block minimax algorithms retain non-asymptotic convergence under bidirectional error-feedback compression, and how do local steps and compression affect the performances?*

Our approach and contributions. We develop FedBCD-Max for global-y problems, combining approximate server-side maximization with Q client-local descent steps, and FedBCDA for local-y problems, performing Q client-local descent–ascent steps. Their compressed variants replace exact uplink and downlink exchanges by error-feedback tracking. Table 1 summarizes the resulting communication-round guarantees.

- **Local block minimax methods and convergence guarantees.** We establish finite-time convergence under both strongly-convex–strongly-concave (SC–SC) and nonconvex–strongly-concave (NC–SC) objectives. With exact communication and controlled residual errors, both methods attain $\widetilde{O}(\kappa^2 \log(1/\epsilon))$ rounds in SC–SC, and $O(\epsilon^{-1})$ rounds in NC–SC. The results hold for any $Q \ge 1$ under explicit effective-step conditions; for the underlying fixed-step SC–SC specializations, the κ^2 dependence is worst-case tight.

- **Bidirectional error-feedback compression.** We extend both methods with compressed uplink and downlink. Under contractive compression with parameter ω, the resulting algorithms preserve the uncompressed dependence on ϵ with an explicit $O(\omega^{-2})$ communication round overhead.
- **Numerical evaluation.** Experiments on fair vertical classification and vertical adversarial learning illustrate the two maximizer placements and local-update schemes. Bidirectional error feedback substantially mitigates the degradation caused by direct sparsification under the same compressed communication budget.

2 Related Work

Block-coordinate and minimax optimization in VFL. FedBCD develops multiple local updates for vertically partitioned minimization, with clients updating distinct parameter blocks using stale cross-client information (Liu et al., 2022b). Communication-efficient extensions include representation compression and hybrid first/zeroth-order schemes (Castiglia et al., 2022; Wang et al., 2023; Stanko et al., 2026). Minimax formulations arise in fair VFL and vertical adversarial learning (Liu et al., 2022a; Qi et al., 2022; Man et al., 2025; 2026), while Beznosikov et al. (2026) study saddle-point reformulations of vertical minimization. We instead focus on local block minimax methods with shared or partitioned maximizers and their communication-efficient variants.

Local updates for distributed minimax optimization. Local computation is a standard approach to reducing communication in horizontal federated learning (McMahan et al., 2017; Stich, 2019; Karimireddy et al., 2020), and minimax optimization has been extensively studied in centralized and distributed settings (Jin et al., 2020; Lin et al., 2020a;b; Zhang et al., 2021; Yang et al., 2022; Zhang & Hu, 2025; Deng & Mahdavi, 2021; Sharma et al., 2022). Zhang et al. (2024) further develop local-update methods for distributed variational inequalities through consensus reformulations. In contrast, VFL clients optimize distinct blocks of one coupled objective, so our analysis must control stale-block drift together with maximization error rather than consensus disagreement among replicated models.

Error feedback and compressed saddle-point methods. Error feedback and compressed-difference tracking are widely used to compensate for biased compression (Stich et al., 2018; Karimireddy et al., 2019; Richtarik et al., 2021), including bidirectional schemes for distributed minimization and variational inequalities (Liu et al., 2020; Beznosikov et al., 2022; Fatkhullin et al., 2025). In VFL, EF-VFL analyzes representation tracking for nonconvex and Polyak–Łojasiewicz objectives with one local update per round (Valdeira et al., 2025). We study bidirectional error feedback together with multiple stale-block minimax updates for both global-y and local-y formulations.

3 Minimax VFL with Exact Block Communication

Problem classes. Client $i \in [K] := \{1, \dots, K\}$ owns an unconstrained primal block $x_i \in \mathbb{R}^{d_i}$, and $x = (x_1, \dots, x_K)$. In the global-y model, the server maintains $y \in \mathcal{Y}_g$ and we solve

$$\min_{x} \max_{y \in \mathcal{Y}_g} F_g(x, y), \qquad F_g(x, y) := \mathbb{E}_\xi f_g(x_1, \dots, x_K, y; \xi). \tag{2}$$

In the local-y model, client i also maintains $y_i \in \mathcal{Y}_i$. With $z_i = (x_i, y_i)$, $z = (z_1, \dots, z_K)$, and $\mathcal{Y}_\ell = \prod_{i=1}^K \mathcal{Y}_i$, the problem is

$$\min_{x} \max_{y \in \mathcal{Y}_\ell} F_\ell(x, y), \qquad F_\ell(x, y) := \mathbb{E}_\xi f_\ell(z_1, \dots, z_K; \xi). \tag{3}$$

Define $\Phi(x) := \max_{y \in \mathcal{Y}} F(x, y)$, $y^\star(x) := \arg\max_{y \in \mathcal{Y}} F(x, y)$, $\Phi^\star := \inf_x \Phi(x)$. Next we introduce the regularity conditions in this work.

Assumption 1 (Regularity conditions)**.** The set $\mathcal{Y}$ is nonempty, closed, and convex, and F is continuously differentiable. There exist $L \geq \mu > 0$ such that:

1. **Function geometry:** $\|\nabla F(x, y) - \nabla F(x', y')\| \leq L\|(x, y) - (x', y')\|$ for all feasible (x, y) and (x', y'). $F(x, \cdot)$ is μ-strongly concave on $\mathcal{Y}$ for every x.
2. **Primal geometry:** in SC–SC, $F(\cdot, y)$ is μ-strongly convex for every $y \in \mathcal{Y}$; in NC–SC, no convexity in x is assumed and $\Phi^\star > -\infty$.

Convergence criteria. Assumption 1 implies $y^\star(x)$ unique. Write $\kappa = L/\mu$ and $\Delta_\Phi = \Phi(x^0) - \Phi^\star$. In SC–SC, let $(x^\star, y^\star)$ be the unique saddle point and define

$$\Psi(y) := \min_x F(x, y), \qquad \mathrm{Gap}(x, y) := \Phi(x) - \Psi(y), \qquad \mathrm{Gap}(z) := \mathrm{Gap}(x, y). \tag{4}$$

In NC–SC, we use the primal stationarity $\|\nabla\Phi(\cdot)\|^2$ as the measurement. Note that the envelope has smoothness constant $L_\Phi = L(1+\kappa)$ (refer to Appendix A.4.1).

Local directions and sampling. Round t uses a mini-batch $\mathcal{B}_t$ of size B for $Q \geq 1$ local substeps. Set $u_i = x_i$ in global-y and $u_i = z_i$ in local-y. Let $g_i^t(v)$ be the sampled direction for client i with local argument v and all other blocks fixed at u_{-i}^t. The population directions are $\nabla_{x_i} F_g(v, x_{-i}^t, y^t)$ and $\nabla_{z_i} F_\ell(v, z_{-i}^t)$, respectively; (v, u_{-i}^t) replaces block i only. The analysis requires the following properties of the sampled maps.

Assumption 2 (Stochastic local directions)**.** Let $\mathcal{F}_t$ be the history immediately before drawing $\mathcal{B}_t$, including the synchronized state and, in global-y, the computed y^t. Write $g^t = (g_i^t(u_i^t))_{i=1}^K$ and $\bar{g}^t = \mathbb{E}[g^t \mid \mathcal{F}_t]$. Under exact communication,

$$\bar{g}^t = \begin{cases} \nabla_x F_g(x^t, y^t), & \text{global-}y, \\ \nabla F_\ell(z^t), & \text{local-}y, \end{cases} \qquad \mathbb{E}\big[\|g^t - \bar{g}^t\|^2 \mid \mathcal{F}_t\big] \leq \begin{cases} \sigma_g^2/B, & \text{global-}y, \\ \sigma_\ell^2/B, & \text{local-}y. \end{cases}$$

Also $\|g_i^t(v) - g_i^t(w)\| \leq L\|v - w\|$ for all admissible v, w.

Algorithms. Next we propose FedBCD-Max and FedBCDA algorithms for the global and local cases, respectively. Both Algorithm 1 and 2 take (T, Q, B) as common inputs and use effective steps $\alpha_x = Q\eta_x$, $\alpha_y = Q\eta_y$. An ε-accurate maximizer at x means a point $y \in \mathcal{Y}_g$ satisfying $\Phi(x) - F_g(x, y) \leq \varepsilon$. FedBCD-Max computes y^t before drawing $\mathcal{B}_t$ and holds it fixed during local descent. Note that y^t is easily attainable by many existing optimizers in classical optimization like GD or Nesterov GD (Nesterov, 2013). FedBCDA updates x_i and y_i simultaneously, with $\Pi_{\mathcal{Y}_i}$ denoting Euclidean projection. One round provides the fixed cross-client information for $g_i^t(\cdot)$ and synchronizes the updated blocks; server-only inner iterations are excluded.

Algorithm 1 FedBCD-Max (global-y)

Require: $x^0, \eta_x, \{\varepsilon_t\}_{t<T}, \varepsilon_{\text{out}}$
Synchronize x^0.
for $t = 0, \ldots, T-1$ **do**
  Find ε_t-accurate y^t at x^t.
  Draw $\mathcal{B}_t$; fix $g_i^t(\cdot)$.
  $x_i^{t,0} \leftarrow x_i^t$.
  for $q = 0, \ldots, Q-1$ **do**
    $g_{x_i}^{t,q} \leftarrow g_i^t(x_i^{t,q})$.
    $x_i^{t,q+1} \leftarrow x_i^{t,q} - \eta_x g_{x_i}^{t,q}$.
  end for
  $x_i^{t+1} \leftarrow x_i^{t,Q}$; synchronize.
end for
At x^T, find ε_{out}-accurate y^T.
Ensure: (x^T, y^T) in SC–SC; x^R in NC–SC

Algorithm 2 FedBCDA (local-y)

Require: $z^0 = (x^0, y^0), \eta_x, \eta_y$
Synchronize z^0.
for $t = 0, \ldots, T-1$ **do**
  Draw $\mathcal{B}_t$; fix $g_i^t(\cdot)$.
  $z_i^{t,0} \leftarrow z_i^t$.
  for $q = 0, \ldots, Q-1$ **do**
    $(g_{x_i}^{t,q}, g_{y_i}^{t,q}) \leftarrow g_i^t(z_i^{t,q})$.
    $x_i^{t,q+1} \leftarrow x_i^{t,q} - \eta_x g_{x_i}^{t,q}$.
    $y_i^{t,q+1} \leftarrow \Pi_{\mathcal{Y}_i}(y_i^{t,q} + \eta_y g_{y_i}^{t,q})$.
  end for
  $z_i^{t+1} \leftarrow z_i^{t,Q}$.
  Synchronize the updated states.
end for
Ensure: $z^T = (x^T, y^T)$ in SC–SC; x^R in NC–SC

Theorem 1 (Communication complexity)**.** *Suppose Assumptions 1 and 2 hold. For any $Q \geq 1$, choose the effective steps below with the universal prefactors specified in Appendix A.7, we have*

(i) ***SC–SC, global-y.*** *For $\alpha_x \asymp [L(1+\kappa)]^{-1}$, FedBCD-Max attains $\mathbb{E}\,\mathrm{Gap}(x^T, y^T) \leq \epsilon$ in $T = \widetilde{O}(\kappa^2 \log(1/\epsilon))$ rounds.*

(ii) ***SC–SC, local-y.*** *For $\alpha_x = \alpha_y \asymp (L\kappa)^{-1}$, FedBCDA attains $\mathbb{E}\,\mathrm{Gap}(z^T) \leq \epsilon$ in $T = \widetilde{O}(\kappa^2 \log(1/\epsilon))$ rounds.*

(iii) ***NC–SC, global-y.*** *For $\alpha_x \asymp [L(1+\kappa)]^{-1}$, FedBCD-Max attains $\mathbb{E}\|\nabla\Phi(x^R)\|^2 \leq \epsilon$ in $T = O(L(1+\kappa)\Delta_\Phi/\epsilon)$ rounds.*

(iv) ***NC–SC, local-y.*** *Choose $\alpha_y \asymp L^{-1}$ and $\alpha_x \asymp (L\kappa^2)^{-1}$ with $\lambda = c_\lambda \kappa^2 \alpha_x/\alpha_y \in [1,2]$, where c_λ is the universal constant from Proposition 4. FedBCDA attains $\mathbb{E}\|\nabla\Phi(x^R)\|^2 \le \epsilon$ in $T = O(L\kappa^2\mathcal{W}_0/\epsilon)$ rounds, where $\mathcal{W}_0 = \Delta_\Phi + \lambda[\Phi(x^0) - F_\ell(x^0, y^0)]$.*

Here $\widetilde{O}$ hides additional logarithmic dependence on initialization and problem parameters. The complete finite-time bounds retain the stochastic and inner-solver residuals.

Remark 1 (Interpretation of local steps)**.** When $Q = 1$ and all gradients and inner maximization steps are exact, FedBCD-Max reduces to GD-Max (Jin et al., 2020), while FedBCDA reduces to projected simultaneous GDA (Lin et al., 2020a). Accordingly, the κ^2 dependence in the SC–SC results is worst-case tight (Zhang et al., 2022; Lee et al., 2024), see Appendix A.7.1 for more details. In the NC–SC local-y case, the prescribed primal and dual step sizes recover the standard two-time-scale GDA scaling (Lin et al., 2020a). For $Q > 1$, our theory permits multiple local updates by scaling the per-step sizes so that the effective steps $\alpha_x = Q\eta_x$ and $\alpha_y = Q\eta_y$ remain admissible. Hence, for fixed effective steps, increasing Q does not yield an additional Q-fold improvement in communication rounds; if the per-step sizes are instead fixed, increasing Q can enlarge the effective step only until the admissible step-size limit is reached.

4 Bidirectional Error Feedback for Block Communication

We add bidirectional error feedback (Bi-EF) to FedBCD-Max and FedBCDA. The local updates are unchanged; we only replace the exact uplink and downlink exchanges in Section 3 by compressed tracked messages, following the EF21 mechanism (Richtarik et al., 2021; Valdeira et al., 2025).

Bidirectional tracking. Let $u_i = x_i$ for FedBCD-Max and $u_i = z_i = (x_i, y_i)$ for FedBCDA. In Section 3, the updated client states u_i^{t+1} are synchronized exactly after the Q local steps. We instead let $\widehat{u}_i^t$ denote the server's reconstruction and update it by

$$\widehat{u}_i^{t+1} = \widehat{u}_i^t + \mathcal{C}_i^{\mathrm{up}}\big(u_i^{t+1} - \widehat{u}_i^t\big)\,. \tag{5}$$

For FedBCD-Max, the server therefore computes y^t using the reconstructed state: $\widehat{F}_t(y) := F_g(\widehat{u}^t, y)$, $\widehat{\Phi}_t := \max_{y\in\mathcal{Y}_g} \widehat{F}_t(y)$. The downlink information used to construct the frozen local direction $g_i^t(\cdot)$ is compressed similarly. Let $m_i^t = M_i(\widehat{u}^t)$ denote its exact target and $\widehat{m}_i^t$ its reconstruction. Client i uses

$$\widetilde{g}_i^t(v) := \begin{cases} R_i(v; \widehat{m}_i^t, y^t, \mathcal{B}_t), & \text{global-}y, \\ R_i(v; \widehat{m}_i^t, \mathcal{B}_t), & \text{local-}y, \end{cases} \tag{6}$$

throughout its Q local steps, and the downlink reconstruction is updated by

$$\widehat{m}_i^{t+1} = \widehat{m}_i^t + \mathcal{C}_i^{\mathrm{down}}\big(m_i^{t+1} - \widehat{m}_i^t\big)\,, \qquad m_i^{t+1} = M_i(\widehat{u}^{t+1}). \tag{7}$$

The sender and receiver maintain identical copies of each tracker. No terminal downlink is required.

Assumption 3 (Regularity and contractive compression)**.** The communication maps and compressors satisfy the following conditions.

1. **Message regularity.** After fixed coordinate normalization, $M = (M_1, \dots, M_K)$ satisfies $\|M(a) - M(a')\| \le \|a - a'\|$ on its domain, including all reachable trackers.
2. The reconstructed direction equals the exact stale-block direction under exact tracking, and its perturbation is L-Lipschitz in the local-block and message errors, with the stacked bounds specified in Appendix A.2.
3. **Conditional contractivity.** For some common $\omega \in (0, 1]$, every uplink or downlink compressor call satisfies $\mathbb{E}\|\mathcal{C}(v) - v\|^2 \le (1-\omega)\|v\|^2$.

These conditions are standard or analogous to those used in error-feedback and compressed VFL analyses (Richtarik et al., 2021; Fatkhullin et al., 2025; Valdeira et al., 2025). Next we introduce the EF variant algorithms.

Two EF variants. Algorithms 3–4 share (T, Q, B) and the compressors. Exact initialization sets $\widehat{u}_i^0 = u_i^0$ and $\widehat{m}_i^0 = M_i(u^0)$ at both endpoints. A surrogate solve to tolerance ε means $\widehat{\Phi}_t - \widehat{F}_t(y) \le \varepsilon$. Global-$y$ solves precede the mini-batch draw; local-y updates remain simultaneous. Client-indexed instructions are parallel, q-loops are sequential, and R is as in Section 3.

Algorithm 3 EF-FedBCD-Max

Require: $x^0, \eta_x, \{\varepsilon_t\}_{t<T}, \varepsilon_{\text{out}}$
Initialize both trackers exactly.
for $t = 0, \dots, T-1$ **do**
  Solve $\widehat{F}_t$ to tolerance ε_t, giving y^t.
  Draw $\mathcal{B}_t$; set $x_i^{t,0} \leftarrow x_i^t$.
  for $q = 0, \dots, Q-1$ **do**
    $g_{x_i}^{t,q} \leftarrow \widetilde{g}_i^t(x_i^{t,q})$.
    $x_i^{t,q+1} \leftarrow x_i^{t,q} - \eta_x g_{x_i}^{t,q}$.
  end for
  $x_i^{t+1} \leftarrow x_i^{t,Q}$.
  Update $\widehat{u}_i^{t+1}$ by Equation (5).
  if $t < T-1$ **then**
    Update $\widehat{m}_i^{t+1}$ by Equation (7).
  end if
end for
Final solve of $\widehat{F}_T$ to ε_{out} gives y^T.
Ensure: (x^T, y^T) in SC–SC; x^R in NC–SC

Algorithm 4 EF-FedBCDA

Require: $z^0 = (x^0, y^0), \eta_x, \eta_y$
Initialize both trackers exactly.
for $t = 0, \dots, T-1$ **do**
  Draw $\mathcal{B}_t$.
  $z_i^{t,0} \leftarrow z_i^t$.
  for $q = 0, \dots, Q-1$ **do**
    $(g_{x_i}^{t,q}, g_{y_i}^{t,q}) \leftarrow \widetilde{g}_i^t(z_i^{t,q})$.
    $x_i^{t,q+1} \leftarrow x_i^{t,q} - \eta_x g_{x_i}^{t,q}$.
    $y_i^{t,q+1} \leftarrow \Pi_{\mathcal{Y}_i}(y_i^{t,q} + \eta_y g_{y_i}^{t,q})$.
  end for
  $z_i^{t+1} \leftarrow z_i^{t,Q}$.
  Update $\widehat{u}_i^{t+1}$ by Equation (5).
  if $t < T-1$ **then**
    Update $\widehat{m}_i^{t+1}$ by Equation (7).
  end if
end for
Ensure: $z^T = (x^T, y^T)$ in SC–SC; x^R in NC–SC

Theorem 2 (Communication complexity with bidirectional error feedback)**.** *Under Assumptions 1–3, we have*

(i) ***SC–SC, global-y.*** *With $\alpha_x \asymp \omega^2/[L(1+\kappa)]$, EF-FedBCD-Max attains $\mathbb{E}\,\mathrm{Gap}(x^T, y^T) \le \epsilon$ in $T = \widetilde{O}(\kappa^2\omega^{-2}\log(1/\epsilon))$ rounds.*

(ii) ***SC–SC, local-y.*** *With $\alpha_x = \alpha_y \asymp \omega^2/(L\kappa)$, EF-FedBCDA attains $\mathbb{E}\,\mathrm{Gap}(z^T) \le \epsilon$ in $T = \widetilde{O}(\kappa^2\omega^{-2}\log(1/\epsilon))$ rounds.*

(iii) ***NC–SC, global-y.*** *With $\alpha_x \asymp \omega^2/[L(1+\kappa)]$, EF-FedBCD-Max attains $\mathbb{E}\|\nabla\Phi(x^R)\|^2 \le \epsilon$ in $T = O(L(1+\kappa)\Delta_\Phi/(\omega^2\epsilon))$ rounds.*

(iv) ***NC–SC, local-y.*** *Choose $\alpha_y \asymp \omega^2/L$ and $\alpha_x \asymp \omega^2/(L\kappa^2)$ with $\lambda = c_\lambda\kappa^2\alpha_x/\alpha_y \in [1, 2]$, as in Theorem 1. EF-FedBCDA attains $\mathbb{E}\|\nabla\Phi(x^R)\|^2 \le \epsilon$ in $T = O(L\kappa^2\mathcal{W}_0/(\omega^2\epsilon))$ rounds.*

Remark 2 (Relation to standard methods)**.** At $\omega = 1$, the EF variants recover the uncompressed algorithms (Proposition 7). Without maximization they reduce to FedBCD-style local block descent (Liu et al., 2022b), while with an exact downlink Equation (5) coincides with EF21-style state tracking (Richtarik et al., 2021).

Remark 3 (Local steps and communication)**.** Increasing Q can reduce communication rounds only until the effective steps $\alpha_x = Q\eta_x$ and $\alpha_y = Q\eta_y$ reach their admissible limits; at fixed effective steps, the bounds are uniform in Q. For fixed $\omega > 0$, EF introduces no additional error floor, but incurs an $O(\omega^{-2})$ round overhead that must be traded against smaller messages (Appendix A.7.2).

Remark 4 (Private labels)**.** Under standard private-label VFL, the server computes the backward message using its private labels and server-side model. After a client changes its local model, this message generally must be recomputed before another exact local gradient can be evaluated; hence the private-label interface of Valdeira et al. (2025) directly supports only $Q = 1$. Our $Q > 1$ theory applies when the protocol provides a reusable round-level message from which clients can evaluate the prescribed local directions after changing their local blocks.

5 Experiment

We evaluate the two maximizer placements using fair vertical classification for global-y and vertical adversarial learning for local-y. In both cases, clients perform Q local updates from a synchronized round-start state. For compression experiments, we apply the same bidirectional residual-tracking

mechanism at the VFL communication interface; these representation-level experiments are therefore not literal instantiations of Theorem 2.

5.1 y-GLOBAL: FAIR VERTICAL CLASSIFICATION

Formulation. Let $x = (x_1, \ldots, x_C)$ be the client model blocks and let $D(x) := \mathcal{L}_{a=0,b=1}(x) - \mathcal{L}_{a=1,b=1}(x)$ denote the differentiable equality-of-opportunity surrogate disparity. With classification loss $\mathcal{L}(x)$ and penalty threshold ϵ, we optimize

$$\min_{x} \max_{\lambda_1,\lambda_2 \geq 0} \Phi_{\text{fair}}(x, \lambda) = \mathcal{L}(x) + \lambda_1(D(x) - \epsilon) + \lambda_2(-D(x) - \epsilon) - \frac{c}{2}(\lambda_1^2 + \lambda_2^2). \tag{8}$$

Maximizing over λ yields $\mathcal{L}(x) + [(|D(x)| - \epsilon)_+]^2/(2c)$; thus finite c gives a soft fairness penalty, not a guarantee that $|D(x)| \leq \epsilon$. Here $y = (\lambda_1, \lambda_2)$ is server-held; six parties split the features, and the active party holds labels and the sensitive attribute. We evaluate ℓ_2-regularized logistic regression and a width-32 one-hidden-layer MLP.

Experiment results. We evaluate Adult, COMPAS, and Crime (Appendix B) using logistic regression and a one-hidden-layer MLP. We compare FairVFL (Liu et al., 2022a) with FedBCD-Max using $Q = 5$ local updates. We report predictive accuracy together with the surrogate disparity $|D(x)|$ and the discrete fairness metrics DFP and DFN. Tables 2–3 show that FedBCD-Max generally improves predictive accuracy while maintaining comparable group disparities. Figure 2 further compares the loss and surrogate-disparity trajectories. Implementation details and hyperparameters are given in Appendix C.

Table 2: Linear-model accuracy and group disparity on Adult, COMPAS, and Crime.

Method	Adult			COMPAS			Crime		
	ACC	DFP	DFN	ACC	DFP	DFN	ACC	DFP	DFN
FairVFL ($\epsilon = 0.01$)	80.81%	**0.04**	0.06	65.40%	0.08	0.08	79.63%	0.11	0.02
FairVFL ($\epsilon = 0.05$)	80.87%	**0.04**	0.06	65.15%	0.15	0.14	79.47%	0.15	0.01
FairVFL ($\epsilon = 0.10$)	80.95%	0.05	0.02	65.72%	0.26	0.22	80.13%	0.17	0.04
FedBCD-Max ($\epsilon = 0.01$)	82.04%	0.06	**0.01**	**67.42%**	**0.03**	**0.07**	79.97%	**0.09**	0.01
FedBCD-Max ($\epsilon = 0.05$)	82.06%	0.06	0.02	**67.42%**	0.09	0.11	81.30%	0.11	**0.00**
FedBCD-Max ($\epsilon = 0.10$)	**82.19%**	0.07	0.05	67.36%	0.16	0.17	**82.14%**	0.14	0.03

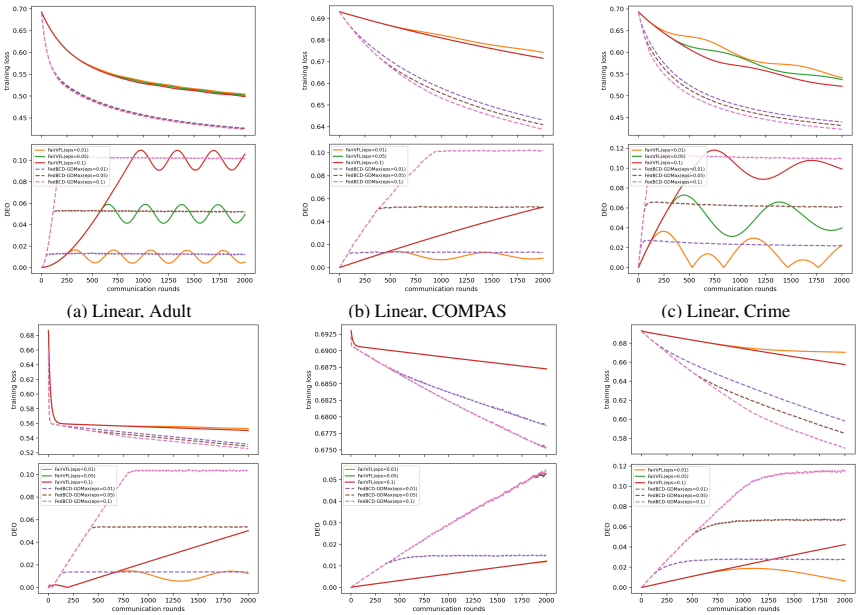
Table 3: MLP accuracy and group disparity on Adult, COMPAS, and Crime.

Method	Adult			COMPAS			Crime		
	ACC	DFP	DFN	ACC	DFP	DFN	ACC	DFP	DFN
FairVFL ($\epsilon = 0.01$)	75.99%	**0.00**	0.13	65.03%	0.09	**0.08**	77.46%	0.09	**0.01**
FairVFL ($\epsilon = 0.05$)	76.02%	**0.00**	0.11	65.15%	0.20	0.17	78.30%	0.18	0.05
FairVFL ($\epsilon = 0.10$)	76.04%	0.01	0.11	**65.59%**	0.33	0.24	79.63%	0.22	0.07
FedBCD-Max ($\epsilon = 0.01$)	78.69%	0.02	0.12	64.08%	**0.07**	**0.08**	79.47%	**0.08**	**0.01**
FedBCD-Max ($\epsilon = 0.05$)	79.85%	0.02	0.04	64.20%	0.16	0.15	79.47%	0.18	**0.01**
FedBCD-Max ($\epsilon = 0.10$)	**80.08%**	0.03	**0.03**	65.09%	0.31	0.23	**80.30%**	0.21	0.05

Communication compression with EF. On Adult, we compare dense FedBCD-Max with direct Bi-Top-K compression and BiEF-FedBCD-Max, where both compressed methods retain 30% of uplink and downlink coordinates. Unlike Theorem 2, the implementation computes the analytic dual response from the sampled mini-batch, so this experiment compares compression schemes under a common stochastic dual protocol rather than directly verifying the population theorem. Both compressed variants reduce recorded traffic by 38.2%, while BiEF largely recovers the degradation caused by direct Top-K.

Table 4: Adult results under bidirectional compression.

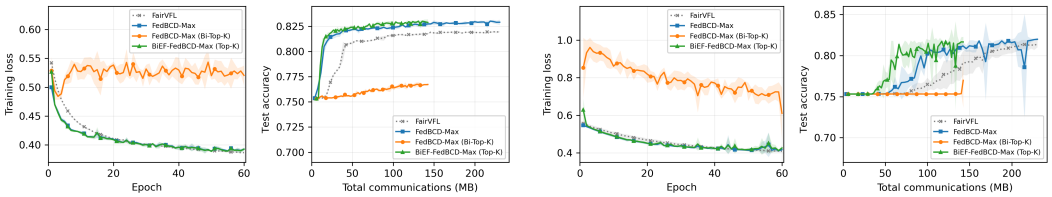
Method	Total MB	Linear loss	Linear ACC	MLP loss	MLP ACC
FairVFL	230.43	$\mathbf{0.3868 \pm 0.0008}$	0.8198 ± 0.0004	$\mathbf{0.4086 \pm 0.0063}$	0.8142 ± 0.0041
FedBCD-Max	230.43	0.3928 ± 0.0025	$\mathbf{0.8291 \pm 0.0026}$	0.4174 ± 0.0096	$\mathbf{0.8198 \pm 0.0016}$
FedBCD-Max (Bi-Top-K, no EF)	**142.43**	0.5204 ± 0.0171	0.7673 ± 0.0013	0.6104 ± 0.1387	0.7698 ± 0.0331
BiEF-FedBCD-Max	**142.43**	0.3927 ± 0.0028	0.8289 ± 0.0032	0.4263 ± 0.0147	0.8165 ± 0.0062



(a) Linear, Adult (b) Linear, COMPAS (c) Linear, Crime

(d) MLP, Adult (e) MLP, COMPAS (f) MLP, Crime

Figure 2: Loss and equality-of-opportunity surrogate-disparity $|D(x)|$ (DEO) trajectories.



(a) Linear model. (b) MLP model.

Figure 3: Adult training-loss and test-accuracy curves under bidirectional compression.

5.2 y-CLIENTS: VERTICAL ADVERSARIAL LEARNING

Formulation. For sample j, let ξ_j and η_j collect the client feature and perturbation blocks, let b_j be the server-held label, and let $\Theta = (\theta_0, \theta_1, \ldots, \theta_C)$ collect all model parameters. We solve

$$\min_{\Theta} \max_{\substack{\|\eta_{j,c}\|_\infty \le \epsilon \\ j=1,\ldots,S;\ c=1,\ldots,C}} \frac{1}{S}\sum_{j=1}^{S} \ell(f_\Theta(\xi_j + \eta_j), b_j) + \sum_{c=0}^{C} R_c(\theta_c). \tag{9}$$

For client c, define $y_c = (\eta_{1,c}, \ldots, \eta_{S,c})$, so $y = (y_1, \ldots, y_C)$ has exactly the client-partitioned structure of the local-y model. We use a two-client ResNet-18, with the labels and top classifier held by the server.

Experiment results. We evaluate CIFAR-10 following the VFAL protocol (Man et al., 2025), with PGD-r (Madry et al., 2018), FreeAT-r (Shafahi et al., 2019), FreeLB-r (Zhu et al., 2020), YOPO-m-n (Zhang et al., 2019), and DecVFAL-m-n (Man et al., 2025) implemented in synchronous VFL-CZOFO (Wang et al., 2023). Our FedBCDA performs $Q = 3$ client-local descent–ascent steps per communication round. Figure 4 shows that FedBCDA reaches competitive clean accuracy while converging rapidly in both epochs and wall-clock time, supporting the practical benefit of parallel block-local updates. Additional implementation details are provided in Appendix B and C.

Communication efficiency with EF. We compare dense FedBCDA, bidirectional Top-K without EF, and BiEF-FedBCDA under the same optimization settings, retaining 20% of uplink and downlink coordinates for both compressed variants. Figure 5 reports the mean and one standard deviation over four seeds. Direct compression degrades both training loss and clean accuracy, whereas er-

ror feedback substantially recovers this loss while retaining the same compressed communication budget.

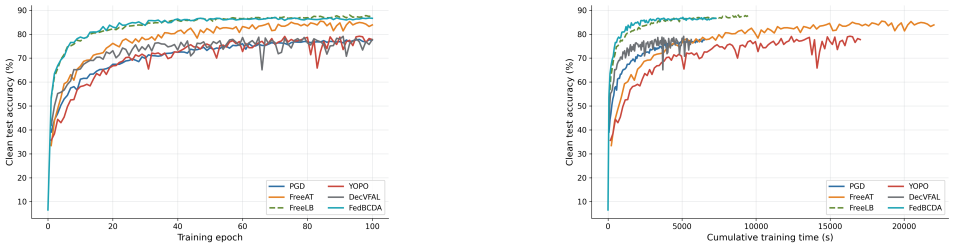


Figure 4: y-clients CIFAR-10 clean test accuracy versus training epoch (left) and wall-clock time (right).

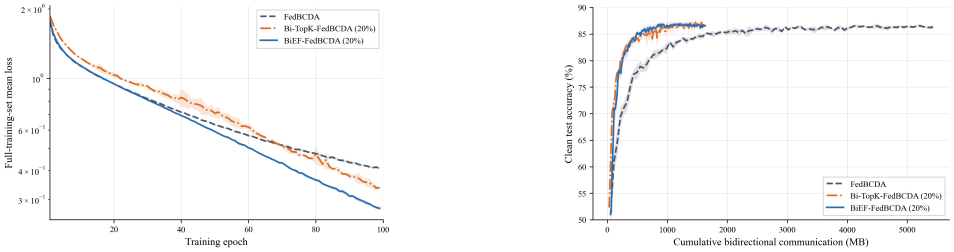


Figure 5: y-clients bidirectional compression on CIFAR-10: training loss versus epoch (left) and clean test accuracy versus communicated data (right), averaged over four seeds.

Robustness evaluation. Table 5 compares robustness under FGSM (Kurakin et al., 2017), PGD (Madry et al., 2018), CW (Carlini & Wagner, 2017), and the zeroth-order attacks ZO-FGSM and ZO-PGD based on ZOO (Chen et al., 2017), together with the CER evaluation based on randomized smoothing (Cohen et al., 2019).Complete attack parameters are provided in Appendix D. Overall, no method dominates across all attacks. The proposed methods remain competitive in both clean accuracy and adversarial robustness, while the relative advantage varies with the attack.

Table 5: y-clients CIFAR-10 robustness (%) under the unified attack protocol.

Method	Clean	FGSM	PGD-10	CW	CER	ZO-FGSM	ZO-PGD-10	s/epoch
PGD	77.44	72.48	**61.97**	77.36	73.28	73.09	75.47	78.84
FreeAT	83.88	72.36	46.84	83.04	**75.89**	76.82	80.04	220.11
FreeLB	**87.44**	51.16	35.06	81.14	66.97	69.33	75.51	109.65
YOPO	77.78	68.59	49.20	77.50	73.75	73.33	75.72	170.37
DecVFAL	77.77	69.20	52.94	77.65	73.23	72.77	74.10	66.21
FedBCDA(**Ours**)	86.13	**74.79**	47.95	85.08	75.47	**78.49**	**84.94**	**30.95**
BiEF-FedBCDA(**Ours**)	87.01	74.46	45.46	**85.93**	73.24	75.89	84.60	31.68

More experimental results on MNIST are provided in Appendix D.

6 Conclusion

We develop local block-coordinate methods for vertical minimax learning with either server-side or client-partitioned maximization variables. For SC–SC and NC–SC problems, we establish communication-round guarantees under both exact exchange and bidirectional error-feedback compression. The exact methods attain $\widetilde{O}(\kappa^2 \log(1/\epsilon))$ rounds in SC–SC and $O(1/\epsilon)$ dependence for squared primal-envelope stationarity in NC–SC, while error feedback preserves the same ϵ dependence with an explicit $O(\omega^{-2})$ round overhead. Experiments on fair classification and vertical adversarial learning show that local updates provide competitive optimization performance and that error feedback substantially mitigates the degradation caused by direct compression.

AI USE STATEMENT

In this work, we used generative AI tools, primarily ChatGPT, to assist with exploring proof strategies and analytical approaches, checking and refining mathematical arguments, surveying relevant literature, implementing parts of the experimental code, and drafting and polishing parts of the manuscript. The authors formulated the research questions and problem settings, developed the proposed methods, and evaluated and selected among candidate approaches. All mathematical claims and proofs in the final paper were manually verified by the authors. AI-assisted code used in the final experiments was reviewed and tested by the authors, and cited literature was checked against the original sources. We have reviewed all AI-assisted work and take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

Detailed algorithmic descriptions, theoretical proofs, and experimental settings are provided in the main text and appendix. An anonymized implementation and scripts for reproducing the experiments are available at https://anonymous.4open.science/r/VFL-MINIMAX-C4F7.

REFERENCES

Aleksandr Beznosikov, Peter Richtárik, Michael Diskin, Max Ryabinin, and Alexander Gasnikov. Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/5ac1428c23b5da5e66d029646ea3206d-Abstract-Conference.html. 2

Aleksandr Beznosikov, Georgiy Kormakov, Alexander Grigorievskiy, Mikhail Rudakov, Ruslan Nazykov, Alexander Rogozin, Anton Vakhrushev, Andrey Savchenko, Martin Takáč, and Alexander Gasnikov. Exploring new frontiers in vertical federated learning: the role of saddle point reformulation. *Journal of Optimization Theory and Applications*, 209, 2026. doi: 10.1007/s10957-026-02961-z. URL https://doi.org/10.1007/s10957-026-02961-z. Article 18. 1, 2

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy*, pp. 39–57. IEEE, 2017. 5.2, D

Timothy J Castiglia, Anirban Das, Shiqiang Wang, and Stacy Patterson. Compressed-VFL: Communication-efficient learning with vertically partitioned data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2738–2766. PMLR, 2022. URL https://proceedings.mlr.press/v162/castiglia22a.html. 1, 2

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26. ACM, 2017. 5.2, D

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019. 5.2, D

Yuyang Deng and Mehrdad Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1387–1395. PMLR, 2021. URL https://proceedings.mlr.press/v130/deng21a.html. 1, 2

Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. EF21 with bells & whistles: Six algorithmic extensions of modern error feedback. *Journal of Machine Learning Research*, 26(189):1–50, 2025. URL https://jmlr.org/papers/v26/24-0059.html. 2, 4

Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International conference on machine learning*, pp. 4880–4889. PMLR, 2020. 2, 1

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3252–3261. PMLR, 2019. URL https://proceedings.mlr.press/v97/karimireddy19a.html. 2

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 2020. URL https://proceedings.mlr.press/v119/karimireddy20a.html. 2

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *International Conference on Learning Representations*, 2017. 5.2, D

Jaewook Lee, Hanseul Cho, and Chulhee Yun. Fundamental benefit of alternating updates in minimax optimization. *arXiv preprint arXiv:2402.10475*, 2024. 1

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6083–6093. PMLR, 2020a. URL https://proceedings.mlr.press/v119/lin20a.html. 2, 1, A.7.1

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on learning theory*, pp. 2738–2779. PMLR, 2020b. 2

Changxin Liu, Zhenan Fan, Zirui Zhou, Yang Shi, Jian Pei, Lingyang Chu, and Yong Zhang. Achieving model fairness in vertical federated learning. *arXiv preprint arXiv:2109.08344*, 2022a. URL https://arxiv.org/abs/2109.08344v3. Version 3; first posted in 2021. 1, 1, 1, 2, 5.1

Xiaorui Liu, Yao Li, Jiliang Tang, and Ming Yan. A double residual compression algorithm for efficient distributed learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 133–143. PMLR, 2020. URL https://proceedings.mlr.press/v108/liu20a.html. 2

Yang Liu, Xinwei Zhang, Yan Kang, Liping Li, Tianjian Chen, Mingyi Hong, and Qiang Yang. FedBCD: A communication-efficient collaborative learning framework for distributed features. *IEEE Transactions on Signal Processing*, 70:4277–4290, 2022b. doi: 10.1109/TSP.2022.3198176. URL https://doi.org/10.1109/TSP.2022.3198176. 1, 1, 2, 2

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. 1

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb. 5.2, 5.2, D

Tianxing Man, Yu Bai, Ganyu Wang, Jinjie Fang, Haoran Fang, Bin Gu, and Yi Chang. Accelerated vertical federated adversarial learning through decoupling layer-wise dependencies. In *Advances in Neural Information Processing Systems*, volume 38, pp. 147561–147593, 2025. doi: 10.52202/085713-4934. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d8ace30c68b085556ccce04ed4ae4ebb-Abstract-Conference.html. 1, 2, 2, 5.2

Tianxing Man, Jinjie Fang, Ganyu Wang, Yu Bai, Zhaogeng Liu, Bin Gu, and Yi Chang. CE-VFAL: A novel framework for communication-efficient vertical federated adversarial learning. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 697–705. International Joint Conferences on Artificial Intelligence Organization, 8 2026. doi: 10.24963/ijcai.2026/79. URL https://doi.org/10.24963/ijcai.2026/79. Main Track. 2, 1, 2

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL https://proceedings.mlr.press/v54/mcmahan17a.html. 2

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013. 3

Tao Qi, Fangzhao Wu, Chuhan Wu, Lingjuan Lyu, Tong Xu, Hao Liao, Zhongliang Yang, Yongfeng Huang, and Xing Xie. FairVFL: A fair vertical federated learning framework with contrastive adversarial learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/333a7697dbb67f09249337f81c27d749-Abstract-Conference.html. 2

Peter Richtarik, Igor Sokolov, and Ilyas Fatkhullin. EF21: A new, simpler, theoretically better, and practically faster error feedback. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4384–4396. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/231141b34c82aa95e48810a9d1b33a79-Paper.pdf. 2, 4, 4, 2

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, volume 32, 2019. 5.2

Pranay Sharma, Rohan Panda, Gauri Joshi, and Pramod Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19683–19730. PMLR, 2022. URL https://proceedings.mlr.press/v162/sharma22c.html. 1, 2

Sergey Stanko, Timur Karimullin, Aleksandr Beznosikov, and Alexander Gasnikov. Accelerated methods with compression for horizontal and vertical federated learning. *Journal of Optimization Theory and Applications*, 208, 2026. doi: 10.1007/s10957-025-02893-0. URL https://doi.org/10.1007/s10957-025-02893-0. Article 68. 2

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1805.09767. 2

Sebastian U. Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/hash/b440509a0106086a67bc2ea9df0a1dab-Abstract.html. 2

Pedro Valdeira, João Xavier, Cláudia Soares, and Yuejie Chi. Communication-efficient vertical federated learning via compressed error feedback. *IEEE Transactions on Signal Processing*, 73: 1065–1080, 2025. doi: 10.1109/TSP.2025.3540655. URL https://doi.org/10.1109/TSP.2025.3540655. 1, 1, 2, 4, 4, 4

Ganyu Wang, Bin Gu, Qingsong Zhang, Xiang Li, Boyu Wang, and Charles X. Ling. A unified solution for privacy and communication efficiency in vertical federated learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 13480–13491, 2023. doi: 10.52202/075280-0593. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/2b5af479527167d4af78847a9b9b645f-Abstract.html. 2, 5.2

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International conference on artificial intelligence and statistics*, pp. 5485–5517. PMLR, 2022. 2

Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 5.2

Guodong Zhang, Yuanhao Wang, Laurent Lessard, and Roger B. Grosse. Near-optimal local convergence of alternating gradient descent-ascent for minimax optimization. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 7659–7679. PMLR, 2022. URL https://proceedings.mlr.press/v151/zhang22e.html. 1, A.7.1

Siqi Zhang and Yifan Hu. Avoid overclaims: Summary of complexity bounds for algorithms in minimization and minimax optimization. In *The Fourth Blogpost Track at ICLR 2025*, 2025. 2

Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in artificial intelligence*, pp. 482–492. PMLR, 2021. 2

Siqi Zhang, Sayantan Choudhury, Sebastian U. Stich, and Nicolas Loizou. Communication-efficient gradient descent-accent methods for distributed variational inequalities: Unified analysis and local updates. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hORCalGn3Z. 1, 2

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. FreeLB: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BygzbyHFvB. 5.2

A Complete Proofs and Technical Details

A.1 Notation used in the theory

Table 6 collects the symbols that recur across Sections 3–4 and the appendix. Proof-only potentials, tracking errors, and numerical constants are introduced locally.

Table 6: Notation used in the theory.

Notation	Meaning	Notation	Meaning
Problem and minimax geometry			
$K, [K]$	Number of clients; $[K] = \{1, \dots, K\}$	x_i, x	Client-i primal block; stacked primal variable
y, y_i	Global/local maximization variables	z_i, z	Local and stacked primal–dual variables
$\mathcal{Y}, \mathcal{Y}_i$	Maximization feasible sets	F_{g}, F_ℓ	Global-y/local-y objectives
$\Phi(x)$	Primal envelope $\max_y F(x, y)$	$y^\star(x)$	Maximizer of $F(x, \cdot)$
$\Phi^\star$	$\inf_x \Phi(x)$	Δ_Φ	$\Phi(x^0) - \Phi^\star$
$\Psi(y)$	$\min_x F(x, y)$ in SC–SC	Gap	SC–SC primal–dual gap; cf. Equation (4)
$z^\star$	Unique SC–SC saddle point	$\mathcal{W}_t$	Local-y NC–SC proof potential
Smoothness, rounds, and stochastic oracles			
L, μ	Smoothness and strong-curvature constants	κ	Condition number L/μ
Q	Local substeps per communication round	T	Number of communication rounds
η_x, η_y	Per-substep steps	α_x, α_y	Effective steps $Q\eta_x, Q\eta_y$
t, q	Round and local-substep indices	$B, \mathcal{B}_t$	Batch size and round-t mini-batch
R	Uniform random output round	$\sigma_{\mathrm{g}}^2, \sigma_\ell^2$	Global/local oracle variances
ε_t	Round-t inner-solve tolerance	$\varepsilon_{\mathrm{out}}$	Final inner-solve tolerance
Bidirectional error feedback			
u_i	Communicated client state: x_i or z_i	$\widehat{u}_i^t$	Server tracker of u_i^t
m_i^t	Exact downlink sufficient message	$\widehat{m}_i^t$	Downlink EF tracker
$\mathcal{M}_i$	Sufficient-message map	$\mathcal{R}_i$	Client oracle reconstruction map
$\mathcal{C}_i^{\mathrm{up}}, \mathcal{C}_i^{\mathrm{down}}$	Uplink/downlink compressors	ω	Common contractivity parameter

A.2 Proof-facing stochastic and communication conditions

The main text states the optimization assumptions and the communication interface; here we make the conditioning used in the proofs explicit. Let $\mathcal{F}_t$ contain all iterates, the EF trackers $\{\widehat{u}_i^t, \widehat{m}_i^t\}_{i=1}^K$, and all randomness revealed before round t. In the global-y model, y^t is $\mathcal{F}_t$-measurable and the outer mini-batch is sampled afterwards. The uplink compressor is applied only after the local trajectory u^{t+1} has been formed; conditional on that uplink realization, the server forms $m^{t+1} = \mathcal{M}(\widehat{u}^{t+1})$ and then applies the downlink compressor. Thus every use of contraction conditions (Assumption 3 (3)) on the full pre-compression history together with the realized input.

Write $g^{t,q} := (g_{x_i}^{t,q})_{i=1}^K$ in the global-y model and $g^{t,q} := (g_{x_i}^{t,q}, g_{y_i}^{t,q})_{i=1}^K$ in the local-y model. Let $g^t := g^{t,0}$ and $\overline{g}^t := \mathbb{E}[g^t \mid \mathcal{F}_t]$, where the conditioning fixes whichever synchronized or tracked communication state is used to build the round oracle. In the global-y model, for both exact and EF communication,

$$\mathbb{E}[\|g^t - \overline{g}^t\|^2 \mid \mathcal{F}_t] \le \frac{\sigma_{\mathrm{g}}^2}{B};$$

under exact exchange $\overline{g}^t = \nabla_x F(x^t, y^t)$, whereas under EF $\overline{g}^t$ is the corresponding population reconstructed oracle and its deterministic tracker bias is controlled separately below. In the local-y model, for either communication mode, define $\xi^t := g^t - \overline{g}^t$ and assume

$$\mathbb{E}[\xi^t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\xi^t\|^2 \mid \mathcal{F}_t] \le \frac{\sigma_\ell^2}{B}.$$

Under exact exchange, $\overline{g}^t$ is the population stale-block oracle at the synchronized state; under EF, it is the population reconstructed oracle at the tracked state. In both cases the centered inequalities above remove only mini-batch noise. The same mini-batch may be reused over all Q substeps. Population, sample, stale-block, and reconstructed oracle maps are uniformly L-Lipschitz in the blocks updated within the round. These statements are imposed before projection; no post-projection unbiasedness or stepsize-uniform Lipschitz property of a projected-gradient mapping is required.

The normalized communication interface in Assumption 3, including its extension to all EF-reachable tracker states, is used through

$$\|\mathcal{M}(a) - \mathcal{M}(a')\| \le \|a - a'\|.$$

Let $\overline{g}^t_\circ$ denote the conditional-mean client oracle built from the current uplink tracker but with the exact downlink target m_i^t. Consistency and L-Lipschitzness of the reconstruction imply

$$\|\overline{g}^t_\circ - \nabla_x F(x^t, y^t)\| \le L\sqrt{D_t^{\mathrm{up}}} \qquad \text{(global-}y\text{)},$$

$$\|\overline{g}^t_\circ - \nabla F(z^t)\| \le L\sqrt{D_t^{\mathrm{up}}} \qquad \text{(local-}y\text{)},$$

$$\|\nabla_y \widehat{F}_t(y) - \nabla_y F(x^t, y)\| \le L\sqrt{D_t^{\mathrm{up}}} \qquad \text{(global-}y\text{)},$$

and replacing m_i^t by $\widehat{m}_i^t$ changes the stacked client oracle by at most $L\sqrt{D_t^{\mathrm{down}}}$. These are conditions on the stacked interface: coordinatewise Lipschitz bounds must be combined with their actual dimension and client-count factors before selecting L. The exact message $m_i^t = \mathcal{M}_i(\widehat{u}^t)$ is formed independently of $\mathcal{B}_t$; the mini-batch and, in the global-y model, y^t enter only the reconstruction map $\mathcal{R}_i$. The clients must have access to these arguments, and any transmissions needed to supply them count toward the communication payload in Appendix A.7.2.

For every global-y surrogate reachable by the EF recursion, Assumption 3 gives a unique maximizer $\widehat{y}_t^\star$ and

$$\widehat{F}_t(\widehat{y}_t^\star) - \widehat{F}_t(y) \ge \frac{\mu}{2}\|y - \widehat{y}_t^\star\|^2, \qquad y \in \mathcal{Y}_{\mathrm{g}}. \tag{10}$$

Only this quadratic-growth inequality is used to turn a certified surrogate maximization gap into a distance; strong concavity of every surrogate is sufficient but not needed.

For the EF proofs define

$$D_t^{\mathrm{up}} := \sum_i \|u_i^t - \widehat{u}_i^t\|^2, \qquad D_t^{\mathrm{down}} := \sum_i \|m_i^t - \widehat{m}_i^t\|^2,$$

and

$$\beta_\omega := (1-\omega)(1+2/\omega), \qquad \theta_\omega := \frac{\omega}{\omega + 24\beta_\omega},$$

$$D_t := D_t^{\mathrm{up}} + \theta_\omega D_t^{\mathrm{down}}, \qquad \bar{\beta}_\omega := \beta_\omega\left(1 + \frac{12\theta_\omega}{\omega}\right), \qquad \Lambda_\omega := L\sqrt{1 + \theta_\omega^{-1}}. \tag{11}$$

At $t = T$ no further downlink is needed, so we set $D_T := D_T^{\mathrm{up}}$. These quantities are proof-local and do not enlarge the algorithmic state.

A.3 Finite-time statements corresponding to the main theorems

The following statements retain the stochastic floors, inner tolerances, and initialization factors suppressed in the headline communication orders.

Theorem 3 (Exact SC–SC finite-time primal–dual-gap bounds)**.** *Let $\varepsilon_{\max} := \max_{t<T}\varepsilon_t$. For FedBCD-Max, assume the effective global step satisfies Proposition 1. After round T, run one server-only inner solve at x^T and return y^T with inner gap at most $\varepsilon_{\mathrm{out}}$. Then*

$$\mathbb{E}\mathrm{Gap}(x^T, y^T) \le \rho_{\mathrm{ex}}\left[e^{-\alpha_x \mu T/2}\Delta_\Phi + \frac{4\sigma_{\mathrm{g}}^2}{\mu B} + \frac{8L^2}{\mu^2}\varepsilon_{\max}\right] + (1+2\kappa^2)\varepsilon_{\mathrm{out}}, \tag{12}$$

where $\rho_{\mathrm{ex}} := 2\kappa(1+\kappa)$. For FedBCDA, for every $Q \ge 1$, under Equation (64),

$$\mathbb{E}\mathrm{Gap}(z^T) \le e^{-c\mu\alpha T}(1 + 128\kappa^2)\mathrm{Gap}(z^0) + C\frac{\kappa^2\sigma_\ell^2}{\mu B}, \tag{13}$$

where $c, C > 0$ are universal.

Theorem 4 (Exact NC–SC finite-time stationarity bounds). *Let $R \sim \mathrm{Unif}\{0, \dots, T-1\}$ be independent of the algorithm and $\bar{\varepsilon}_T := T^{-1} \sum_{t<T} \varepsilon_t$. Under the step conditions of Propositions 1 and 4,*

$$\mathbb{E}\|\nabla\Phi(x^R)\|^2 \le \frac{16\Delta_\Phi}{7\alpha_x T} + \frac{32\sigma_{\mathrm{g}}^2}{7B} + \frac{64L^2}{7\mu}\bar{\varepsilon}_T, \tag{14}$$

$$\mathbb{E}\|\nabla\Phi(x^R)\|^2 \le \frac{C\mathcal{W}_0}{\alpha_x T} + \frac{C\kappa^2\sigma_\ell^2}{B}, \tag{15}$$

for global-y and local-y, respectively, where C is universal and $\mathcal{W}_0$ is the local-y potential in Proposition 4.

Theorem 5 (EF SC–SC finite-time primal–dual-gap bounds). *For EF-FedBCD-Max, suppose that α_x satisfies Equation (45), and write $V_T^{\mathrm{g}} := \mathbb{E}\mathcal{V}_T^{\mathrm{g}}$ for the potential in Proposition 3. Then*

$$V_T^{\mathrm{g}} \le e^{-\alpha_x\mu T/8} V_0^{\mathrm{g}} + \frac{8\sigma_{\mathrm{g}}^2}{\mu B} + \frac{24C_\varepsilon}{\mu}\varepsilon_{\max}. \tag{16}$$

After a final server-only surrogate inner solve with certificate $\varepsilon_{\mathrm{out}}$,

$$\mathbb{E}\mathrm{Gap}(x^T, y^T) \le \rho_{\mathrm{ef}}(\alpha_x, \omega) V_T^{\mathrm{g}} + (4\kappa^2 + 2)\varepsilon_{\mathrm{out}}, \tag{17}$$

where

$$\rho_{\mathrm{ef}}(\alpha, \omega) := 2\kappa(1+\kappa) + \left(2\kappa^2 + \frac{3}{2}\right)\frac{\omega}{48\alpha\mu(1+\kappa)^2}. \tag{18}$$

For EF-FedBCDA, for every $Q \ge 1$, under Equation (66),

$$\mathbb{E}\mathrm{Gap}(z^T) \le e^{-c\mu\alpha T}(1 + 128\kappa^2)\mathrm{Gap}(z^0) + C\frac{\kappa^2\sigma_\ell^2}{\mu B}, \tag{19}$$

provided the trackers are initialized exactly, so $D_0^{\mathrm{up}} = D_0^{\mathrm{down}} = D_0 = 0$.

Theorem 6 (EF NC–SC finite-time stationarity bounds). *Under the step conditions of Propositions 3 and 4,*

$$\mathbb{E}\|\nabla\Phi(x^R)\|^2 \le \frac{16V_0^{\mathrm{g}}}{\alpha_x T} + 32C_\varepsilon\bar{\varepsilon}_T + \frac{8\sigma_{\mathrm{g}}^2}{B}, \tag{20}$$

$$\mathbb{E}\|\nabla\Phi(x^R)\|^2 \le \frac{CV_0^\ell}{\alpha_x T} + \frac{C\kappa^2\sigma_\ell^2}{B}, \tag{21}$$

where $C_\varepsilon := 4L^2/\mu$, V_0^{g} is the initial global EF Lyapunov value, and V_0^ℓ is the initial local EF Lyapunov value. With exact tracker initialization, $D_0 = 0$, hence $V_0^{\mathrm{g}} = \Delta_\Phi$ and $V_0^\ell = \mathcal{W}_0$.

A.4 Common technical tools

This appendix proves every analytic estimate used later. All conditional expectations are taken with respect to the filtration in Assumption 2. For local-y, stochasticity is analyzed before projection through the smooth signed saddle gradient $\mathcal{T}_F = (\nabla_x F, -\nabla_y F)$.

A.4.1 Standard constrained-envelope facts

We use the standard variational inequality and firm nonexpansiveness of Euclidean projection onto a nonempty closed convex set. Under Assumption 1, the constrained maximizer $y^\star(x)$ is unique and the usual strong-concavity and sensitivity bounds give, for every x, x' and $y \in \mathcal{Y}$,

$$F(x, y^\star(x)) - F(x, y) \ge \frac{\mu}{2}\|y - y^\star(x)\|^2, \tag{22}$$

$$\|y^\star(x) - y^\star(x')\| \le \frac{L}{\mu}\|x - x'\| = \kappa\|x - x'\|. \tag{23}$$

The single-valued constrained Danskin formula yields

$$\nabla\Phi(x) = \nabla_x F(x, y^\star(x)), \qquad \|\nabla\Phi(x) - \nabla\Phi(x')\| \le L_\Phi\|x - x'\|, \quad L_\Phi = L(1+\kappa). \tag{24}$$

In SC–SC, Φ is μ-strongly convex and hence

$$2\mu(\Phi(x) - \Phi^\star) \le \|\nabla\Phi(x)\|^2 \le 2L_\Phi(\Phi(x) - \Phi^\star). \tag{25}$$

Moreover, the maximization gap controls both the dual distance and the primal gradient mismatch:

$$\|y - y^\star(x)\|^2 \le \frac{2}{\mu}\,[\Phi(x) - F(x,y)], \tag{26}$$

$$\|\nabla_x F(x,y) - \nabla\Phi(x)\|^2 \le \frac{2L^2}{\mu}\,[\Phi(x) - F(x,y)]. \tag{27}$$

These facts apply to both models; in global-y, an ε_t-certified inner solve simply makes the bracketed quantity at most ε_t.

Lemma 1 (Perturbed projected ascent at fixed x)**.** *Let ψ be L-smooth and μ-strongly concave on $\mathcal{Y}$, let $y_\star = \arg\max_{\mathcal{Y}} \psi$, and set*

$$y^+ = \Pi_{\mathcal{Y}}\big(y + \eta(\nabla\psi(y) + d)\big).$$

If $\eta L \le 1/4$, then

$$\psi(y_\star) - \psi(y^+) \le \Big(1 - \frac{\mu\eta}{4}\Big)\,[\psi(y_\star) - \psi(y)] - \frac{\|y^+ - y\|^2}{8\eta} + 4\eta\|d\|^2. \tag{28}$$

Proof. Set $q = -\psi$, $e = -d$, $s = y^+ - y$, $\theta = \mu\eta$, and $w = (1-\theta)y + \theta y_\star \in \mathcal{Y}$. Smoothness, strong convexity, and the projection variational inequality give

$$q(y^+) \le q(w) + \left(\frac{1}{2\eta} - \frac{\mu}{2}\right)\|w - y\|^2 - \frac{\|y^+ - w\|^2}{2\eta} - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|s\|^2 - \langle e, y^+ - w\rangle.$$

By strong convexity, $q(w) \le (1-\theta)q(y) + \theta q(y_\star) - (\mu/2)\theta(1-\theta)\|y - y_\star\|^2$. Since $\theta = \mu\eta$, the resulting coefficient of $\|y - y_\star\|^2$ cancels. Completing the square in $y^+ - w$ yields

$$q(y^+) - q(y_\star) \le (1-\theta)[q(y) - q(y_\star)] - \left(\frac{1}{2\eta} - \frac{L}{2}\right)\|s\|^2 + \frac{\eta}{2}\|e\|^2.$$

Use $\eta L \le 1/4$ and weaken the coefficients to obtain Equation (28). □

A.4.2 One local-round drift bound

Lemma 2 (Global-y Q-step drift)**.** *For global-y, let $\alpha = Q\eta_x$. If each round-start stale smooth x_i-gradient is L-Lipschitz in its updated block and $\alpha L \le 1/4$, then*

$$x^{t+1} = x^t - \alpha g^t + r^t, \qquad \|r^t\| \le 2\alpha^2 L\|g^t\|. \tag{29}$$

Proof. Write $g_{i,q} = g_{x_i}^{t,q}$. Telescoping the local steps gives the first identity with $r_i^t = -\eta_x \sum_{q<Q}(g_{i,q} - g_{i,0})$. The usual discrete Gronwall bound for an L-Lipschitz local gradient gives

$$\max_{q\le Q}\|x_i^{t,q} - x_i^t\| \le \frac{\alpha}{1 - \alpha L}\|g_{i,0}\| \le 2\alpha\|g_{i,0}\|.$$

Thus $\|r_i^t\| \le 2\alpha^2 L\|g_{i,0}\|$; stacking clients proves Equation (29). □

A.4.3 Communication bias and bidirectional EF tracking

Lemma 3 (Bidirectional communication-state bias)**.** *Let $\overline{g}^t$ be the conditional mean of the round-start oracle actually used by the clients, i.e., with the downlink memory $\widehat{m}_i^t$. Then for global-y,*

$$\|\overline{g}^t - \nabla_x F(x^t, y^t)\| \le \Lambda_\omega\sqrt{D_t}, \qquad \|\nabla_y \widehat{F}_t(y) - \nabla_y F(x^t, y)\| \le \Lambda_\omega\sqrt{D_t}, \tag{30}$$

and for local-y,

$$\|\overline{g}^t - \nabla F(z^t)\| \le \Lambda_\omega\sqrt{D_t}. \tag{31}$$

Consequently, in global-y EF, with

$$C_D := 2\Lambda_\omega^2(1+\kappa)^2, \qquad C_\varepsilon := \frac{4L^2}{\mu}, \tag{32}$$

we have

$$\|\overline{g}^t - \nabla\Phi(x^t)\|^2 \le C_D D_t + C_\varepsilon\varepsilon_t. \tag{33}$$

Proof. For either model, let $\overline{g}_\circ^t$ denote the conditional-mean client oracle built from the same uplink memories but with the exact downlink targets m_i^t. The proof-facing communication conditions in Appendix A.2 give

$$\|\overline{g}_\circ^t - \nabla_x F(x^t, y^t)\| \le L\sqrt{D_t^{\mathrm{up}}}, \qquad \|\overline{g}_\circ^t - \nabla F(z^t)\| \le L\sqrt{D_t^{\mathrm{up}}},$$

and replacing m_i^t by $\widehat{m}_i^t$ changes the corresponding stacked client oracle by at most $L\sqrt{D_t^{\mathrm{down}}}$. Therefore

$$L\sqrt{D_t^{\mathrm{up}}} + L\sqrt{D_t^{\mathrm{down}}} \le L\sqrt{1+\theta_\omega^{-1}}\,\sqrt{D_t^{\mathrm{up}} + \theta_\omega D_t^{\mathrm{down}}} = \Lambda_\omega\sqrt{D_t}.$$

The server y-gradient is unaffected by downlink compression and satisfies

$$\|\nabla_y \widehat{F}_t(y) - \nabla_y F(x^t, y)\| \le L\sqrt{D_t^{\mathrm{up}}} \le \Lambda_\omega\sqrt{D_t}.$$

This proves the global- and local-state bias bounds.

For the global envelope comparison, let $\widehat{y}_t^\star = \arg\max_{y\in\mathcal{Y}_{\mathrm{g}}} \widehat{F}_t(y)$. The reachable-surrogate quadratic-growth condition Equation (10) and the ε_t-maximization guarantee give

$$\|y^t - \widehat{y}_t^\star\| \le \sqrt{\frac{2\varepsilon_t}{\mu}}. \tag{34}$$

The two constrained optimality conditions and strong concavity of $F(x^t, \cdot)$ give

$$\|\widehat{y}_t^\star - y^\star(x^t)\| \le \frac{L}{\mu}\sqrt{D_t^{\mathrm{up}}} \le \frac{\Lambda_\omega}{\mu}\sqrt{D_t}. \tag{35}$$

Combining the first estimate in Equation (30), Equation (34), and Equation (35) yields

$$\|\overline{g}^t - \nabla\Phi(x^t)\| \le \Lambda_\omega(1+\kappa)\sqrt{D_t} + L\sqrt{2\varepsilon_t/\mu}.$$

Finally, $(a+b)^2 \le 2a^2 + 2b^2$ proves Equation (33). □

Lemma 4 (Bidirectional EF memory recursion)**.** *With the proof-level quantities in Appendix A.2, for every $t < T$,*

$$\mathbb{E}[D_{t+1} \mid u^{t+1}, \mathcal{F}_t] \le \left(1 - \frac{\omega}{4}\right) D_t + \bar{\beta}_\omega \|u^{t+1} - u^t\|^2, \tag{36}$$

where at $t = T-1$ the terminal convention $D_T = D_T^{\mathrm{up}}$ is used. Moreover,

$$\bar{\beta}_\omega \le \beta_\omega + \frac{1}{2} \le \frac{5}{2\omega}. \tag{37}$$

Proof. Set $e_i^t = u_i^t - \widehat{u}_i^t$ and $\delta_i^t = u_i^{t+1} - u_i^t$. The uplink EF update and common-ω contractivity give

$$\mathbb{E}[D_{t+1}^{\mathrm{up}} \mid u^{t+1}, \mathcal{F}_t] \le \left(1 - \frac{\omega}{2}\right) D_t^{\mathrm{up}} + \beta_\omega \|u^{t+1} - u^t\|^2. \tag{38}$$

For $t < T-1$, condition first on the realized uplink update. The downlink EF update gives

$$\mathbb{E}[D_{t+1}^{\mathrm{down}} \mid \widehat{u}^{t+1}, u^{t+1}, \mathcal{F}_t] \le \left(1 - \frac{\omega}{2}\right) D_t^{\mathrm{down}} + \beta_\omega \|m^{t+1} - m^t\|^2.$$

By nonexpansiveness of the sufficient-message map,

$$\|m^{t+1} - m^t\|^2 \le \|\widehat{u}^{t+1} - \widehat{u}^t\|^2.$$

For the uplink compressor input $v^t = e^t + \delta^t$, contractivity implies $\mathbb{E}\|\mathcal{C}^{\mathrm{up}}(v^t)\|^2 \le 4\|v^t\|^2$. Young's inequality with parameter $\omega/2$ and $\|\delta^t\| \le \|u^{t+1} - u^t\|$ therefore give

$$\mathbb{E}\|\widehat{u}^{t+1} - \widehat{u}^t\|^2 \le 6D_t^{\mathrm{up}} + \frac{12}{\omega}\|u^{t+1} - u^t\|^2. \tag{39}$$

Taking the tower expectation yields

$$\mathbb{E}[D_{t+1}^{\mathrm{down}} \mid u^{t+1}, \mathcal{F}_t] \le \left(1 - \frac{\omega}{2}\right) D_t^{\mathrm{down}} + 6\beta_\omega D_t^{\mathrm{up}} + \frac{12\beta_\omega}{\omega}\|u^{t+1} - u^t\|^2. \tag{40}$$

Multiply Equation (40) by θ_ω and add Equation (38). By the definition of θ_ω, $6\theta_\omega\beta_\omega \le \omega/4$, so both memory coordinates contract at least at rate $\omega/4$. The path coefficient is exactly $\bar{\beta}_\omega$. This proves Equation (36) for $t < T-1$. At $t = T-1$, no downlink update is performed; using $D_T = D_T^{\mathrm{up}}$, Equation (38), $D_{T-1}^{\mathrm{up}} \le D_{T-1}$, and $\beta_\omega \le \bar{\beta}_\omega$ gives the same inequality. Finally,

$$\bar{\beta}_\omega = \beta_\omega + \frac{12\beta_\omega}{\omega + 24\beta_\omega} \le \beta_\omega + \frac{1}{2},$$

and $\beta_\omega \le 2/\omega$ proves Equation (37). □

A.4.4 Exact global-y recursions

Proposition 1 (Exact global-y recursions). *Let $\alpha \le \min\{L_\Phi^{-1}, (8L)^{-1}\}$. Then*

$$\mathbb{E}[\Phi(x^{t+1}) - \Phi^\star \mid \mathcal{F}_t] \le \left(1 - \frac{\alpha\mu}{2}\right)(\Phi(x^t) - \Phi^\star) + \frac{2\alpha\sigma_{\mathrm{g}}^2}{B} + \frac{4\alpha L^2}{\mu}\varepsilon_t \quad \textit{(SC–SC)}, \tag{41}$$

$$\mathbb{E}[\Phi(x^{t+1}) \mid \mathcal{F}_t] \le \Phi(x^t) - \frac{7\alpha}{16}\|\nabla\Phi(x^t)\|^2 + \frac{2\alpha\sigma_{\mathrm{g}}^2}{B} + \frac{4\alpha L^2}{\mu}\varepsilon_t \quad \textit{(NC–SC)}. \tag{42}$$

Proof. Let $\delta_t = \nabla_x F(x^t, y^t) - \nabla\Phi(x^t)$ and $e_{\mathrm{stat}}^t = g^t - \overline{g}^t$. Combining the drift bound with conditional centering gives the standard inexact-gradient estimate

$$\mathbb{E}[\Phi(x^{t+1}) \mid \mathcal{F}_t] \le \Phi(x^t) - \frac{7\alpha}{16}\|\nabla\Phi(x^t)\|^2 + 2\alpha\|\delta_t\|^2 + \frac{2\alpha\sigma_{\mathrm{g}}^2}{B}.$$

Indeed, this is the L_Φ-smooth descent inequality after using $\|r^t\| \le 2\alpha^2 L\|g^t\|$ and $\mathbb{E}\|\delta_t + e_{\mathrm{stat}}^t\|^2 = \|\delta_t\|^2 + \mathbb{E}\|e_{\mathrm{stat}}^t\|^2$. By Equation (27) and the inner certificate, $\|\delta_t\|^2 \le 2L^2\varepsilon_t/\mu$, which proves Equation (42). In SC–SC, $\|\nabla\Phi(x^t)\|^2 \ge 2\mu(\Phi(x^t) - \Phi^\star)$ yields the first recursion. □

A.4.5 Global-y EF recursions

Proposition 2 (Global-y EF recursions). *Let C_D, C_ε be as in Equation (32) and suppose $\alpha L \le 1/8$. Then*

$$\mathbb{E}[D_{t+1} \mid \mathcal{F}_t] \le \left(1 - \frac{\omega}{4} + 4\bar{\beta}_\omega\alpha^2 C_D\right) D_t + 4\bar{\beta}_\omega\alpha^2\|\nabla\Phi(x^t)\|^2 + 4\bar{\beta}_\omega\alpha^2 C_\varepsilon\varepsilon_t + \frac{2\bar{\beta}_\omega\alpha^2\sigma_{\mathrm{g}}^2}{B}, \tag{43}$$

$$\mathbb{E}[\Phi(x^{t+1}) \mid \mathcal{F}_t] \le \Phi(x^t) - \frac{\alpha}{8}\|\nabla\Phi(x^t)\|^2 + \frac{3\alpha C_D}{2}D_t + \frac{3\alpha C_\varepsilon}{2}\varepsilon_t + \frac{\alpha\sigma_{\mathrm{g}}^2}{4B}, \tag{44}$$

provided additionally $\alpha \le (12L_\Phi)^{-1}$ for Equation (44). In SC–SC, the gradient term in Equation (44) is at most $-(\alpha\mu/4)(\Phi(x^t) - \Phi^\star)$.

Proof. Let $b_t = \overline{g}^t - \nabla\Phi(x^t)$ and $e_{\mathrm{stat}}^t = g^t - \overline{g}^t$. Then $g^t = \nabla\Phi(x^t) + b_t + e_{\mathrm{stat}}^t$ and $\|b_t\|^2 \le C_D D_t + C_\varepsilon\varepsilon_t$. Consequently,

$$\mathbb{E}[\|g^t\|^2 \mid \mathcal{F}_t] \le 2\|\nabla\Phi(x^t)\|^2 + 2C_D D_t + 2C_\varepsilon\varepsilon_t + \frac{\sigma_{\mathrm{g}}^2}{B}, \qquad \|x^{t+1} - x^t\|^2 \le 2\alpha^2\|g^t\|^2.$$

The first estimate in the proposition follows by substitution in the memory recursion. For the second, apply the standard L_Φ-smooth descent inequality to $x^{t+1} = x^t - \alpha g^t + r^t$, use $\|r^t\| \le 2\alpha^2 L\|g^t\|$, and collect terms under $\alpha L \le 1/8$ and $\alpha L_\Phi \le 1/12$. This gives exactly Equation (44); Equation (25) gives its SC–SC form. □

Proposition 3 (Global-y EF Lyapunov recursion). *Let C_D, C_ε be as in Equation (32). Suppose*

$$\alpha \le \min\left\{\frac{1}{8L}, \frac{1}{12L_\Phi}, \frac{\omega}{\mu}, \sqrt{\frac{\omega}{1536\bar{\beta}_\omega C_D}}\right\}, \tag{45}$$

where the square-root restriction is omitted when $\bar{\beta}_\omega = 0$, and define

$$\mathcal{V}_t^{\mathrm{g}} := \Phi(x^t) - \Phi^\star + \frac{24\alpha C_D}{\omega}D_t.$$

Then the global-y EF iterates satisfy

$$\mathbb{E}[\mathcal{V}_{t+1}^{\mathrm{g}} \mid \mathcal{F}_t] \le \left(1 - \frac{\alpha\mu}{8}\right) \mathcal{V}_t^{\mathrm{g}} + 3\alpha C_\varepsilon \varepsilon_t + \alpha \frac{\sigma_{\mathrm{g}}^2}{B} \qquad \textit{in SC–SC}, \tag{46}$$

$$\mathbb{E}[\mathcal{V}_{t+1}^{\mathrm{g}} \mid \mathcal{F}_t] \le \mathcal{V}_t^{\mathrm{g}} - \frac{\alpha}{16}\|\nabla\Phi(x^t)\|^2 + 2\alpha C_\varepsilon \varepsilon_t + \frac{\alpha}{2}\frac{\sigma_{\mathrm{g}}^2}{B} \qquad \textit{in NC–SC}. \tag{47}$$

Proof. Multiply Equation (43) by $24\alpha C_D/\omega$ and add Equation (44). Under Equation (45), the retained memory decrease is at least $\omega/8$, and the stationarity coefficient is at least $\alpha/16$. The resulting perturbation is bounded by $3\alpha C_\varepsilon \varepsilon_t + \alpha\sigma_{\mathrm{g}}^2/B$ in SC–SC and by $2\alpha C_\varepsilon \varepsilon_t + \alpha\sigma_{\mathrm{g}}^2/(2B)$ in NC–SC. For SC–SC, use $\|\nabla\Phi(x^t)\|^2 \ge 2\mu[\Phi(x^t) - \Phi^\star]$ and $\alpha\mu \le \omega$ to contract both terms in $\mathcal{V}_t^{\mathrm{g}}$. □

A.4.6 Local-y NC–SC substep analysis

The local-y NC–SC proof keeps the Q projected substeps instead of collapsing them into a single round update. This preserves the negative path length that is later used to absorb both stale-block drift and EF memory.

Lemma 5 (Primal envelope decrease for one local substep)**.** *Let $x^+ = x - \eta_x(\nabla_x F_\ell(x,y) + d_x)$ and $\Delta^y = \Phi(x) - F_\ell(x,y)$. If $\eta_x L_\Phi \le 1/8$, then*

$$\Phi(x^+) - \Phi^\star \le \Phi(x) - \Phi^\star - \frac{\eta_x}{8}\|\nabla\Phi(x)\|^2 - \frac{\|x^+ - x\|^2}{8\eta_x} + C\frac{\eta_x L^2}{\mu}\Delta^y + C\eta_x\|d_x\|^2. \tag{48}$$

Proof. Set $a = \nabla_x F_\ell(x,y) - \nabla\Phi(x)$ and $s = x^+ - x = -\eta_x(\nabla\Phi(x) + a + d_x)$. The standard smooth-descent inequality, obtained by averaging its displacement and gradient forms, gives

$$\Phi(x^+) \le \Phi(x) - \frac{\eta_x}{8}\|\nabla\Phi(x)\|^2 - \frac{\|s\|^2}{8\eta_x} + 2\eta_x\|a + d_x\|^2.$$

Use $\|a\|^2 \le 2L^2\Delta^y/\mu$ from Equation (27) and Young's inequality to conclude. □

Lemma 6 (Dual-gap decrease and transport)**.** *Let*

$$y^+ = \Pi_{\mathcal{Y}_\ell}\big(y + \eta_y(\nabla_y F_\ell(x,y) + d_y)\big)$$

and let x^+ be arbitrary. If $\eta_y L \le 1/4$, then

$$\begin{aligned}\Phi(x^+) - F_\ell(x^+,y^+) \le \left(1 - \frac{\mu\eta_y}{8}\right)[\Phi(x) - F_\ell(x,y)] - \frac{\|y^+ - y\|^2}{8\eta_y} \\ + C\frac{L^2}{\mu^2\eta_y}\|x^+ - x\|^2 + C\eta_y\|d_y\|^2.\end{aligned} \tag{49}$$

Proof. Apply Lemma 1 to $F_\ell(x,\cdot)$, then transport the resulting gap from x to x^+. Smoothness and Equation (27), followed by Young's inequality, bound this transport by

$$\frac{\mu\eta_y}{16}[\Phi(x) - F_\ell(x,y^+)] + C\frac{L^2}{\mu^2\eta_y}\|x^+ - x\|^2.$$

Combining the two estimates and weakening $(1+\theta/16)(1-\theta/4)$ to $1-\theta/8$ proves Equation (49). □

Proposition 4 (Local-y NC–SC substep dissipation)**.** *Let*

$$\Delta_t^y := \Phi(x^t) - F_\ell(x^t,y^t), \qquad \lambda := c_\lambda\frac{\alpha_x L^2}{\mu^2\alpha_y}, \qquad \mathcal{W}_t := \Phi(x^t) - \Phi^\star + \lambda\Delta_t^y,$$

where $c_\lambda > 0$ is a sufficiently large universal constant. There are universal constants $c_0, C > 0$ such that, whenever

$$\alpha_y L \le c_0, \qquad 1 \le \lambda \le 2, \tag{50}$$

the exact local round satisfies

$$\mathbb{E}[\mathcal{W}_{t+1} \mid \mathcal{F}_t] \le \mathcal{W}_t - \frac{\alpha_x}{C}\|\nabla\Phi(x^t)\|^2 - \frac{1}{C}\mathbb{E}[\mathcal{P}_t \mid \mathcal{F}_t] + C\alpha_x\kappa^2\frac{\sigma_\ell^2}{B}, \tag{51}$$

where

$$\mathcal{P}_t := \sum_{q=0}^{Q-1}\left(\frac{\|x^{t,q+1}-x^{t,q}\|^2}{\eta_x} + \lambda\frac{\|y^{t,q+1}-y^{t,q}\|^2}{\eta_y}\right).$$

Under EF communication, the same estimate holds with an additional $C\alpha_x\kappa^2\Lambda_\omega^2 D_t$ on the right, and

$$\mathbb{E}[D_{t+1} \mid \mathcal{F}_t] \le \left(1-\frac{\omega}{4}\right)D_t + 2\bar{\beta}_\omega\alpha_y\mathbb{E}[\mathcal{P}_t \mid \mathcal{F}_t]. \tag{52}$$

Consequently, if additionally

$$\alpha_y \le c_0\frac{\omega}{\Lambda_\omega}, \qquad \tau := \frac{8C\alpha_x\kappa^2\Lambda_\omega^2}{\omega}, \tag{53}$$

then $\mathcal{V}_t := \mathcal{W}_t + \tau D_t$ obeys

$$\mathbb{E}[\mathcal{V}_{t+1} \mid \mathcal{F}_t] \le \mathcal{V}_t - \frac{\alpha_x}{C}\|\nabla\Phi(x^t)\|^2 + C\alpha_x\kappa^2\frac{\sigma_\ell^2}{B}. \tag{54}$$

The first condition in Equation (53) is vacuous if $\Lambda_\omega = 0$.

Proof. Fix round t and abbreviate $z_q = z^{t,q} = (x_q, y_q)$. The same mini-batch is used throughout the round. Let $\xi^t = g^t - \overline{g}^t$ be the centered round-start sample error. Uniform Lipschitzness of the sample, stale-block, and population maps gives

$$g^{t,q} = \nabla F(z_q) + \xi^t + b_q, \qquad \|b_q\| \le 2L\|z_q - z_0\| + \Lambda_\omega\sqrt{D_t}. \tag{55}$$

The communication-state term is absent under exact communication. Let $d_q = \xi^t + b_q$, $\Delta_q^y = \Phi(x_q) - F_\ell(x_q, y_q)$, and $\mathcal{W}_q = \Phi(x_q) - \Phi^\star + \lambda\Delta_q^y$. Since $\lambda = c_\lambda\eta_x L^2/(\mu^2\eta_y)$, a sufficiently large universal c_λ absorbs the dual-gap and primal-displacement cross terms when the two preceding substep lemmas are added; the stated step restrictions ensure that both lemmas apply. Thus

$$\begin{aligned}\mathcal{W}_{q+1} \le \mathcal{W}_q &- \frac{\eta_x}{C}\|\nabla\Phi(x_q)\|^2 \\ &- \frac{1}{C}\left(\frac{\|x_{q+1}-x_q\|^2}{\eta_x} + \lambda\frac{\|y_{q+1}-y_q\|^2}{\eta_y}\right) + C(\eta_x+\lambda\eta_y)\|\xi^t + b_q\|^2.\end{aligned} \tag{56}$$

Use the standard discrete path bound

$$\sum_{q<Q}\|z_q - z_0\|^2 \le Q^2\sum_{q<Q}\|z_{q+1} - z_q\|^2$$

in Equation (55). Since $Q(\eta_x + \lambda\eta_y) \le C\alpha_x\kappa^2$, its stale-path contribution is $C(\alpha_y L)^2\mathcal{P}_t$, which is absorbed for small c_0. The equally standard gradient-transfer bound is

$$\sum_{q<Q}\eta_x\|\nabla\Phi(x_q)\|^2 \ge \frac{\alpha_x}{2}\|\nabla\Phi(x_0)\|^2 - C(\alpha_x L_\Phi)^2\mathcal{P}_t;$$

its path term is absorbed under the same restriction. Summing Equation (56) and taking conditional expectation gives Equation (51). Under EF, the sole additional term is $C\alpha_x\kappa^2\Lambda_\omega^2 D_t$.

By Cauchy–Schwarz and the definition of $\mathcal{P}_t$, $\|z^{t+1} - z^t\|^2 \le 2\alpha_y\mathcal{P}_t$. The memory recursion therefore gives Equation (52). Multiply it by τ and add the EF version of the exact recursion. The choice of τ absorbs the positive state-bias term, while the added path coefficient is $O(\bar{\beta}_\omega\Lambda_\omega^2\alpha_y^2/\omega)$; it is absorbed by $\alpha_y \le c_0\omega/\Lambda_\omega$ and Equation (37). Dropping the residual nonpositive terms proves Equation (54). □

The displayed recursions are the complete model-specific inputs used by the finite-time theorems below.

A.5 SC–SC PRIMAL–DUAL-GAP ANALYSIS

This subsection gives the complete SC–SC argument used by both exact and EF methods. It has two parts. Global-y first contracts the primal envelope and then converts the final server-side maximization certificate to the primal–dual gap. Local-y uses a gap-compatible merit at every projected substep; this is what permits arbitrary Q without introducing projected-mapping regularity.

A.5.1 A GAP-COMPATIBLE LOCAL MERIT

Let $z^\star = (x^\star, y^\star)$ denote the unique constrained saddle point and define

$$G_\star(z) := F(x, y^\star) - F(x^\star, y), \qquad W^{\mathrm{sc}}(z) := G_\star(z) + \frac{64L^2}{\mu}\|z - z^\star\|^2. \tag{57}$$

Lemma 7 (Comparison with the primal–dual gap)**.** *Under SC–SC,*

$$G_\star(z) \le \mathrm{Gap}(z) \le W^{\mathrm{sc}}(z) \le (1 + 128\kappa^2)\mathrm{Gap}(z). \tag{58}$$

Proof. The two saddle optimality conditions give $G_\star(z) \le \mathrm{Gap}(z)$ and $G_\star(z) \ge (\mu/2)\|z - z^\star\|^2$. The standard perturbed-optimality bound for a strongly convex–concave saddle problem gives

$$\mathrm{Gap}(z) - G_\star(z) \le \frac{L^2}{2\mu}\|z - z^\star\|^2.$$

The middle inequality in Equation (58) follows immediately. Finally, substitute $\|z - z^\star\|^2 \le 2\mathrm{Gap}(z)/\mu$ into Equation (57). □

Lemma 8 (One projected SC–SC substep)**.** *Let $\mathcal{Z} = \mathbb{R}^{d_x} \times \mathcal{Y}$ and*

$$z^+ = \Pi_{\mathcal{Z}}(z - \eta(\mathcal{T}_F(z) + e)), \tag{59}$$

where projection acts only on the y block. If $\eta \le \mu/(4L^2)$, then

$$W^{\mathrm{sc}}(z^+) \le \left(1 - \frac{\mu\eta}{8}\right) W^{\mathrm{sc}}(z) - \frac{\|z^+ - z\|^2}{8\eta} + 328\eta\kappa^2\|e\|^2. \tag{60}$$

Proof. Let $s = z^+ - z$. Apply the standard inexact smooth descent inequality to $F(\cdot, y^\star)$ and Lemma 1 to $F(x^\star, \cdot)$. The respective perturbations are bounded by $L\|y - y^\star\| + \|e_x\|$ and $L\|x - x^\star\| + \|e_y\|$. Adding the two inequalities, using Young's inequality, gives

$$G_\star(z^+) \le \left(1 - \frac{\mu\eta}{4}\right) G_\star(z) - \frac{\|s\|^2}{8\eta} + 8\eta L^2\|z - z^\star\|^2 + 8\eta\|e\|^2. \tag{61}$$

For the distance term, use the saddle inclusion and nonexpansiveness of projection in the usual inexact projected forward-step estimate. Strong monotonicity, Lipschitzness, and $\eta \le \mu/(4L^2)$ yield

$$\|z^+ - z^\star\|^2 \le (1 - \mu\eta)\|z - z^\star\|^2 + \frac{4\eta}{\mu}\|e\|^2. \tag{62}$$

Multiply Equation (62) by $64L^2/\mu$ and add Equation (61). The distance decrease absorbs the coupling term and supplies the required $\mu\eta/8$ contraction. Finally,

$$8 + 256\kappa^2 \le 328\kappa^2 \qquad (\kappa \ge 1),$$

which proves Equation (60). □

A.5.2 ARBITRARY-Q LOCAL ROUNDS

For a local round let $z_q = z^{t,q}$, $\eta = \alpha/Q$, and define the path energy

$$\mathcal{P}_t^{\mathrm{sc}} := \sum_{q=0}^{Q-1} \frac{\|z_{q+1} - z_q\|^2}{\eta}. \tag{63}$$

Proposition 5 (Exact local SC–SC round recursion for arbitrary Q)**.** *There are universal constants $c_0, c_1, C > 0$ such that for every $Q \geq 1$, if $\alpha_x = \alpha_y = \alpha$, $\eta = \alpha/Q$, and*

$$\alpha \leq c_0 \frac{\mu}{L^2}, \tag{64}$$

then FedBCDA satisfies

$$\mathbb{E}[W^{\mathrm{sc}}(z^{t+1}) \mid \mathcal{F}_t] \leq (1 - c_1\mu\alpha)W^{\mathrm{sc}}(z^t) - c_1\mathbb{E}[\mathcal{P}_t^{\mathrm{sc}} \mid \mathcal{F}_t] + C\alpha\kappa^2\frac{\sigma_\ell^2}{B}. \tag{65}$$

One admissible explicit choice is $c_0 = 1/256$ after enlarging C.

Proof. At substep q, write the reused stochastic direction as $g^{t,q} = \nabla F(z_q) + \xi^t + b_q$, where $\|b_q\| \leq 2L\|z_q - z_0\|$. Negate the dual components of the direction and its error to apply Lemma 8; the error norm remains $\|\xi^t + b_q\|$. Sum the resulting geometric recursion. The negative increment terms are bounded below by a fixed multiple of $\mathcal{P}_t^{\mathrm{sc}}$, while $r^Q \leq 1 - \mu\alpha/16$ for $r = 1 - \mu\eta/8$.

It remains only to absorb the stale-gradient term. The standard discrete path bound

$$\sum_{q<Q} \|z_q - z_0\|^2 \leq Q^2 \sum_{q<Q} \|z_{q+1} - z_q\|^2$$

and $\|\xi^t + b_q\|^2 \leq 2\|\xi^t\|^2 + 8L^2\|z_q - z_0\|^2$ show that its coefficient relative to $\mathcal{P}_t^{\mathrm{sc}}$ is $\eta^2\kappa^2L^2Q^2 = (\alpha L^2/\mu)^2$. It is absorbed under Equation (64); the remaining term is $C\alpha\kappa^2\|\xi^t\|^2$. Conditional expectation proves Equation (65). No substep independence is used. □

Proposition 6 (EF local SC–SC round recursion for arbitrary Q)**.** *Under the same equal effective steps, suppose*

$$\alpha \leq c_0 \frac{\mu}{L^2}, \qquad \mu\alpha \leq c_0\omega, \qquad \frac{\bar{\beta}_\omega\alpha^2\kappa^2\Lambda_\omega^2}{\omega} \leq c_0. \tag{66}$$

Then, for a universal $\tau \asymp \alpha\kappa^2\Lambda_\omega^2/\omega$,

$$\mathcal{V}_t^{\mathrm{sc}} := W^{\mathrm{sc}}(z^t) + \tau D_t \tag{67}$$

satisfies

$$\mathbb{E}[\mathcal{V}_{t+1}^{\mathrm{sc}} \mid \mathcal{F}_t] \leq (1 - c\mu\alpha)\mathcal{V}_t^{\mathrm{sc}} + C\alpha\kappa^2\frac{\sigma_\ell^2}{B}. \tag{68}$$

Proof. Now $\|b_q\| \leq 2L\|z_q - z_0\| + \Lambda_\omega\sqrt{D_t}$. Repeating the preceding path argument, with $\|\xi^t + b_q\|^2 \leq 2\|\xi^t\|^2 + 16L^2\|z_q - z_0\|^2 + 4\Lambda_\omega^2 D_t$, gives

$$\mathbb{E}[W^{\mathrm{sc}}(z^{t+1}) \mid \mathcal{F}_t] \leq (1 - c\mu\alpha)W^{\mathrm{sc}}(z^t) - c\mathbb{E}[\mathcal{P}_t^{\mathrm{sc}} \mid \mathcal{F}_t] + C\alpha\kappa^2\Lambda_\omega^2 D_t + C\alpha\kappa^2\frac{\sigma_\ell^2}{B}. \tag{69}$$

Here the Q^2 in the path bound again cancels with $\eta^2 = (\alpha/Q)^2$. Moreover, Lemma 4 and Cauchy–Schwarz give

$$\mathbb{E}[D_{t+1} \mid \mathcal{F}_t] \leq \left(1 - \frac{\omega}{4}\right) D_t + \bar{\beta}_\omega\alpha\mathbb{E}[\mathcal{P}_t^{\mathrm{sc}} \mid \mathcal{F}_t].$$

Choose $\tau = C_\tau\alpha\kappa^2\Lambda_\omega^2/\omega$. The first condition in Equation (66) absorbs the stale-path term, the third absorbs the extra path term created by the memory recursion, and the second ensures that the retained EF contraction is at least as fast as the target $c\mu\alpha$ contraction. Adding the two recursions and dropping the nonpositive residuals proves Equation (68). □

A.5.3 Global-y terminal primal–dual-gap conversion

Define

$$\rho_{\mathrm{ex}} := 2\kappa(1 + \kappa). \tag{70}$$

Lemma 9 (Exact global terminal conversion)**.** *If $F(x, y^\star(x)) - F(x, y) \leq \varepsilon$, then*

$$\mathrm{Gap}(x, y) \leq \rho_{\mathrm{ex}}[\Phi(x) - \Phi^\star] + (1 + 2\kappa^2)\varepsilon. \tag{71}$$

Proof. Strong concavity gives $\|y-y^\star(x)\|^2 \le 2\varepsilon/\mu$. Combine the standard strong-convexity bound $F(x,y)-\Psi(y) \le \|\nabla_x F(x,y)\|^2/(2\mu)$ with

$$\|\nabla_x F(x,y)\|^2 \le 2\|\nabla\Phi(x)\|^2 + 4L^2\varepsilon/\mu, \qquad \|\nabla\Phi(x)\|^2 \le 2L_\Phi[\Phi(x)-\Phi^\star].$$

Adding the certified inner gap proves Equation (71). □

Lemma 10 (EF global terminal conversion)**.** *Let $\widehat{y}_\star$ maximize the reachable surrogate $\widehat{F}$ and suppose $\widehat{F}(\widehat{y}_\star) - \widehat{F}(y) \le \varepsilon$. If*

$$\sup_{v\in\mathcal{Y}_g} \|\nabla_y \widehat{F}(v) - \nabla_y F(x,v)\| \le \delta,$$

then

$$\mathrm{Gap}(x,y) \le \rho_{\mathrm{ex}}[\Phi(x)-\Phi^\star] + \left(2\kappa^2 + \frac{3}{2}\right)\frac{\delta^2}{\mu} + (4\kappa^2+2)\varepsilon. \tag{72}$$

Proof. Let $y^\star = y^\star(x)$. Comparing the true and surrogate optimality conditions gives $\|\widehat{y}_\star - y^\star\| \le \delta/\mu$; surrogate quadratic growth gives $\|y-\widehat{y}_\star\|^2 \le 2\varepsilon/\mu$. Thus

$$\|y-y^\star\|^2 \le \frac{2\delta^2}{\mu^2} + \frac{4\varepsilon}{\mu}. \tag{73}$$

The same comparison, the surrogate certificate, and Young's inequality yield

$$\Phi(x) - F(x,y) \le \frac{3\delta^2}{2\mu} + 2\varepsilon. \tag{74}$$

Finally use $F(x,y)-\Psi(y) \le \|\nabla\Phi(x)\|^2/\mu + L^2\|y-y^\star\|^2/\mu$, the envelope smoothness bound, and the two displays above to obtain Equation (72). □

A.5.4 Proofs of the SC–SC finite-time theorems

Proof of Theorem 3. Geometrically unroll Equation (41), then apply Lemma 9 to the final server-only certificate; this gives Equation (12). For local-y, unroll Equation (65) after dropping its path term, and use the two sides of Equation (58) to convert the terminal and initial merits. This gives Equation (13). □

Proof of Theorem 5. Unrolling Equation (46) gives Equation (16). At the terminal round, Lemma 3 gives $\delta^2 \le \Lambda_\omega^2 D_T$, while $D_T \le \omega\mathcal{V}_T^{\mathrm{g}}/(24\alpha_x C_D)$; substitution in Lemma 10 yields Equation (17). For local-y, exact initialization gives $\mathcal{V}_0^{\mathrm{sc}} = W^{\mathrm{sc}}(z^0)$; unroll Equation (68) and use $\mathrm{Gap}(z^T) \le W^{\mathrm{sc}}(z^T) \le \mathcal{V}_T^{\mathrm{sc}}$ together with Equation (58) to obtain Equation (19). □

Corollary 1 (Headline SC–SC communication orders)**.** *Ignoring universal numerical constants and choosing batch sizes and inner tolerances so that the additive terms in Theorems 3 and 5 are at most $O(\epsilon)$ gives*

$$\begin{aligned}
\textit{FedBCD-Max:}\quad & T = O\left(\kappa(1+\kappa)\log\frac{\kappa(1+\kappa)\Delta_\Phi}{\epsilon}\right),\\
\textit{FedBCDA:}\quad & T = O\left(\kappa^2\log\frac{\kappa^2\mathrm{Gap}(z^0)}{\epsilon}\right),\\
\textit{EF-FedBCD-Max:}\quad & T = O\left(\frac{\kappa(1+\kappa)}{\omega^2}\log\frac{\kappa(1+\kappa)\Delta_\Phi}{\omega^2\epsilon}\right),\\
\textit{EF-FedBCDA:}\quad & T = O\left(\frac{\kappa^2}{\omega^2}\log\frac{\kappa^2\mathrm{Gap}(z^0)}{\epsilon}\right).
\end{aligned}$$

All four orders hold for arbitrary $Q \ge 1$ when the corresponding per-substep step is the effective step divided by Q.

Proof. For exact local SC–SC, choose $\alpha = \Theta(\mu/L^2)$ in Equation (64); then $(\mu\alpha)^{-1} = \Theta(\kappa^2)$. For EF local SC–SC, Appendix A.7 shows that the sharper choice $\alpha \asymp \omega^2/(L\kappa)$ already satisfies all three conditions in Equation (66); hence $(\mu\alpha)^{-1} = O(\kappa^2/\omega^2)$. The global EF choice $\alpha_x \asymp \omega^2/[L(1+\kappa)]$ gives inverse contraction scale $O(\kappa(1+\kappa)/\omega^2)$. Substitution into the geometric factors and terminal metric conversions proves the displayed orders; universal numerical constants are absorbed into $O(\cdot)$. □

A.6 Proofs of the NC–SC finite-time theorems

A.6.1 Exact communication

Proof of Theorem 4. For global-y, telescope Equation (42) after subtracting $\Phi^\star$, use $\Phi(x^T) - \Phi^\star \geq 0$, and divide by $7\alpha_x T/16$. The independent uniform choice of R gives Equation (14). For local-y, telescope Equation (51), drop the nonnegative terminal potential $\mathcal{W}_T$, and divide by $\alpha_x T/C$. The same uniform-index argument gives Equation (15). □

A.6.2 Identity-compressor recovery

Proposition 7 (Identity-compressor recovery)**.** *Suppose both the uplink and downlink compressors are the identity, the EF methods use the stated exact-memory initialization, and the exact/EF pairs use the same mini-batches, local-oracle randomness, and inner-solver choices. Then $D_t^{\mathrm{up}} = D_t^{\mathrm{down}} = D_t = 0$ whenever the corresponding memories are used. Consequently, EF-FedBCD-Max coincides round by round with FedBCD-Max, and EF-FedBCDA coincides round by round with FedBCDA.*

Proof. The stated initialization makes both memories exact. If they are exact at round t, identity uplink compression gives $\widehat{u}_i^{t+1} = u_i^{t+1}$; for $t < T-1$, identity downlink compression then gives $\widehat{m}_i^{t+1} = m_i^{t+1}$. Induction proves exactness throughout. By Equation (6), the reconstructed maps, inner points, and local updates therefore coincide with those of the coupled exact methods. □

A.6.3 Error feedback

Proof of Theorem 6. For global-y, telescope Equation (47), use $\mathcal{V}_T^{\mathrm{g}} \geq 0$, and divide by $\alpha_x T/16$ to obtain Equation (20). For local-y, the same argument applied to Equation (54), with $\mathcal{V}_T \geq 0$, gives Equation (21). □

A.7 Condition-number and compression specialization

Throughout, $\kappa = L/\mu \geq 1$ and $c > 0$ denotes a sufficiently small universal constant. Fixed communication-interface regularity is absorbed into L.

Proposition 8 (Feasible effective steps and headline orders)**.** *The effective-step choices in Table 1, with sufficiently small universal prefactors, satisfy all proof-level admissibility conditions. The local-y choices remain valid for every $Q \geq 1$ when $\eta_x = \alpha_x/Q$ and $\eta_y = \alpha_y/Q$. In particular, the EF local SC–SC branch permits*

$$\alpha_x = \alpha_y = \Theta\left(\frac{\omega^2}{L\kappa}\right),$$

so its compression dependence is ω^{-2}. Both local NC–SC branches have stochastic stationarity floor $O(\kappa^2\sigma_\ell^2/B)$.

Proof. From the definitions in Equation (11) and Equation (37), for $0 < \omega \leq 1$,

$$\beta_\omega \leq \frac{2}{\omega}, \qquad \theta_\omega^{-1} = 1 + \frac{24\beta_\omega}{\omega} \leq \frac{49}{\omega^2}, \qquad \Lambda_\omega \leq \frac{CL}{\omega}, \qquad \bar{\beta}_\omega \leq \frac{5}{2\omega}.$$

For exact communication take

$$\alpha_x = \frac{c}{L(1+\kappa)}, \qquad \alpha = \frac{c\mu}{L^2} = \frac{c}{L\kappa}, \qquad \alpha_y = \frac{c}{L}, \quad \alpha_x = \frac{\alpha_y}{c_\lambda \kappa^2}$$

in the global-y, local SC–SC, and local NC–SC cases, respectively. Then $\alpha_x L_\Phi = O(c)$, $\alpha L^2/\mu = c$, and the local NC–SC weight λ in Proposition 4 is a universal constant in $[1, 2]$ after fixing c_λ and the prefactors. Hence the exact restrictions hold for small enough c, independently of Q.

For global-y EF use

$$\alpha_x = \frac{c\omega^2}{L(1+\kappa)}.$$

Since $C_D = 2\Lambda_\omega^2(1+\kappa)^2 = O(L^2(1+\kappa)^2/\omega^2)$,

$$\frac{\bar{\beta}_\omega \alpha_x^2 C_D}{\omega} = O(c^2),$$

while $\alpha_x L_\Phi = O(c\omega^2)$ and $\alpha_x \mu \le c\omega$; thus Equation (45) holds after decreasing c.

For local SC–SC EF, take the sharper choice

$$\alpha = \frac{c\omega^2}{L\kappa} = \frac{c\omega^2\mu}{L^2}. \tag{75}$$

The first two conditions in Equation (66) follow immediately from $\omega \le 1$ and $\kappa \ge 1$:

$$\alpha \le c\frac{\mu}{L^2}, \qquad \mu\alpha = \frac{c\omega^2}{\kappa^2} \le c\omega.$$

For the mixed EF-memory condition,

$$\frac{\bar{\beta}_\omega \alpha^2 \kappa^2 \Lambda_\omega^2}{\omega} \le C\left(\frac{1}{\omega}\right)\left(\frac{c^2\omega^4}{L^2\kappa^2}\right)\kappa^2\left(\frac{L^2}{\omega^2}\right)\left(\frac{1}{\omega}\right) = O(c^2).$$

Thus the mixed memory restriction is satisfied with an effective step of order $\omega^2/(L\kappa)$.

Finally, for local NC–SC EF use

$$\alpha_y = \frac{c\omega^2}{L}, \qquad \alpha_x = \frac{\alpha_y}{c_\lambda \kappa^2}.$$

Because $\omega/\Lambda_\omega = \Omega(\omega^2/L)$, the EF restriction Equation (53) holds, and the same choice of relative prefactors keeps $\lambda \in [1, 2]$.

All admissibility restrictions therefore hold for sufficiently small universal c and arbitrary $Q \ge 1$. Inverting the effective steps and substituting into Theorems 3–6 gives the eight orders in Table 1; in particular, Equation (75) yields $(\mu\alpha)^{-1} = \Theta(\kappa^2\omega^{-2})$ for EF-FedBCDA in SC–SC. □

A.7.1 Comparison with the underlying centralized algorithms

The relevant reference algorithms are determined by the update schedule. With $Q = 1$, exact communication, and deterministic gradients, FedBCDA is projected simultaneous GDA on the stacked blocks. FedBCD-Max with exact inner maximization is gradient descent on Φ. The following examples show that their SC–SC κ^2 dependence cannot in general be improved for these fixed-step specializations. They do not give lower bounds for all Q, adaptive steps, or other minimax algorithms.

Simultaneous GDA. Let $\kappa \ge 2$, $a = \sqrt{L^2 - \mu^2}$, and consider the unconstrained scalar game

$$F(x, y) = \frac{\mu}{2}x^2 + axy - \frac{\mu}{2}y^2.$$

Its joint gradient is L-Lipschitz, both strong-curvature constants are μ, and

$$\mathrm{Gap}(x, y) = \frac{L^2}{2\mu}(x^2 + y^2).$$

One simultaneous GDA step with a common stepsize η multiplies this gap by

$$r(\eta) = (1 - \eta\mu)^2 + \eta^2 a^2, \qquad \min_{\eta > 0} r(\eta) = 1 - \kappa^{-2}.$$

Thus even the best common fixed step requires $\Omega(\kappa^2 \log(1/\epsilon))$ iterations for a relative gap reduction by ϵ. This agrees with the Sim-GDA conditioning discussed in Zhang et al. (2022, Theorem 4); their blockwise smoothness convention changes only universal constants. Our arbitrary-Q upper bound retains this conditioning order while allowing stale local trajectories.

Envelope gradient descent. For $x = (x_1, x_2)$, consider

$$F(x,y) = \frac{\mu}{2}(x_1^2 + x_2^2) + a x_2 y - \frac{\mu}{2} y^2, \qquad \Phi(x) = \frac{\mu}{2} x_1^2 + \frac{L^2}{2\mu} x_2^2.$$

Again F has joint smoothness L and strong-curvature constants μ, but the envelope has condition number κ^2. The worst-coordinate contraction of fixed-step gradient descent, measured in the envelope gap, is

$$\min_{\eta>0} \max\{(1-\eta\mu)^2, (1-\eta L^2/\mu)^2\} = \left(\frac{\kappa^2-1}{\kappa^2+1}\right)^2.$$

Moreover, $\mathrm{Gap}(x, y^\star(x)) = \mu x_1^2/2 + L^4 x_2^2/(2\mu^3)$ has the same worst-coordinate contraction. Hence $\Omega(\kappa^2 \log(1/\epsilon))$ rounds are necessary in this specialization as well. The example explains why an exact inner maximizer alone does not improve the SC–SC order to κ: eliminating y can square the conditioning seen by primal gradient descent.

NC–SC and initialization. Write δ for a gradient-norm tolerance. Theorem 4.4 of Lin et al. (2020a) gives $O((L\kappa^2\Delta_\Phi + \kappa L^2 D_{\mathcal{Y}}^2)/\delta^2)$ iterations for deterministic two-time-scale GDA on a bounded dual set of diameter $D_{\mathcal{Y}}$. Our criterion uses the squared norm, so $\delta^2 = \epsilon$. The steps $\alpha_y = \Theta(L^{-1})$, $\alpha_x = \Theta((L\kappa^2)^{-1})$ and the leading $L\kappa^2\Delta_\Phi/\epsilon$ term agree. Our bound instead retains

$$\mathcal{W}_0 = \Delta_\Phi + \lambda[\Phi(x^0) - F_\ell(x^0, y^0)], \qquad 1 \le \lambda \le 2,$$

and does not require a bounded dual set. The two initialization terms are not interchangeable: an uncontrolled initial dual gap can itself depend on κ. With $y^0 = y^\star(x^0)$, $\mathcal{W}_0 = \Delta_\Phi$; the cost of obtaining such an initialization must be counted if it requires an inner solve. This is a comparison of upper bounds, not an NC–SC optimality claim. The global-y NC–SC bound is the standard $O(L_\Phi\Delta_\Phi/\epsilon)$ outer-gradient bound; computing its certified maximizer can require additional inner work.

A.7.2 Local computation, compression, and communication accounting

For a fixed admissible ratio of primal and dual steps, the transient bounds scale as $1/\alpha_x$ in NC–SC and $1/(\mu\alpha_x)$ in SC–SC, up to logarithms. Thus increasing Q at fixed per-substep steps gives a proportional improvement in the round upper bound only while all effective-step restrictions hold and the same target residual budget is maintained. At fixed effective steps, the bounds are uniform in Q and do not guarantee further speedup. This trades additional local oracle evaluations for fewer synchronizations; it is not a claim of reduced total gradient work.

For fixed $\omega > 0$, exact tracker initialization, noiseless oracles, and exact inner solves, the EF finite-time bounds converge to zero without a separate compression floor. In the stochastic setting, B and the inner certificates must control the displayed residuals. Let b_t^{up} and b_t^{down} be the actual encoded payloads (including sparse indices and headers), and let b_t^{aux} include any other required transmissions, such as a shared global dual variable. Total traffic is

$$\mathrm{Bytes}(T) = b_{\mathrm{init}} + \sum_{t=0}^{T-1} b_t^{\mathrm{up}} + \sum_{t=1}^{T-1} b_t^{\mathrm{down}} + \sum_{t=0}^{T} b_t^{\mathrm{aux}}.$$

If b_ω and b_1 denote comparable per-round payloads, inserting the headline round bounds yields a sufficient-budget scaling proportional to $b_\omega \omega^{-2}$ versus b_1, apart from initialization and logarithms. Contractivity alone gives no inequality between these byte budgets. For example, Top-k in dimension d admits $\omega = k/d$ but transmits k values and their indices, so its worst-case round bound does not by itself prove a total-byte gain. Such gains must be evaluated using the realized payload and rounds to the same target accuracy.

B DATASETS

Fair vertical classification. Following the FairVFL protocol, we use Adult, COMPAS, and Crime. Their training/test splits contain $40{,}000/5{,}222$, $3{,}694/1{,}584$, and $1{,}395/599$ examples, respectively, after preprocessing; each slash separates the number of training examples from the number of test examples. The corresponding processed input dimensions are 104, 15, and 123 features per example.

Vertical adversarial learning. We use CIFAR-10, which contains 50,000 training and 10,000 test images.

C HYPERPARAMETER TUNING

How do we tune hyperparameters? We use the same staged principle for the two placements of the maximization variable in our unified minimax VFL framework. In the y-global setting, the server holds one shared maximizer and the clients update only their minimization blocks; in the y-clients setting, each client maintains a coupled minimax block. For both settings, we first tune the uncompressed base optimizer, rerun promising configurations over a longer horizon, and only then freeze the optimizer while tuning representation compression. This separation prevents a favorable compressor setting from masking a poorly tuned base method.

The two settings share three evaluation rules. First, we use unsmoothed full-dataset measurements whenever available: full-training loss, surrogate disparity $|D(x)|$ (DEO), its violation $(|D(x)|-\epsilon)_+$, and discrete DFP/DFN for fairness, and clean validation accuracy for adversarial learning. Second, one mini-batch is one outer communication round. Each client performs Q local block updates from the round-start synchronized state before the next communication, so increasing Q trades additional local computation for fewer synchronizations but does not multiply the outer-round count. Third, comparisons state whether the budget is fixed in epochs, communication rounds, local updates, or transmitted MB. Baselines retain their original update rules and are not assigned the multi-step, damping, proximal, or error-feedback mechanisms introduced for our FedBCD variants.

C.1 y-GLOBAL: FAIR VERTICAL CLASSIFICATION

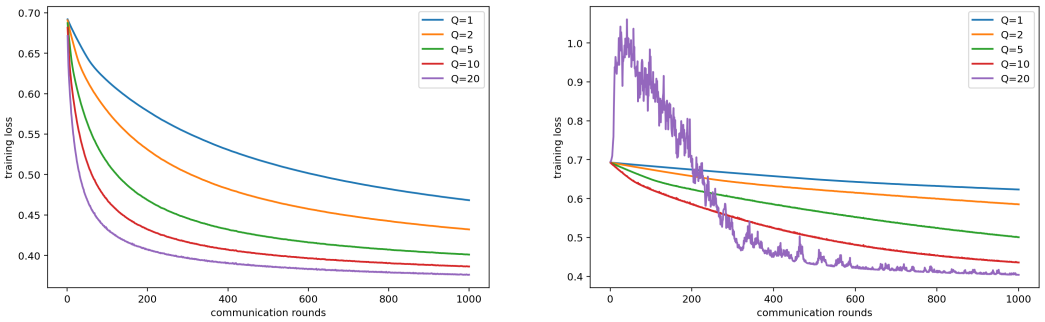
Base-optimizer search. We tune FedBCD-Max on the linear and one-hidden-layer MLP tasks for Adult, COMPAS, and Crime. Adult, COMPAS, and Crime contain $40{,}000$, $3{,}694$, and $1{,}395$ training examples, respectively; the MLP hidden width is 32. The search varies one factor at a time while holding the remaining parameters fixed. Table 7 records the complete ranges. Main experiments retain $\alpha = 10^{-3}$ and batch size 512 because this pair provides consistent progress across all six dataset-model combinations. Smaller learning rates make little progress within 1,000 rounds, whereas 10^{-1} amplifies stale-block and dual-correction oscillations. Batches 128 and 256 give noisy group-conditioned estimates on the smaller datasets, while full-batch training no longer reflects the intended mini-batch regime.

Table 7: y-global search space for FedBCD-Max. Bold entries are used in the main experiment or the corresponding compression study.

Hyperparameter	Search space	Selected use
Penalty threshold ϵ	$0.01, \mathbf{0.05}, 0.1$	Report all; 0.05 for EF
Dual regularization c	$0.001, 0.01, \mathbf{0.05}, \mathbf{0.1}$	Main: 0.05; EF: 0.10
Local steps Q	$1, 2, \mathbf{3}, \mathbf{5}, 7, 10, 20, 30$	Main: 5; EF: 3
Learning rate α	$10^{-5}, 10^{-4}, \mathbf{10^{-3}}, \mathbf{10^{-2}}, 10^{-1}$	Main: 10^{-3}; EF: 10^{-2}
Batch size	$128, 256, \mathbf{512}, 1024$, full	Main and EF
EF Top-K retention	$\{0.1, 1, 5, 10, 15, 20, 25, \mathbf{30}, 35, 40\}\%$	Screen: 20%; final: 30%
EF learning rate	$0.001, 0.0025, 0.003, 0.005, 0.0075, \mathbf{0.01}$	EF: 0.01

Local steps and stability. We compare $Q \in \{1, 2, 3, 5, 7, 10\}$ for 1,000 fixed communication rounds with $\epsilon = 0.05$, $c = 0.05$, $\alpha = 0.01$, and batch size 512, and extend the diagnostic runs to $Q \in \{20, 30\}$. Figure 6 shows the representative Crime results. A larger Q monotonically accelerates the linear model on the communication-round axis, but the MLP becomes increasingly sensitive to stale-block drift; $Q = 20$ produces a large transient and $Q = 30$ is still less stable. The

differences also shrink when runs are compared at a fixed number of local updates. We therefore use $Q = 5$ as a cross-task compromise instead of selecting the largest value from the linear task.



(a) Crime linear (b) Crime MLP

Figure 6: y-global sensitivity to the number of local updates Q under 1,000 fixed communication rounds.

Penalty and stabilization parameters. For the regularized equality-of-opportunity objective, the optimized dual response has the form

$$\lambda_1^\star = \left[\frac{D(x) - \epsilon}{c}\right]_+, \qquad \lambda_2^\star = \left[\frac{-D(x) - \epsilon}{c}\right]_+.$$

Eliminating the dual variables yields $\mathcal{L}(x) + [(|D(x)| - \epsilon)_+]^2/(2c)$; hence finite c defines a soft penalty and does not guarantee $|D(x)| \le \epsilon$. We treat ϵ as the penalty threshold and distinguish $|D(x)|$, its violation $(|D(x)| - \epsilon)_+$, and discrete DFP/DFN. In the dense 2,000-round screen, $c = 0.001$ responds too aggressively, $c = 0.1$ too conservatively, and $c = 0.05$ gives the best loss–disparity balance. We also test $\lambda_{\max} \in \{5, 10, 20\}$, dual damping $\rho \in \{0.05, 0.1, 0.2\}$, proximal weights $\mu \in \{10^{-3}, 10^{-2}, 10^{-1}\}$, and square-root or linear learning-rate scaling with Q. None improves all settings, so the main configuration uses no finite dual cap, damping, proximal term, or Q-dependent learning-rate scaling.

Compression-specific tuning. The Adult compression study uses a stochastic round-level dual protocol that differs from the full-data dual evaluation used in the main fairness experiment. At the beginning of each communication round, the server computes the analytic dual response from the synchronized mini-batch, and this dual variable is held fixed throughout the subsequent Q local client updates. The same protocol is used for dense FedBCD-Max, direct bidirectional Top-K, and BiEF-FedBCD-Max, so that the compression comparison changes only the communication mechanism rather than the dual update rule. For the compression study, all FedBCD-Max variants use the same round-wise mini-batch analytic dual computed from the shared mini-batch and reused across the Q local updates. Thus, the dense, direct Top-K, and BiEF variants differ only in their communication mechanism.Preliminary sparsification studies consider both individual retention rates and independent uplink/downlink rates, from which we select a moderate common operating point $k_u = k_d = 30\%$ for the final controlled comparison between direct bidirectional Top-K and BiEF. We further refine the learning rate, local-step count, dual regularization, and initialization strategy. To make the no-EF versus BiEF comparison controlled, both compressed methods use the same dense round-0 initialization of the uplink surrogates and downlink memories; thereafter they use the same bidirectional Top-K operator, while only BiEF carries residual errors forward. The retained compression configuration is

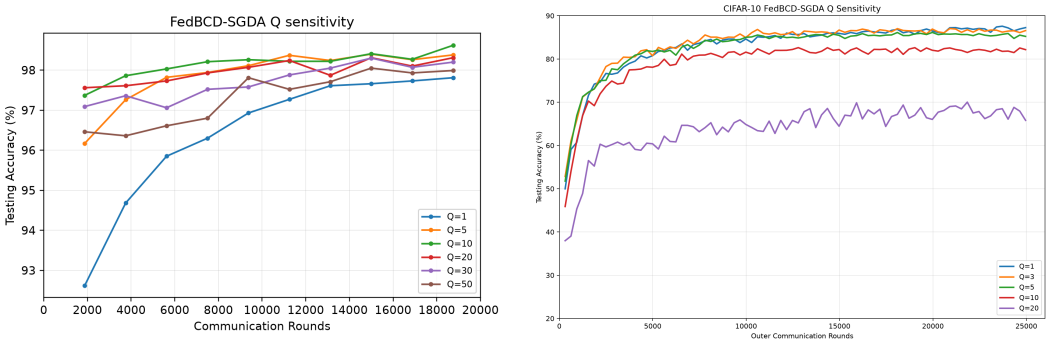
$$Q = 3, \quad \alpha = 0.01, \quad k_u = k_d = 30\%, \quad \epsilon = 0.05, \quad c = 0.10, \quad \text{BS} = 512.$$

The final comparison runs 60 data epochs, or 4,688 communication rounds, for each of seeds 0–3. The communication counter includes the common one-time dense initialization and the realized encoded uplink and downlink payloads, including sparse indices. Dense communication uses 230.43 MB, while both direct bidirectional $30\%/30\%$ Top-K and BiEF use 142.43 MB. Under this accounting, the compressed variants reduce recorded communication by 38.2% relative to dense communication while using the same communication budget. Averaged over the four seeds, BiEF

closely matches dense FedBCD-Max on the linear model and remains close on the MLP, whereas direct Top-K is substantially worse and more variable. These results indicate that error feedback can recover information repeatedly omitted by sparsification and achieve near-dense optimization performance under the same sparse bidirectional payload.

C.2 y-CLIENTS: VERTICAL ADVERSARIAL LEARNING

Sensitivity to the number of local updates. We first examine Q, which controls the amount of local computation performed between two outer communication rounds. On MNIST, we compare $Q \in \{1, 2, 5, 10, 20\}$ and extend the diagnostic range to $Q \in \{1, 5, 10, 20, 30, 50\}$. On CIFAR-10, we fix batch size 160 and compare $Q \in \{1, 3, 5, 10, 20\}$ for 80 epochs. Figure 7 collects both sensitivity studies. On MNIST, $Q = 5$–20 enter the high-accuracy regime quickly and $Q = 10$ gives the best displayed endpoint. On CIFAR-10, $Q = 1, 3, 5$ finish in a similar 85%–87% range, whereas $Q = 10$ is lower and $Q = 20$ is unstable. Together, the results show a stable region at moderate Q and degradation when too many local updates amplify stale-block error. We therefore retain $Q \in \{5, 10, 20\}$ in the MNIST grid and use $Q = 3$ for the final CIFAR-10 experiment.



(a) MNIST (b) CIFAR-10

Figure 7: y-clients accuracy sensitivity to the number of local updates Q on the outer communication-round axis.

MNIST grid search. We use MNIST to tune the remaining optimization parameters jointly. The client and server learning rates are selected independently from

$$\eta_c, \eta_s \in \{10^{-4}, 5 \times 10^{-4}, 10^{-3}\},$$

the batch-size candidates are $\{32, 64, 128\}$, and $Q \in \{5, 10, 20\}$. Each configuration is trained for 10 epochs and ranked by mean clean accuracy over the final three epochs, with clean loss as a tie-breaker. We then extend the 20 strongest configurations to 20 epochs and compare their complete accuracy and loss trajectories.

The grid-search results form a broad high-performing region rather than an isolated optimum: mean clean accuracy across the shortlisted configurations ranges from 98.32% to 98.64%, and the top ten configurations are within 0.10 percentage points of one another. The highest-ranked configurations also use different values of Q and batch size. We therefore treat small changes in ranking as negligible and select a configuration that combines the highest observed clean accuracy with a stable full training trajectory:

$$Q = 10, \qquad \mathrm{BS} = 64, \qquad \eta_c = 10^{-4}, \qquad \eta_s = 5 \times 10^{-4}.$$

We then rerun the selected configuration for 15 epochs across four independent runs on a fixed 5,000-example validation subset. Averaged over the final three epochs, it achieves 98.67% validation accuracy and 0.0582 validation loss.

Compression-specific tuning. With the MNIST base optimizer fixed, we compare direct Top-K and EF-TopK for

$$r \in \{10\%, 20\%, 25\%, 30\%, 35\%\}.$$

Early runs used the interim $Q = 5$ candidate, but every formal sweep is rerun with the selected $Q = 10$, batch size 64, and learning-rate pair $10^{-4}/5\times10^{-4}$. Retaining only 10%–25% strongly degrades

direct Top-K in the final implementation, whereas EF remains stable across the sweep. Direct Top-K recovers substantially at $r = 30\%$; increasing the retention to 35% improves it further but raises the recorded traffic from 531.8 to 620.6 MB without changing the qualitative EF conclusion. We therefore select $r = 30\%$ as the smallest tested retention that avoids the severe low-retention regime while preserving a meaningful communication reduction, and confirm this choice over four training seeds.

The final experiment uses the same four-run protocol and records full-training-set mean loss and validation accuracy every 100 mini-batches.

At epoch 15, the four-seed mean validation accuracies for dense, direct Top-K 30%, and EF-TopK 30% are 98.56%, 97.15%, and 98.81%; their mean training losses are 8.48×10^{-4}, 5.42×10^{-2}, and 4.47×10^{-4}. The final recorded traffic is 1205.3 MB for dense communication, 531.8 MB for direct Top-K, and 567.3 MB for EF-TopK. At the common 531.8 MB budget, their validation accuracies are 98.54%, 97.62%, and 98.79%, respectively.

C.3 Final configurations and reporting boundaries

Table 8 summarizes all retained main and bidirectional compression experiments. A different Q in a compression experiment is not a contradiction: the base solver and the compressor are tuned in separate stages, and compression can change the stability of repeated local updates.

Table 8: Final experimental configurations used after tuning.

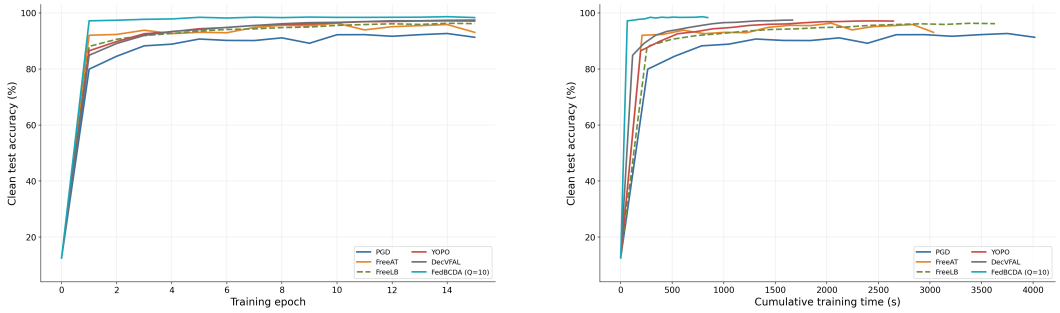
Setting	Experiment	C	Q	BS	Learning rate(s)	Other settings
y-global	Fairness main	6	5	512	$\alpha = 10^{-3}$	$c = 0.05$; $\epsilon \in \{0.01, 0.05, 0.1\}$
y-global	Adult EF	6	3	512	$\alpha = 10^{-2}$	$c = 0.10$, $\epsilon = 0.05$; Bi-Top-K/BiEF, $k_u = k_d = 30\%$
y-clients	MNIST main	2	10	64	$\eta_c = 10^{-4}$, $\eta_s = 5 \times 10^{-4}$	$\epsilon = 0.02$, $\sigma = 0.002$; exact
y-clients	MNIST EF	2	10	64	$\eta_c = 10^{-4}$, $\eta_s = 5 \times 10^{-4}$	$\epsilon = 0.02$, $\sigma = 0.002$; Bi-Top-K/BiEF, $k_u = k_d = 30\%$
y-clients	CIFAR-10 main	2	3	160	$\eta_c = 10^{-4}$, $\eta_s = 5 \times 10^{-4}$	$\epsilon = 8/255$, $\sigma = 1/255$; exact
y-clients	CIFAR-10 EF	2	3	160	$\eta_c = 10^{-4}$, $\eta_s = 5 \times 10^{-4}$	$\epsilon = 8/255$, $\sigma = 1/255$; Bi-Top-K/BiEF, $k_u = k_d = 20\%$

Communication is measured from the realized encoded uplink and downlink payloads using the same accounting rule for direct Top-K and BiEF. The MNIST coarse screen also used clean test accuracy because a validation split was not yet available and is therefore exploratory. Formal confirmation uses a fixed validation subset and multiple independent runs.

D More Experimental Results

This section provides additional MNIST results for framework effectiveness, communication efficiency, and adversarial robustness.

Framework effectiveness and convergence efficiency. Figure 8 reports clean accuracy against both training epoch and elapsed time. FedBCDA with $Q = 10$ reaches 97.18% after the first epoch and finishes at 98.34% after 15 epochs. The two axes respectively show optimization progress and practical wall-clock convergence against the same five baselines.



(a) Training epoch. (b) Elapsed time.

Figure 8: y-clients MNIST-MLP clean test accuracy.

Communication efficiency with EF. We compare dense FedBCDA, bidirectional Top-K, and BiEF-FedBCDA under the same training and communication settings. Figure 9 reports four-seed mean trajectories with one-standard-deviation shading. BiEF effectively recovers information lost through direct sparsification, closely matching dense convergence while achieving the best clean test accuracy under the shared communication budget.

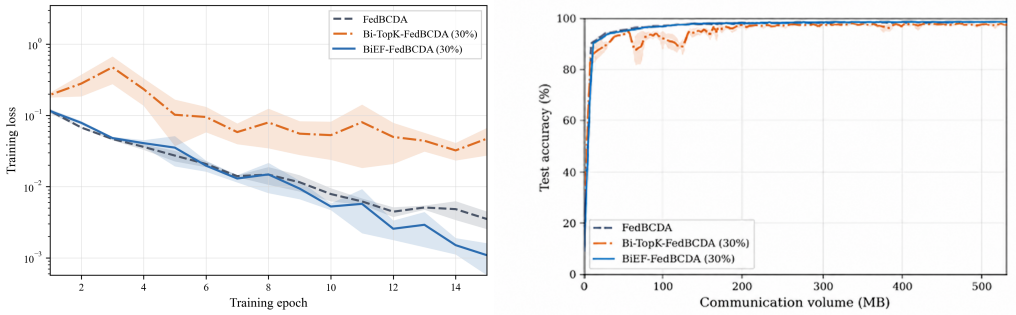


Figure 9: y-clients MNIST bidirectional compression with and without error feedback ($Q = 10$, $k_u = k_d = 30\%$, four seeds).

Robustness evaluation. We evaluate FGSM (Kurakin et al., 2017), PGD (Madry et al., 2018), CW (Carlini & Wagner, 2017), and zeroth-order variants based on ZOO (Chen et al., 2017). In all tests, one of the two clients is malicious, and only its feature partition is perturbed. FGSM, PGD, and CW have white-box access to the end-to-end model, whereas ZO-FGSM and ZO-PGD use loss queries. PGD and ZO-PGD use 10 steps, one random initialization, and no additional restarts. On MNIST/CIFAR-10, ZOO uses 100/200 directions per gradient estimate with $\mu = 0.05$, and CW uses at most 100 update steps.

CER follows the randomized-smoothing CERTIFY procedure of Cohen et al. (Cohen et al., 2019). Gaussian noise is applied only to the malicious client's feature block. For each example, we use $n_0 = 100$ samples to select the predicted class and $n = 100{,}000$ independent samples to compute a one-sided Clopper–Pearson lower bound with failure probability $\alpha = 0.001$. We report certified accuracy at a fixed ℓ_2 radius, with $\sigma = R = 128/255$ on MNIST and $\sigma = R = 64/255$ on CIFAR-10. An example is counted as correct only when its smoothed prediction is correct and its certified radius is at least R; abstentions count as incorrect.

Table 9: y-clients MNIST robustness (%).

Method	Clean	FGSM	PGD-10	CW	CER	ZO-FGSM	ZO-PGD-10	s/epoch
PGD	91.33	66.61	41.26	90.79	77.79	72.02	67.77	267.74
FreeAT	93.05	81.78	72.71	92.99	88.20	87.08	86.46	202.35
FreeLB	96.18	47.86	10.63	95.40	75.22	71.72	66.84	242.06
YOPO	97.09	93.44	90.19	97.07	95.57	94.94	94.20	176.33
DecVFAL	97.45	**94.52**	**91.82**	97.43	**96.02**	**95.16**	94.30	111.09
FedBCDA (**Ours**)	98.35	90.50	75.42	98.24	94.89	94.94	94.33	**56.30**
BiEF-FedBCDA (**Ours**)	**98.69**	91.53	77.04	**98.56**	94.50	94.63	**94.42**	57.67