Switching Gradient Methods for Constrained Federated Optimization

Antesh Upadhyay AANTESH@PURDUE.EDU
Sang Bin Moon MOON182@PURDUE.EDU
Abolfazl Hashemi ABOLFAZL@PURDUE.EDU
School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

Abstract

Constrained optimization problems arise in federated learning (FL) settings, where a global objective must be minimized subject to a functional constraint aggregated across clients. We introduce *Federated Switching Gradient Methods* (FEDSGM), a primal-only, projection-free algorithm for federated constrained optimization. By extending switching gradient methods to the federated setting, FEDSGM avoids the inner solves and penalty tuning required by dual or penalty-based methods, enabling lightweight and scalable deployment. Our analysis addresses three practical challenges simultaneously: (i) multi-step local updates to accommodate heterogeneous client compute, (ii) unbiased uplink compression to mitigate communication costs, and (iii) both hard and soft switching between objective and constraint gradients. We provide the first convergence guarantees for constrained FL that hold under these combined settings, recovering known centralized rates in special cases. In particular, we show that soft switching, recently proposed in the centralized literature, retains convergence guarantees while offering improved empirical stability near the constraint boundary.

1. Introduction

We study federated constrained optimization problems of the form

$$w^* = \arg\min_{w\in\mathbb{R}^d} \left\{ f(w) := \frac{1}{n}\sum_{j=1}^{n} f_j(w) \quad \text{s.t.} \quad g(w) := \frac{1}{n}\sum_{j=1}^{n} g_j(w) \le 0 \right\}, \tag{1}$$

where f_j and g_j represent the local objective and constraint functions on the client j. This framework captures many real-world applications of federated learning [18, 22, 24], where models must be trained across private or geo-distributed data without central collection. Constraints $g(w) \le 0$ encode feasibility conditions such as fairness mandates, energy budgets, or safety margins in autonomous systems and battery management [1, 5, 19, 38].

Still, solving (1) in realistic deployment scenarios remains challenging. We highlight three primary difficulties that simultaneously shape the design space for federated constrained optimization:

(i) Functional constraints. Federated tasks increasingly involve feasibility criteria beyond minimizing a loss: fairness across subpopulations [1], bounding risk exposure in financial models, or safety limits in batteries (e.g., maximum temperature rise). Enforcing such constraints requires algorithms that provide guarantees on feasibility without resorting to expensive projections or inner constrained solves each round.

(ii) Severe bandwidth limits. Deep neural network models involve millions of parameters, yet FL often operates on commodity wireless or edge networks, where it is infeasible to send full-precision parameter updates every round. Communication-efficient training requires compression techniques such as Top-K, Rand-K, or quantization [2, 9, 33].

(iii) Heterogeneous on-device compute. Devices participating in FL varies in orders of magnitude with their FLOPS, memory, and energy capacity. A common strategy is to allow clients to perform multiple local updates ($E > 1$) before each communication round [20], amortizing latency and increasing utilization. Yet, local updates cause *drifts* between client and global iterates, complicating convergence analysis, especially when constraints must be satisfied globally.

Despite the importance of these challenges for real-world FL applications, no existing method provides provable guarantees that addresses them simultaneously.

Limitations of existing approaches. Constrained versions of FEDAVG [14] and primal–dual or AL/ADMM methods [4, 6, 8, 13, 21, 25, 26, 39] can certify feasibility but rely on dual variable tuning, penalty scheduling, or inner projection steps. These methods also typically assume synchronous full participation and uncompressed updates, which is unrealistic in FL applications. Error-feedback compression algorithms, such as EF-SGD and SAFE-EF [2, 9, 17, 33], provide robustness against biased quantization, but do not incorporate local update drift ($E > 1$). Local-SGD methods [20] directly address compute heterogeneity but are unconstrained, offering neither feasibility nor compression robustness. Finally, switching-gradient methods (SGM) [23, 27, 30, 36] provide a primal-only, projection-free mechanism for constrained optimization: if w_t is nearly feasible, update along ∇f; if not, update along ∇g. This design achieves the optimal $\mathcal{O}(\epsilon^{-2})$ rate for convex, possibly non-smooth problems, and recent work [7, 15] extends optimality guarantees to weakly convex objectives. However, all existing analyses assume centralized, synchronous, full-gradient access, which is not suitable for federated systems with compression and local updates.

Contributions.

- We investigate FEDSGM as a unifying backbone for constrained FL by extending the primal-only philosophy of SGM. Unlike AL/ADMM-based methods, FEDSGM avoids dual-variable tuning and penalty scheduling, ensuring lightweight per-round computation while certifying the feasibility of the averaged iterate.
- We analyze FEDSGM under multiple local steps ($E > 1$) by bounding the drift between local and global iterates. This yields rates of the form $\mathcal{O}\left(\frac{DG\sqrt{E}}{\sqrt{T}}\sqrt{1+\frac{q_u}{n}}\right)$, where $D := \|w_0 - w^*\|$, T is the total number of rounds, G is the Lipschitz constant, and q_u the uplink compression factor. The analysis recovers the canonical $1/\sqrt{T}$ rate from centralized SGM, while explicitly quantifying the impact of multi-step local updates and compression.
- We extend both *hard switching*, a binary choice of update, and *soft switching* [36], a smooth interpolation between ∇f and ∇g via a smooth weighting function based on the magnitude of violation. Theorem 2 recovers the hard switching regime while improving stability near the feasibility boundary, without impacting the convergence rate.

Together, these results provide the first convergence guarantees for constrained FL with uplink-unbiased compression, multiple local updates, and soft switching.

2. Problem Setup and Preliminaries

Notation. We write $[T] := \{0, 1, \ldots, T-1\}$ for an index set of length T, and $[n] := \{1, 2, \ldots, n\}$ for the set of all clients. For a set $\mathcal{A}$, $|\mathcal{A}|$ denotes its cardinality. We use **Id** to denote the identity mapping, i.e., $\mathbf{Id}(x) = x$ for all x. For a scalar $z \in \mathbb{R}$, we define the positive part operator $[z]_+ := \max\{0, z\}$. Additionally, $\mathbb{1}_{\{\cdot\}}$ denotes the standard binary indicator function.

In this work, we consider the problem defined in (1), which is a standard constrained optimization problem in the FL setting, where $f_j : \mathbb{R}^d \to \mathbb{R}$ denotes the local objective of client j and $g_j : \mathbb{R}^d \to \mathbb{R}$ denotes its local constraint function. Here, n is the total number of clients and $w \in \mathbb{R}^d$ are the model parameters. The functions $f(\cdot)$ and $g(\cdot)$ represent the global objective and global constraint, respectively. The goal is to compute an ϵ-solution $\bar{w}$ that satisfies $f(\bar{w}) - f(w^*) \leq \epsilon$ and $g(\bar{w}) \leq \epsilon$. Following are the assumptions used in our analysis.

Assumption 1 *Each function f_j and g_j is convex and G-Lipschitz continuous in $\mathbb{R}^d$. Consequently, f and g are also convex and G-Lipschitz.*

Assumption 2 (Unbiased Compression) *Each client $j \in [n]$ uses an independent stochastic operator $\mathcal{C}_j : \mathbb{R}^d \to \mathbb{R}^d$ such that for all $w \in \mathbb{R}^d$, $\mathbb{E}[\mathcal{C}_j(w)] = w, \mathbb{E}[\|\mathcal{C}_j(w) - w\|^2] \leq q_u \|w\|^2$, where $q_u \geq 0$ is the compression factor.*

3. FEDSGM: Federated Switching Gradient Methods

We introduce FEDSGM, a projection-free and duality-free algorithm for constrained optimization in FL. FEDSGM combines **Fed**erated learning with the **S**witching **G**radient **M**ethod, focusing here on the setting with: (i) full client participation, (ii) both hard and soft switching between objective and constraint updates, (iii) multiple local updates per communication round, and (iv) uplink compression.

At round t, the server collects constraint values $\{g_j(w_t)\}_{j=1}^n$ and broadcasts the global average $g(w_t)$. Each client j initializes $w_{j,0}^t = w_t$ and performs E local steps. The update direction is

$$\nu_{j,\tau}^t \;=\; (1 - \sigma_\beta(g(w_t) - \epsilon))\nabla f_j(w_{j,\tau}^t) + \sigma_\beta(g(w_t) - \epsilon)\nabla g_j(w_{j,\tau}^t),$$

where σ_β is the switching rule: hard switching $\sigma_\beta(z) = \mathbb{1}_{\{z>0\}}$, or soft switching $\sigma_\beta(z) = \min\{1, [1 + \beta z]_+\}$. Clients then set $w_{j,\tau+1}^t = w_{j,\tau}^t - \eta \nu_{j,\tau}^t$. After E steps, each client transmits the compressed difference $\Delta_{j,E}^t = \mathcal{C}((w_t - w_{j,E}^t)/\eta)$, with $\mathcal{C} \equiv \mathbf{Id}$ denoting no compression. The server updates

$$w_{t+1} \;\leftarrow\; w_t - \eta \cdot \tfrac{1}{n} \sum_{j=1}^n \Delta_{j,E}^t.$$

This unified design (Alg. 1, see Appendix A) cleanly subsumes both hard and soft switching under compression.

Motivation for Soft Switching. Hard switching enforces a binary rule: each client either optimizes the objective or the constraint, depending on whether the global violation $g(w_t)$ is within the tolerance ϵ. While simple, this approach is highly sensitive when $g(w_t)$ fluctuates around ϵ as already shown in the centralized case [36]. Such fluctuations can trigger frequent back-and-forth updates, amplifying client drift and resulting in unstable trajectories.

Soft switching provides a remedy by introducing a smooth interpolation between the two gradients [36]. Specifically, the switching weight $\sigma_\beta(g(w_t)-\epsilon)$ ensures that near the feasibility boundary, the update direction is a convex combination of ∇f_j and ∇g_j, rather than an abrupt choice. This smooth relaxation suppresses oscillations while preserving convergence guarantees. Our analysis demonstrates that, even under compression, soft switching attains the same order of convergence as hard switching, while empirical validation confirms its superior stability near the feasibility boundary. To the best of our knowledge, this is the first convergence guarantee establishing the effectiveness of soft switching for constrained FL with compression, thereby extending beyond prior results restricted to hard switching and single local updates [17]. Next, we state the theorems that provide the convergence guarantees for FEDSGM in Algorithm 1.

Theorem 1 (Hard switching FEDSGM) *Consider the problem in (1) and Algorithm 1, under Assumptions 1 and 2 (only for compression). Let $D := \|w_0 - w^*\|$ and define*

$$\mathcal{A} := \{t \in [T] \mid g(w_t) \le \epsilon\}, \quad \bar{w} := \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t.$$

Let

$$\Gamma = 2E^2 + \frac{Eq_u}{2n}\mathbb{1}_{\{\textit{Unbiased Compression}\}},$$

where q_u is client-to-server compression factor. Now, set the constraint threshold and step size as

$$\epsilon = \sqrt{\frac{2D^2G^2\Gamma}{ET}}, \quad \eta = \sqrt{\frac{D^2}{2G^2ET\Gamma}},$$

then $\mathcal{A}$ is nonempty, $\bar{w}$ is well-defined, and $\bar{w}$ is an ϵ-solution of (1).

Theorem 2 (Soft switching FEDSGM) *Consider the problem in (1) and Algorithm 1, under Assumptions 1 and 2 (only for compression). Let $D := \|w_0 - w^*\|$ and define*

$$\mathcal{A} = \{t \in [T] \mid g(w_t) < \epsilon\}, \quad \bar{w} = \sum_{t\in\mathcal{A}} \alpha_t w_t, \quad \textit{where} \quad \alpha_t = \frac{1-\sigma_\beta(g(w_t)-\epsilon)}{\sum_{s\in\mathcal{A}}\left[1-\sigma_\beta(g(w_s)-\epsilon)\right]}.$$

Let

$$\Gamma = 2E^2 + \frac{Eq_u}{2n}\mathbb{1}_{\{\textit{Unbiased Compression}\}},$$

where q_u is client-to-server compression factor. Now, set the constraint threshold and step size as

$$\epsilon = \sqrt{\frac{2D^2G^2\Gamma}{ET}}, \quad \eta = \sqrt{\frac{D^2}{2G^2ET\Gamma}}, \quad \textit{and} \ \ \beta \ge \frac{2}{\epsilon},$$

then $\mathcal{A}$ is nonempty, $\bar{w}$ is well-defined, and $\bar{w}$ is an ϵ-solution of (1).

The details of the proof are present in the Appendix A. Now, we discuss the implications of the statement of Theorems 1 and 2.

• $n = 1, q_u = 0, \mathcal{C}_j \equiv \mathbf{Id}, E = 1$ **, i.e., centralized with no compression**: In this case, we can infer from Theorems 1 and 2, that rates we receive is $\mathcal{O}\left(DG/\sqrt{T}\right)$, corresponding to the rates present in [23, 28, 36].

• $g_j \equiv 0, \forall j \in [n], E = 1$**, i.e., no constraint**: In this case, our algorithm reduces to the well-known FEDCOM method [12], where we recover the standard rate of $\mathcal{O}\left(\frac{DG\sqrt{E}}{\sqrt{T}}\sqrt{1+\frac{q_u}{n}}\right)$.
• $q_u = 0, \mathcal{C}_j =$ **Id , i.e., FEDSGM w/ full participation w/o compression**: In this case, we can infer from Theorems 1 and 2 that the sub-optimality gap and constraint value diminish at the rate of $\mathcal{O}\left(\frac{DG\sqrt{E}}{\sqrt{T}}\right)$. The scaling of $\sqrt{E}$ captures the effect of client-drift in this federated constrained setting, deviating from the previous centralized results.
• **FEDSGM with uplink compression**: We focus on the case where only uplink transmissions are compressed, which captures the dominant communication bottleneck in federated learning. Most prior works on uplink compression have studied restricted classes of compressors—such as absolute compression [34] or unbiased operators [10, 11, 29, 35]—and largely target unconstrained or single-step settings. Recent progress on biased compressors [3, 16, 31] has established convergence guarantees for unconstrained optimization, while constrained FL with compression has only been addressed under hard switching and single local updates [17]. In this work, we analyze FEDSGM in the unbiased uplink regime, extending the scope to multiple local updates and, crucially, to soft switching between objective and constraint updates. To the best of our knowledge, this is the first convergence result showing that soft switching remains effective in constrained FL under uplink-unbiased compression, thereby laying the groundwork for subsequent extensions to biased and bi-directional compression.

4. Numerical Experiments

We evaluate the proposed method in a federated Neyman–Pearson (NP) classification setting with constrained optimization objectives. The aim is to examine how different switching strategies behave under uplink compression, with particular focus on convergence, stability, and feasibility.

4.1. NP Classification

We consider the constrained optimization problem in (1), where the objective is to minimize the empirical loss on the majority class while ensuring that the loss on the minority class remains below a prescribed tolerance. For each client j, the local objective and constraint are defined as

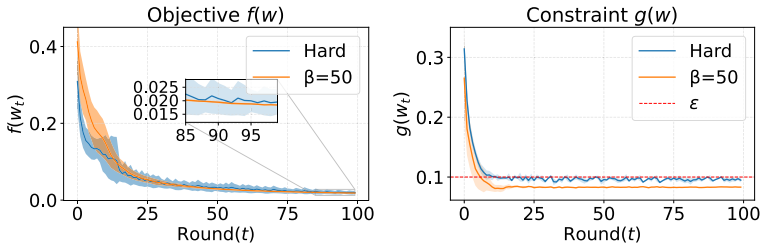
$$f_j(w) := \frac{1}{m_{j0}} \sum_{x \in \mathcal{D}_j^{(0)}} \phi(w; (x, 0)), \qquad g_j(w) := \frac{1}{m_{j1}} \sum_{x \in \mathcal{D}_j^{(1)}} \phi(w; (x, 1)), \tag{2}$$

where $\mathcal{D}_j^{(0)}$ and $\mathcal{D}_j^{(1)}$ denote local samples of class 0 and class 1, respectively, and m_{j0}, m_{j1} are their cardinalities. The function ϕ is the binary logistic loss,

$$\phi(w; (x, y)) = -y\, w^\top x + \log\left(1 + e^{w^\top x}\right), \quad y \in \{0, 1\}. \tag{3}$$

This captures the NP paradigm: $f(w)$ enforces performance on the majority class, while the constraint $g(w) \leq \epsilon$ ensures the minority class loss does not exceed the tolerance.

We use the breast cancer dataset [37], containing 569 samples with 30 features. To simulate the federated setting, the data is split in an IID fashion between $n = 10$ clients, such that each client receives an equal number of samples and the same class ratio. Each client performs local gradient descent updates for $E = 5$ epochs before communication, and the global model is updated



Switching	# Violations
Hard	32.00 ± 8.49
Soft ($\beta = 50$)	7.67 ± 2.62

Figure 1: Federated NP classification results under Rand-K compression with $K = 9$. Left: evolution of the class-0 objective loss $f(w)$; Middle: class-1 constraint loss $g(w)$ relative to the tolerance $\epsilon = 0.1$; Right: number of constraint violations (out of 100 rounds). Results are averaged over three random seeds and reported as mean $\pm$ standard deviation.

for $T = 100$ communication rounds. We compare both hard and soft switching using a tolerance $\epsilon = 0.1$.

For communication efficiency, we evaluate using Rand-K [12, 32] compressor, which transmits a fraction K/d of coordinates with unbiased scaling. Performance is measured in terms of the majority-class objective loss $f(w)$, the minority-class constraint loss $g(w)$, and the number of rounds in which the feasibility condition $g(w) \leq \epsilon$ is violated. Each experiment is repeated with three random seeds, and we report mean trajectories with variance bands. From Figure 1, we observe that both hard and soft switching achieve convergence of the majority-class objective $f(w)$. However, soft switching exhibits markedly improved stability in the constraint $g(w)$: while hard switching frequently oscillates around the tolerance, soft switching maintains feasibility more consistently. This translates into a substantial reduction in the number of violations, with soft switching yielding around $4\times$ fewer violations compared to hard switching. Notably, this improvement in feasibility is achieved without compromising the convergence behavior of the objective loss, and in fact ensures more stable results as evident from the variance plots, highlighting the advantage of soft switching in balancing accuracy and constraint satisfaction under communication compression.

5. Conclusion

FEDSGM provides a unified algorithmic framework for constrained federated learning by integrating projection-free and duality-free switching-gradient methods with multi-step local updates and uplink compression. Our analysis provides the first convergence guarantees in this regime, yielding rates of order $\mathcal{O}\left(\frac{DG\sqrt{E}}{\sqrt{T}}\sqrt{1+\frac{q_u}{n}}\right)$ with explicit dependence on the compression factor, and recovers classical centralized and FEDCOM bounds as special cases. Beyond hard switching, we demonstrated that soft switching recovers the hard regime when $\beta \geq 2/\epsilon$, but is empirically more stable near the feasibility boundary, suppressing oscillations without altering the rate. Overall, FEDSGM robustly balances feasibility, client drift, and communication efficiency, and lays the foundation for extensions to biased/bidirectional compression and partial participation.

Acknowledgements

This work was supported in part by NSF CNS 2313109 and Wistron Corporation.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pages 60–69. PMLR, 2018.

[2] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

[3] Aleksandr Beznosikov, Samuel Horváth, Peter Richtárik, and Mher Safaryan. On biased compression for distributed learning. *Journal of Machine Learning Research*, 24(276):1–50, 2023.

[4] Digvijay Boob, Qi Deng, and Guanghui Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279, 2023.

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[6] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[7] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[8] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Alejandro Ribeiro. Last-iterate convergent policy gradient primal-dual methods for constrained mdps. *Advances in Neural Information Processing Systems*, 36:66138–66200, 2023.

[9] Ilyas Fatkhullin, Igor Sokolov, Eduard Gorbunov, Zhize Li, and Peter Richtárik. Ef21 with bells & whistles: Practical algorithmic extensions of modern error feedback. *arXiv preprint arXiv:2110.03294*, 2021.

[10] Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Ef21-p and friends: Improved theoretical communication complexity for distributed optimization with bidirectional compression. In *International conference on machine learning*, pages 11761–11807. PMLR, 2023.

[11] Kaja Gruntkowska, Alexander Tyurin, and Peter Richtárik. Improving the worst-case bidirectional communication complexity for nonconvex distributed optimization under function similarity. *Advances in Neural Information Processing Systems*, 37:88807–88873, 2024.

[12] Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2350–2358. PMLR, 2021.

[13] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31 (2):1299–1329, 2021.

[14] Chuan He, Le Peng, and Ju Sun. Federated learning with convex global and local constraints. *Transactions on machine learning research*, 2024:https–openreview, 2024.

[15] Yankun Huang and Qihang Lin. Oracle complexity of single-loop switching subgradient methods for non-smooth weakly convex functional constrained optimization. *Advances in Neural Information Processing Systems*, 36:61327–61340, 2023.

[16] Rustem Islamov, Yuan Gao, and Sebastian U Stich. Towards faster decentralized stochastic optimization with communication compression. *arXiv preprint arXiv:2405.20114*, 2024.

[17] Rustem Islamov, Yarden As, and Ilyas Fatkhullin. Safe-ef: Error feedback for nonsmooth constrained optimization. *arXiv preprint arXiv:2505.06053*, 2025.

[18] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[19] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018.

[20] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local gd on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

[21] Jong Gwang Kim, Ashish Chandra, Abolfazl Hashemi, and Christopher Brinton. A fast single-loop primal-dual algorithm for non-convex functional constrained optimization. *arXiv preprint arXiv:2406.17107*, 2024.

[22] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

[23] Guanghui Lan and Zhiqiang Zhou. Algorithms for stochastic optimization with function or expectation constraints. *Computational Optimization and Applications*, 76(2):461–498, 2020.

[24] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[25] Adrian Müller, Pragnya Alatur, Volkan Cevher, Giorgia Ramponi, and Niao He. Truly no-regret learning in constrained mdps. *arXiv preprint arXiv:2402.15776*, 2024.

[26] Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[27] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[28] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[29] Constantin Philippenko and Aymeric Dieuleveut. Preserved central model for faster bidirectional compression in distributed settings. *Advances in Neural Information Processing Systems*, 34:2387–2399, 2021.

[30] Boris T Polyak. A general method for solving extremal problems. In *Soviet Mathematics Doklady*, volume 8, pages 593–597, 1967.

[31] Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. Fednl: Making newton-type methods applicable to federated learning. *arXiv preprint arXiv:2106.02969*, 2021.

[32] Egor Shulgin and Peter Richtárik. Shifted compression framework: Generalizations and improvements. In *Uncertainty in Artificial Intelligence*, pages 1813–1823. PMLR, 2022.

[33] Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

[34] Hanlin Tang, Chen Yu, Xiangru Lian, Tong Zhang, and Ji Liu. Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression. In *International Conference on Machine Learning*, pages 6155–6165. PMLR, 2019.

[35] Alexander Tyurin and Peter Richtarik. 2direction: Theoretically faster distributed training with bidirectional communication compression. *Advances in Neural Information Processing Systems*, 36:11737–11808, 2023.

[36] Antesh Upadhyay, Sang Bin Moon, and Abolfazl Hashemi. Optimization via first-order switching methods: Skew-symmetric dynamics and optimistic discretization. *arXiv preprint arXiv:2505.09146*, 2025.

[37] William H. Wolberg, Olvi L. Mangasarian, Nick Street, and W. N. Street. Breast Cancer Wisconsin (Diagnostic) Dataset. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.

[38] Han Zhang, Xiaofan Gui, Shun Zheng, Ziheng Lu, Yuqi Li, and Jiang Bian. Batteryml: An open-source platform for machine learning on battery degradation. *arXiv preprint arXiv:2310.14714*, 2023.

[39] Zhe Zhang and Guanghui Lan. Solving convex smooth function constrained optimization is almost as easy as unconstrained optimization. *arXiv preprint arXiv:2210.05807*, 2022.

Appendix A. Appendix

Algorithm 1: FedSGM($T, E, \mathcal{C}_j, \eta, \beta$)

Input: Number of rounds T; local updates E; compression operator $\mathcal{C}_j$ (with $\mathcal{C}_j = \mathbf{Id}$ denoting no compression); learning rate η; initial model w_0; switching rule $\sigma_\beta(\cdot)$ (hard or soft)

for $t \leftarrow 0$ **to** $T-1$ **do**
  foreach *client* $j \in [n]$ `// in parallel` **do**
    send $g_j(w_t)$ to server `// cheap single float communication`
  end
  compute $g(w_t) \leftarrow \frac{1}{n}\sum_{j=1}^{n} g_j(w_t)$ and broadcast;
  foreach *client* $j \in [n]$ `// in parallel` **do**
    set $w_{j,0}^t \leftarrow w_t$;
    for $\tau \leftarrow 0$ **to** $E-1$ **do**
      compute switching weight $\alpha_t \leftarrow \sigma_\beta(g(w_t) - \epsilon)$;
      update direction $\nu_{j,\tau}^t \leftarrow (1-\alpha_t)\nabla f_j(w_{j,\tau}^t) + \alpha_t \nabla g_j(w_{j,\tau}^t)$;
      update $w_{j,\tau+1}^t \leftarrow w_{j,\tau}^t - \eta \nu_{j,\tau}^t$;
    end
    send $\Delta_{j,E}^t \leftarrow \mathcal{C}_j\left(\frac{w_t - w_{j,E}^t}{\eta}\right)$ to server;
  end
  server computes $w_{t+1} \leftarrow w_t - \eta \cdot \frac{1}{n}\sum_{j=1}^{n} \Delta_{j,E}^t$ and broadcasts;
end

A.1. Lemmas

Lemma 3 (*Bound on the Expected Norm of Compressed Aggregates*) *Under Assumptions 1 and 2, for all rounds $t \in [T]$, the following bound holds,*

$$\mathbb{E}_{\mathcal{C}_j}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\mathcal{C}_j\left(\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t\right)\right\|^2\right] \le \left(\frac{q_u}{n}+1\right)E^2G^2.$$

Proof Let $X_j^t := \sum_{\tau=0}^{E-1}\nu_{j,\tau}^t$. We aim to bound:

$$\mathbb{E}_{\mathcal{C}_j}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\mathcal{C}_j(X_j^t)\right\|^2\right].$$

Using the identity for the second moment:

$$\mathbb{E}[\|Y\|^2] = \|\mathbb{E}[Y]\|^2 + \mathbb{E}[\|Y - \mathbb{E}[Y]\|^2],$$

we apply this to our setup:

$$\mathbb{E}_{\mathcal{C}_j}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\mathcal{C}_j(X_j^t)\right\|^2\right] = \left\|\frac{1}{n}\sum_{j=1}^{n}X_j^t\right\|^2 + \mathbb{E}_{\mathcal{C}_j}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\left(\mathcal{C}_j(X_j^t)-X_j^t\right)\right\|^2\right]$$
$$\leq \left\|\frac{1}{n}\sum_{j=1}^{n}X_j^t\right\|^2 + \frac{1}{n^2}\sum_{j=1}^{n}\mathbb{E}_{\mathcal{C}_j}\left[\|\mathcal{C}_j(X_j^t)-X_j^t\|^2\right],$$

where the inequality uses independence across j and linearity of expectation. Applying Assumption 2 (variance bound of the compression operator), $\mathbb{E}_{\mathcal{C}_j}\left[\|\mathcal{C}_j(X_j^t)-X_j^t\|^2\right] \leq q_u\|X_j^t\|^2$, we get,

$$\mathbb{E}_{\mathcal{C}_j}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\mathcal{C}_j(X_j^t)\right\|^2\right] \leq \left\|\frac{1}{n}\sum_{j=1}^{n}X_j^t\right\|^2 + \frac{q_u}{n^2}\sum_{j=1}^{n}\|X_j^t\|^2$$
$$\overset{Jensen's}{\leq} \frac{1}{n}\sum_{j=1}^{n}\|X_j^t\|^2 + \frac{q_u}{n^2}\sum_{j=1}^{n}\|X_j^t\|^2$$
$$= \left(\frac{q_u}{n}+1\right)\cdot\frac{1}{n}\sum_{j=1}^{n}\|X_j^t\|^2.$$

Now substitute $X_j^t = \sum_{\tau=0}^{E-1}\nu_{j,\tau}^t$ and apply Jensen's inequality again,

$$\frac{1}{n}\sum_{j=1}^{n}\|X_j^t\|^2 = \frac{1}{n}\sum_{j=1}^{n}\left\|\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t\right\|^2 \leq \frac{1}{n}\sum_{j=1}^{n}E\sum_{\tau=0}^{E-1}\|\nu_{j,\tau}^t\|^2$$
$$\overset{G-Lip}{\leq} E^2G^2.$$

Hence,

$$\mathbb{E}_{\mathcal{C}_j}\left[\left\|\frac{1}{n}\sum_{j=1}^{n}\mathcal{C}_j(X_j^t)\right\|^2\right] \leq \left(\frac{q_u}{n}+1\right)E^2G^2.$$

■

Lemma 4 (*Inner Product Bound under Compression and Switching*) *Under Assumptions 1 and 2, for all rounds $t \in [T]$, the following bound holds,*

$$\mathbb{E}_{\mathcal{C}_j}\left[-2\eta\left\langle w_t - w^*, \frac{1}{n}\sum_{j=1}^{n}\mathcal{C}_j\left(\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t\right)\right\rangle\right] = -2\eta\left\langle w_t - w^*, \frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t\right\rangle$$
$$\leq \frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\begin{cases}\frac{\eta}{\alpha}\|w_t - w_{j,\tau}^t\|^2 + \eta\alpha G^2 + 2\eta(f_j(w^*) - f_j(w_{j,\tau}^t)), & \text{if } t \in \mathcal{A},\\ \frac{\eta}{\alpha}\|w_t - w_{j,\tau}^t\|^2 + \eta\alpha G^2 + 2\eta(g_j(w^*) - g_j(w_{j,\tau}^t)), & \text{if } t \in \mathcal{B}.\end{cases}$$

Proof We now analyze the cross term in the squared distance recursion,

$$-2\eta \left\langle w_t - w^*, \frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t \right\rangle = \frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\left[\underbrace{-2\eta\left\langle w_t - w_{j,\tau}^t, \nu_{j,\tau}^t\right\rangle}_{TermA}\underbrace{-2\eta\left\langle w_{j,\tau}^t - w^*, \nu_{j,\tau}^t\right\rangle}_{TermB}\right].$$

We handle the second term by applying the convexity of f_j or g_j. For $t \in \mathcal{A}$, where updates use ∇f_j we apply:

$$f_j(w^*) \geq f_j(w_{j,\tau}^t) + \left\langle \nabla f_j(w_{j,\tau}^t), w^* - w_{j,\tau}^t \right\rangle,$$

which implies:

$$-\left\langle w_{j,\tau}^t - w^*, \nabla f_j(w_{j,\tau}^t)\right\rangle \leq f_j(w^*) - f_j(w_{j,\tau}^t).$$

Thus,

$$TermB = -2\eta\left\langle w_{j,\tau}^t - w^*, \nabla f_j(w_{j,\tau}^t)\right\rangle \leq 2\eta\left(f_j(w^*) - f_j(w_{j,\tau}^t)\right).$$

A similar argument holds for $t \in \mathcal{B}$ with ∇g_j, resulting in $TermB \leq 2\eta\left(g_j(w^*) - g_j(w_{j,\tau}^t)\right)$. Again, while upper bounding $TermA$, we need to deal with 2 cases depending on whether $t \in \mathcal{A}$ or $t \in \mathcal{B}$. Firstly, we start with the case where $t \in \mathcal{A}$, for any $\alpha > 0$

$$\begin{aligned} TermA = -2\eta\left\langle w_t - w_{j,\tau}^t, \nabla f_j(w_{j,\tau}^t)\right\rangle &\overset{\text{Young's}}{\leq} \frac{\eta}{\alpha}\|w_t - w_{j,\tau}^t\|^2 + \eta\alpha\|\nabla f_j(w_{j,\tau}^t)\|^2 \\ &\overset{G-Lip}{\leq} \frac{\eta}{\alpha}\|w_t - w_{j,\tau}^t\|^2 + \eta\alpha G^2. \end{aligned}$$

Similarly, for $t \in \mathcal{B}$, we get

$$TermA = -2\eta\left\langle w_t - w_{j,\tau}^t, \nabla f_j(w_{j,\tau}^t)\right\rangle \leq \frac{\eta}{\alpha}\|w_t - w_{j,\tau}^t\|^2 + \eta\alpha G^2.$$

Substituting the bounds for both $TermA$ and $TermB$ back into the original expectation furnishes the proof. ■

Lemma 5 (*Global-local iterate bound*) *Under Assumptions 1, for all rounds $t \in [T]$, the following bound holds,*

$$\sum_{\tau=0}^{E-1}\|w_t - w_{j,\tau}^t\|^2 \leq \frac{1}{3}\eta^2 E^3 G^2$$

Proof

$$\begin{aligned} \|w_t - w_{j,\tau}^t\|^2 &= \left\| w_t - \left(w_t - \eta\sum_{k=0}^{\tau-1}\nu_{j,k}^t\right)\right\|^2 \\ &= \eta^2\left\|\sum_{k=0}^{\tau-1}\nu_{j,k}^t\right\|^2 \\ &\overset{Jensen's}{\leq} \eta^2\tau\sum_{k=0}^{\tau-1}\|\nu_{j,k}^t\|^2 \\ &\overset{G-Lip}{\leq} \eta^2\tau^2 G^2 \end{aligned}$$

Therefore,

$$\sum_{\tau=0}^{E-1} \|w_t - w_{j,\tau}^t\|^2 \overset{\text{sum of squares}}{\leq} \frac{1}{3}\eta^2 E^3 G^2.$$

■

A.2. Main Theorem FedSGM—Unbiased Compression

A.2.1. HARD SWITCHING — FULL PARTICIPATION

Theorem 6 (*FedSGCM*) *Consider the problem in eq. (1) and Algorithm 1, under Assumptions 1 and 2. Define* $D := \|w_0 - w^*\|$ *and*

$$\mathcal{A} = \{t \in [T] \mid g(w_t) \leq \epsilon\}, \qquad \bar{w} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t.$$

Then, if

$$\epsilon = \sqrt{\frac{2D^2G^2\Gamma}{ET}}, \text{ and } \eta = \sqrt{\frac{D^2}{2G^2ET\Gamma}}, \text{ where } \Gamma = 2E^2 + \frac{Eq_u}{2n}$$

it holds that $\mathcal{A}$ *is nonempty,* $\bar{w}$ *is well-defined, and* $\bar{w}$ *is an* ϵ*-solution for* P*.*

Proof Using Algorithm 1, the update rule for the global model is

$$w_{t+1} = w_t - \eta \cdot \frac{1}{n}\sum_{j=1}^{n} \mathcal{C}_j\left(\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t\right), \tag{4}$$

where we define the compressed update using a stochastic compression operator $\mathcal{C}_j(\cdot)$, and also

$$\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t = \frac{w_t - w_{j,E}^t}{\eta}.$$

We analyze the squared distance to the optimal point w^* as follows,

$$\begin{aligned}
\|w_{t+1} - w^*\|^2 &= \left\| w_t - \eta \cdot \frac{1}{n}\sum_{j=1}^{n} \mathcal{C}_j\left(\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t\right) - w^* \right\|^2 \\
&= \|w_t - w^*\|^2 + \eta^2 \left\| \frac{1}{n}\sum_{j=1}^{n} \mathcal{C}_j\left(\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t\right) \right\|^2 - 2\eta \left\langle w_t - w^*, \frac{1}{n}\sum_{j=1}^{n} \mathcal{C}_j\left(\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t\right) \right\rangle
\end{aligned}$$

Taking the expectation with respect to the compression operator $\mathcal{C}_j$ and using Lemmas 3 and 4, we get

$$\begin{aligned}\|w_{t+1}-w^*\|^2 \leq \|w_t-w^*\|^2 &+ \eta^2\left(\frac{q_u}{n}+1\right)E^2G^2 + \eta\alpha EG^2 \\ &+ \frac{\eta}{\alpha n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\|w_t - w_{j,\tau}^t\|^2 \\ &+ \frac{2\eta}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}(f_j(w^*)-f_j(w_{j,\tau}^t))\mathbb{1}\{t\in\mathcal{A}\} \\ &+ \frac{2\eta}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}(g_j(w^*)-g_j(w_{j,\tau}^t))\mathbb{1}\{t\in\mathcal{B}\}. \end{aligned} \tag{5}$$

Now our goal is to handle the term $f_j(w^*) - f_j(w_{j,\tau}^t)$, so we first rewrite,

$$f_j(w_{j,\tau}^t) \geq f_j(w_t) + \langle \nabla f_j(w_t), w_{j,\tau}^t - w_t\rangle \quad \text{(by convexity)}$$

$$\Rightarrow f_j(w^*) - f_j(w_{j,\tau}^t) \leq f_j(w^*) - f_j(w_t) - \langle \nabla f_j(w_t), w_{j,\tau}^t - w_t\rangle.$$

Using Young's inequality with parameter $\alpha > 0$, we get

$$-\langle \nabla f_j(w_t), w_{j,\tau}^t - w_t\rangle \leq \frac{1}{2\alpha}\|w_{j,\tau}^t - w_t\|^2 + \frac{\alpha}{2}\|\nabla f_j(w_t^t)\|^2.$$

Thus, we get,

$$\begin{aligned} f_j(w^*) - f_j(w_{j,\tau}^t) &\leq f_j(w^*) - f_j(w_t) + \frac{1}{2\alpha}\|w_t - w_{j,\tau}^t\|^2 + \frac{\alpha}{2}\|\nabla f_j(w_t)\|^2 \\ &\overset{G-Lip}{\leq} f_j(w^*) - f_j(w_t) + \frac{1}{2\alpha}\|w_t - w_{j,\tau}^t\|^2 + \frac{\alpha}{2}G^2. \end{aligned}$$

Similarly, we can handle the other term with g and get,

$$g_j(w^*) - g_j(w_{j,\tau}^t) \leq g_j(w^*) - g_j(w_t) + \frac{1}{2\alpha}\|w_t - w_{j,\tau}^t\|^2 + \frac{\alpha}{2}G^2.$$

So, putting these inequalities back in eq. (5) along with the use of Lemma 5 and eq. (1), we get

$$\begin{aligned}\|w_{t+1}-w^*\|^2 \leq \|w_t-w^*\|^2 &+ \eta^2\left(\frac{q_u}{n}+1\right)E^2G^2 + 2\eta\alpha EG^2 + \frac{2}{3\alpha}\eta^3E^3G^2 \\ &+ 2\eta E(f(w^*)-f(w_t))\mathbb{1}\{t\in\mathcal{A}\} + 2\eta E(g(w^*)-g(w_t))\mathbb{1}\{t\in\mathcal{B}\}.\end{aligned}$$

Now, for $\alpha = \eta$, and rearranging the terms we get

$$\begin{aligned}(f(w_t)-f(w^*))\mathbb{1}\{t\in\mathcal{A}\} + (g(w_t)-g(w^*))\mathbb{1}\{t\in\mathcal{B}\} \leq &\frac{\|w_t-w^*\|^2 - \|w_{t+1}-w^*\|^2}{2\eta E} \\ &+\frac{\eta}{2}\left(\frac{q_u}{n}+1\right)EG^2 + \eta G^2 + \frac{1}{3}\eta E^2G^2.\end{aligned}$$

Defining $D := \|w_0 - w^*\|$ and summing the expression above for $t = 0, 1, \cdots, T-1$ and then dividing by T, we get

$$\frac{1}{T}\sum_{t\in\mathcal{A}}(f(w_t) - f(w^*)) + \frac{1}{T}\sum_{t\in\mathcal{B}}(g(w_t) - g(w^*)) \leq \frac{D^2}{2\eta ET} + \frac{\eta}{2}\left(\frac{q_u}{n} + 1\right)EG^2 + \eta G^2 + \frac{1}{3}\eta E^2 G^2.$$

Now choosing $\eta = \sqrt{\frac{D^2}{2G^2ET\Gamma}}$, where $\Gamma = \frac{1}{2}\left(\frac{q_u}{n} + 1\right)E + 1 + \frac{1}{3}E^2$, we get

$$\frac{1}{T}\sum_{t\in\mathcal{A}}(f(w_t) - f(w^*)) + \frac{1}{T}\sum_{t\in\mathcal{B}}(g(w_t) - g(w^*)) \leq \sqrt{\frac{2D^2G^2\Gamma}{ET}}.$$

Note that when ϵ is sufficiently large, $\mathcal{A}$ is nonempty. Assuming an empty $\mathcal{A}$, we can find the largest "bad" ϵ:

$$\epsilon_{bad} < \frac{1}{T}\sum_{t\in\mathcal{B}} g(w_t) - g(w^*) \leq \sqrt{\frac{2D^2G^2\Gamma}{ET}}$$

Thus, let us set $\epsilon = (N+1)\sqrt{\frac{2D^2G^2\Gamma}{ET}}$ for some $N \geq 0$. With this choice, $\mathcal{A}$ is guaranteed to be nonempty.

Now, we consider two cases. Either $\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) \leq 0$ which implies by the convexity of f and g for $\bar{w} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t$ we have

$$f(\bar{w}) - f(w^*) \leq 0 < \epsilon, \qquad g(\bar{w}) \leq \epsilon. \tag{6}$$

Otherwise, if $\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) > 0$, then

$$\begin{aligned}
\sqrt{\frac{2D^2G^2\Gamma}{ET}} &\geq \frac{1}{T}\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) + \frac{1}{T}\sum_{t\in\mathcal{B}} g(w_t) - g(w^*) \\
&> \frac{1}{T}\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) + \frac{1}{T}\sum_{t\in\mathcal{B}} \epsilon \\
&= \frac{|\mathcal{A}|}{T}\frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) + (1 - \frac{|\mathcal{A}|}{T})\epsilon \\
&\geq \frac{|\mathcal{A}|}{T}\big(f(\bar{w}) - f(w^*)\big) + (1 - \frac{|\mathcal{A}|}{T})\epsilon.
\end{aligned} \tag{7}$$

By rearranging

$$\frac{|\mathcal{A}|}{T}\Big(f(\bar{w}) - f(w^*) - \epsilon\Big) < \sqrt{\frac{2D^2G^2\Gamma}{ET}} - \epsilon \leq -N\sqrt{\frac{2D^2G^2\Gamma}{ET}}, \tag{8}$$

Implying $f(\bar{w}) - f(w^*) < \epsilon$ and further by convexity of g for $\bar{w} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t$, we also have $g(\bar{w}) \leq \epsilon$. ■

A.2.2. SOFT SWITCHING — FULL PARTICIPATION

Theorem 7 (*FedSSGCM*) *Consider the problem in (1) and Algorithm 1, under Assumption 1. Define* $D := \|w_0 - w^*\|$ *and*

$$\mathcal{A} = \{t \in [T] \mid g(w_t) < \epsilon\}, \qquad \bar{w} = \sum_{t \in \mathcal{A}} \alpha_t w_t,$$

where

$$\alpha_t = \frac{1 - \sigma_\beta(g(w_t) - \epsilon)}{\sum_{t \in \mathcal{A}}[1 - \sigma_\beta(g(w_t) - \epsilon)]}. \tag{9}$$

Then, if

$$\epsilon = \sqrt{\frac{2D^2G^2\Gamma}{ET}}, \qquad \eta = \sqrt{\frac{D^2}{2G^2ET\Gamma}}, \quad \text{and} \quad \beta = \frac{2}{\epsilon} \text{ where } \Gamma = 2E^2 + \frac{Eq}{2n}$$

it holds that $\mathcal{A}$ *is nonempty,* $\bar{w}$ *is well-defined, and* $\bar{w}$ *is an* ϵ*-solution for* P*.*

Proof Using Algorithm 1, the update rule for the global model is

$$w_{t+1} = w_t - \eta \cdot \frac{1}{n} \sum_{j=1}^{n} \mathcal{C}_j \left(\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t \right),$$
$$\text{where, } \nu_{j,\tau}^t = \sigma_\beta(g(w_t) - \epsilon)\nabla g_j(w_{j,\tau}^t) + (1 - \sigma_\beta(g(w_t) - \epsilon))\nabla f_j(w_{j,\tau}^t)$$

We analyze the squared distance to the optimal point w^* as follows,

$$\|w_{t+1} - w^*\|^2 = \left\| w_t - \eta \cdot \frac{1}{n} \sum_{j=1}^{n} \mathcal{C}_j \left(\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t \right) - w^* \right\|^2$$
$$= \|w_t - w^*\|^2 + \underbrace{\eta^2 \left\| \frac{1}{n} \sum_{j=1}^{n} \mathcal{C}_j \left(\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t \right) \right\|^2}_{Term-A} \underbrace{-2\eta \left\langle w_t - w^*, \frac{1}{n} \sum_{j=1}^{n} \mathcal{C}_j \left(\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t \right) \right\rangle}_{Term-B}$$

Firstly, we start with upper-bounding $Term - A$. Taking the expectation with respect to the compression operator $\mathcal{C}_j$ and using Lemmas 3 and the fact that $\sigma_\beta(\cdot) \in [0, 1]$, we have

$$\mathbb{E}_C[Term - A] \leq \eta^2 \left(\frac{q_u}{n} + 1 \right) E^2 G^2.$$

Now we aim to upper bound $Term - B$. First, we take the expectation with respect to $\mathcal{C}_j$ and

$$\mathbb{E}_{\mathcal{C}_j}[Term - B] = -2\eta \left\langle w_t - w^*, \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} \nu_{j,\tau}^t \right\rangle$$
$$= \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} -2\eta \left\langle w_t - w^*, \nu_{j,\tau}^t \right\rangle$$
$$= \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} \left[\underbrace{-2\eta \left\langle w_t - w_{j,\tau}^t, \nu_{j,\tau}^t \right\rangle}_{Term-B_1} \underbrace{-2\eta \left\langle w_{j,\tau}^t - w^*, \nu_{j,\tau}^t \right\rangle}_{Term-B_2} \right]$$

Now we define $\sigma_\beta^t = \sigma_\beta(g(w_t) - \epsilon)$ before we start upper bounding $Term - B_1$ for any $\alpha > 0$.

$$
\begin{aligned}
Term - B_1 &= -2\eta \left\langle w_t - w_{j,\tau}^t, \nu_{j,\tau}^t \right\rangle \\
&= -2\eta\sigma_\beta^t \left\langle w_t - w_{j,\tau}^t, g_j(w_{j,\tau}^t) \right\rangle - 2\eta(1-\sigma_\beta^t) \left\langle w_t - w_{j,\tau}^t, f_j(w_{j,\tau}^t) \right\rangle \\
&\overset{Young's}{\leq} \sigma_\beta^t \left[\frac{\eta}{\alpha} \|w_t - w_{j,\tau}^t\|^2 + \eta\alpha \|\nabla g_j(w_{j,\tau}^t)\|^2 \right] + (1-\sigma_\beta^t) \left[\frac{\eta}{\alpha} \|w_t - w_{j,\tau}^t\|^2 + \eta\alpha \|\nabla f_j(w_{j,\tau}^t)\|^2 \right] \\
&\overset{G-Lip}{\leq} \frac{\eta}{\alpha} \|w_t - w_{j,\tau}^t\|^2 + \eta\alpha G^2
\end{aligned}
$$

Similarly, we start to upper bound $Term - B_2$ as well.

$$
\begin{aligned}
Term - B_2 &= -2\eta \left\langle w_{j,\tau}^t - w^*, \nu_{j,\tau}^t) \right\rangle \\
&= -2\eta\sigma_\beta^t \left\langle w_{j,\tau}^t - w^*, g_j(w_{j,\tau}^t) \right\rangle - 2\eta(1-\sigma_\beta^t) \left\langle w_{j,\tau}^t - w^*, f_j(w_{j,\tau}^t) \right\rangle \\
&\overset{Cvx.}{\leq} 2\eta\sigma_\beta^t \left(g_j(w^*) - g_j(w_{j,\tau}^t) \right) + 2\eta(1-\sigma_\beta^t) \left(f_j(w^*) - f_j(w_{j,\tau}^t) \right) \\
&\leq 2\eta\sigma_\beta^t \left[g_j(w^*) - g_j(w_t) + \frac{1}{2\alpha} \|w_t - w_{j,\tau}^t\|^2 + \frac{\alpha}{2} G^2 \right] \qquad \text{(by Cvx., Young's, \& G-Lip)} \\
&\quad + 2\eta(1-\sigma_\beta^t) \left[f_j(w^*) - f_j(w_t) + \frac{1}{2\alpha} \|w_t - w_{j,\tau}^t\|^2 + \frac{\alpha}{2} G^2 \right]
\end{aligned}
$$

Putting $Term - B_1$ and $Term - B_2$ back in $Term - B$, we get

$$
\begin{aligned}
\mathbb{E}_{\mathcal{C}_j}[Term - B] \leq \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} \Bigg[& \frac{2\eta}{\alpha} \|w_t - w_{j,\tau}^t\|^2 + 2\eta\alpha G^2 + 2\eta\sigma_\beta^t \left(g_j(w^*) - g_j(w_t) \right) \\
& + 2\eta(1-\sigma_\beta^t) \left(f_j(w^*) - f_j(w_t) \right) \Bigg] \\
\overset{Lemma\ 5}{\leq} & \frac{2}{3\alpha} \eta^3 E^3 G^2 + 2\eta\alpha E G^2 + 2\eta E \sigma_\beta^t \left(g(w^*) - g(w_t) \right) \\
& + 2\eta E (1-\sigma_\beta^t) \left(f(w^*) - f(w_t) \right)
\end{aligned}
$$

Substituting $Term - A$ and $Term - B$ back in the original expression after taking the expectation w.r.t $\mathcal{C}_j$, we get for $\alpha = \eta$

$$
\begin{aligned}
\|w_{t+1} - w^*\|^2 \leq \|w_t - w^*\|^2 + \eta^2 \left(\frac{q_u}{n} + 1 \right) E^2 G^2 + 2\eta^2 E G^2 + \frac{2}{3} \eta^2 E^3 G^2 \\
+ 2\eta E \sigma_\beta^t \left(g(w^*) - g(w_t) \right) + 2\eta E (1-\sigma_\beta^t) \left(f(w^*) - f(w_t) \right)
\end{aligned}
$$

Let $\mathcal{A} = \{t \in [T] | g(w_t) < \epsilon\}$ and $\mathcal{B} = [T] \backslash \mathcal{A} = \{t \in [T] | g(w_t) \geq \epsilon\}$. Note that for all $t \in \mathcal{B}$ it holds that $\sigma_\beta(g(w_t) - \epsilon) = 1$ and $g(w_t) - g(w^*) \geq \epsilon$. Further, for all $t \in \mathcal{A}$ if $\sigma_\beta(g(w_t) - \epsilon) \geq 0$ it holds that $g(w_t) - g(w^*) \geq g(w_t) \geq \epsilon - 1/\beta$. With these observations, using convexity of f and g and decomposing the sum over t according to the definitions of $\mathcal{A}$ and $\mathcal{B}$ and division by T yields,

$$
\begin{aligned}
\frac{D^2}{2\eta E} + \frac{\eta}{2} \left(\frac{q_u}{n} + 1 \right) E G^2 T + \eta G^2 T + \frac{1}{3} \eta E G^2 T \geq & \sum_{t \in \mathcal{A}} \sigma_\beta^t \left(g(w_t) - g(w^*) \right) + \sum_{t \in \mathcal{B}} \left(g(w_t) - g(w^*) \right) \\
& + \sum_{t \in \mathcal{A}} (1 - \sigma_\beta^t) \left(f(w_t) - f(w^*) \right).
\end{aligned}
$$

Now choosing $\eta = \sqrt{\frac{D^2}{2G^2ET\Gamma}}$, where $\Gamma = \frac{1}{2}\left(\frac{q_u}{n}+1\right)E + 1 + \frac{1}{3}E^2$, we get

$$\begin{aligned}\sqrt{\frac{2D^2G^2T\Gamma}{E}} &\geq \sum_{t\in\mathcal{A}} \sigma_\beta^t \left(g(w_t) - g(w^*)\right) + \sum_{t\in\mathcal{A}} (1-\sigma_\beta^t)\left(f(w_t) - f(w^*)\right) + \sum_{t\in\mathcal{B}} \left(g(w_t) - g(w^*)\right) \\ &\geq \sum_{t\in\mathcal{A}} (1-\sigma_\beta^t)\left(f(w_t) - f(w^*)\right) + \epsilon|\mathcal{B}| + \left(\epsilon - \frac{1}{\beta}\right)\sum_{t\in\mathcal{A}} \sigma_\beta^t. \end{aligned} \tag{10}$$

Similar to the previous proofs, we first need to find the smallest ϵ to ensure $\mathcal{A}$ is non-empty. So, to find a lower bound on ϵ, assume $\mathcal{A}$ is empty in eq. (10) and observe that as long as condition $\sqrt{\frac{2D^2G^2\Gamma}{ET}} < \epsilon$ is met, $\mathcal{A}$ is non-empty. We choose to set $\epsilon = 2\sqrt{\frac{2D^2G^2\Gamma}{ET}}$. Now, like before, we consider two cases based on the sign of $\sum_{t\in\mathcal{A}}\left(1-\sigma_\beta^t\right)\left(f(w_t)-f(w^*)\right)$. As before, when the sum is non-positive we are done by the definition of $\mathcal{A}$, which implies $0 < 1 - \sigma_\beta(g(w_t)-\epsilon) \leq 1$ and the convexity of f and g.

Assuming the sum is positive, dividing eq. (10) by $\sum_{t\in\mathcal{A}}\left(1-\sigma_\beta^t\right)$ (which by the definition of $\mathcal{A}$ is strictly positive), using convexity of f, and the definition of $\bar{w}$, we have

$$\begin{aligned} f(\bar{w}) - f(w^*) &\leq \frac{0.5\epsilon T - \epsilon|\mathcal{B}| - (\epsilon - \frac{1}{\beta})\sum_{t\in\mathcal{A}}\sigma_\beta^t}{|\mathcal{A}| - \sum_{t\in\mathcal{A}}\sigma_\beta(g(w_t)-\epsilon)} \\ &= \epsilon + \frac{-0.5\epsilon T + \beta^{-1}\sum_{t\in\mathcal{A}}\sigma_\beta^t}{|\mathcal{A}| - \sum_{t\in\mathcal{A}}\sigma_\beta^t}, \end{aligned}$$

where we used $|\mathcal{B}| = T - |\mathcal{A}|$.

Let us now find a lower bound on β to ensure the second term in the bound is non-positive. Note this is done for simplicity, and as long as the second term is $\mathcal{O}(\epsilon)$, an ϵ-solution can be found. Immediate calculations show the second term in the bound is non-positive when

$$\beta \geq \frac{2\sum_{t\in\mathcal{A}}\sigma_\beta^t}{\epsilon T}.$$

Since $\sum_{t\in\mathcal{A}}\sigma_\beta^t < T$, a sufficient (and highly conservative) condition for all $T \geq 1$ is to set $\beta \geq 2/\epsilon$. Thus, we proved the suboptimality gap result. The feasibility result is immediate given the definition of $\mathcal{A}$ and the convexity of g. ■

A.3. Main Theorem FedSGM

A.3.1. HARD SWITCHING — FULL PARTICIPATION

Theorem 8 (*FedHSGM*) *Consider the problem in eq. (1), under Assumption 1. Define* $D := \|w_0 - w^*\|$ *and*

$$\mathcal{A} = \{t \in [T] \mid g(w_t) \leq \epsilon\}, \qquad \bar{w} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t.$$

Then, if

$$\epsilon = \sqrt{\frac{4D^2G^2E}{T}}, \text{ and } \eta = \sqrt{\frac{D^2}{4G^2E^3T}}$$

it holds that $\mathcal{A}$ *is nonempty,* $\bar{w}$ *is well-defined, and* $\bar{w}$ *is an* ϵ*-solution for* P.

Proof Using Algorithm 1, the update rule for the global model is

$$w_{t+1} = w_t - \eta \cdot \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} \nu_{j,\tau}^t, \tag{11}$$

We analyze the squared distance to the optimal point w^* as follows,

$$\begin{aligned}
\|w_{t+1} - w^*\|^2 &= \left\| w_t - \eta \cdot \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} \nu_{j,\tau}^t - w^* \right\|^2 \\
&= \|w_t - w^*\|^2 + \underbrace{\eta^2 \left\| \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} \nu_{j,\tau}^t \right\|^2}_{Term-A} \underbrace{-2\eta \left\langle w_t - w^*, \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} \nu_{j,\tau}^t \right\rangle}_{Term-B}
\end{aligned}$$

Firstly, we start with upper-bounding $Term - A$.

$$\begin{aligned}
Term - A = \eta^2 \left\| \frac{1}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} \nu_{j,\tau}^t \right\|^2 &\overset{Jensen's}{\leq} \eta^2 \frac{1}{n} \sum_{j=1}^{n} E \sum_{\tau=0}^{E-1} \left\| \nu_{j,\tau}^t \right\|^2 \\
&\overset{G-Lip}{\leq} \eta^2 E^2 G^2
\end{aligned}$$

Now we can use Lemma 4 to bound $Term - B$. So, putting $Term - A$ and $Term - B$ back into the expression we get

$$\begin{aligned}
\|w_{t+1} - w^*\|^2 \leq \|w_t - w^*\|^2 &+ \eta^2 E^2 G^2 + \eta \alpha E G^2 \\
&+ \frac{\eta}{\alpha n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} \|w_t - w_{j,\tau}^t\|^2 \\
&+ \frac{2\eta}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} (f_j(w^*) - f_j(w_{j,\tau}^t)) \mathbb{1}\{t \in \mathcal{A}\} \\
&+ \frac{2\eta}{n} \sum_{j=1}^{n} \sum_{\tau=0}^{E-1} (g_j(w^*) - g_j(w_{j,\tau}^t)) \mathbb{1}\{t \in \mathcal{B}\}.
\end{aligned} \tag{12}$$

Now our goal is to handle the term $f_j(w^*) - f_j(w_{j,\tau}^t)$, so we first rewrite,

$$\begin{aligned}
f_j(w_{j,\tau}^t) &\geq f_j(w_t) + \langle \nabla f_j(w_t), w_{j,\tau}^t - w_t \rangle \quad \text{(by convexity)} \\
\Rightarrow f_j(w^*) - f_j(w_{j,\tau}^t) &\leq f_j(w^*) - f_j(w_t) - \langle \nabla f_j(w_t), w_{j,\tau}^t - w_t \rangle.
\end{aligned}$$

Using Young's inequality with parameter $\alpha > 0$, we get

$$-\langle \nabla f_j(w_t), w_{j,\tau}^t - w_t \rangle \leq \frac{1}{2\alpha} \|w_{j,\tau}^t - w_t\|^2 + \frac{\alpha}{2} \|\nabla f_j(w_t^t)\|^2.$$

Thus, we get,

$$f_j(w^*) - f_j(w_{j,\tau}^t) \le f_j(w^*) - f_j(w_t) + \frac{1}{2\alpha}\|w_t - w_{j,\tau}^t\|^2 + \frac{\alpha}{2}\|\nabla f_j(w_t)\|^2$$
$$\overset{G-Lip}{\le} f_j(w^*) - f_j(w_t) + \frac{1}{2\alpha}\|w_t - w_{j,\tau}^t\|^2 + \frac{\alpha}{2}G^2.$$

Similarly, we can handle the other term with g and get,

$$g_j(w^*) - g_j(w_{j,\tau}^t) \le g_j(w^*) - g_j(w_t) + \frac{1}{2\alpha}\|w_t - w_{j,\tau}^t\|^2 + \frac{\alpha}{2}G^2.$$

So, putting these inequalities back in eq. (12) along with the use of Lemma 5 and eq. (1), we get

$$\|w_{t+1} - w^*\|^2 \le \|w_t - w^*\|^2 + \eta^2 E^2 G^2 + 2\eta\alpha E G^2 + \frac{2}{3\alpha}\eta^3 E^3 G^2$$
$$+ 2\eta E(f(w^*) - f(w_t))\mathbb{1}\{t \in \mathcal{A}\} + 2\eta E(g(w^*) - g(w_t))\mathbb{1}\{t \in \mathcal{B}\}.$$

Now, for $\alpha = \eta$, and rearranging the terms we get

$$(f(w_t) - f(w^*))\mathbb{1}\{t \in \mathcal{A}\} + (g(w_t) - g(w^*))\mathbb{1}\{t \in \mathcal{B}\} \le \frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2\eta E}$$
$$+\frac{\eta}{2}EG^2 + \eta G^2 + \frac{1}{3}\eta E^2 G^2.$$

Defining $D := \|w_0 - w^*\|$ and summing the expression above for $t = 0, 1, \cdots, T-1$ and then dividing by T, we get

$$\frac{1}{T}\sum_{t\in\mathcal{A}}(f(w_t) - f(w^*)) + \frac{1}{T}\sum_{t\in\mathcal{B}}(g(w_t) - g(w^*)) \le \frac{D^2}{2\eta ET} + \frac{\eta}{2}EG^2 + \eta G^2 + \frac{1}{3}\eta E^2 G^2.$$

Now choosing $\eta = \sqrt{\frac{D^2}{2G^2 ET\Gamma}}$, where $\Gamma = \frac{1}{2}E + 1 + \frac{1}{3}E^2$, we get

$$\frac{1}{T}\sum_{t\in\mathcal{A}}(f(w_t) - f(w^*)) + \frac{1}{T}\sum_{t\in\mathcal{B}}(g(w_t) - g(w^*)) \le \sqrt{\frac{2D^2G^2\Gamma}{ET}}.$$

Note that when ϵ is sufficiently large, $\mathcal{A}$ is nonempty. Assuming an empty $\mathcal{A}$, we can find the largest "bad" ϵ:

$$\epsilon_{bad} < \frac{1}{T}\sum_{t\in\mathcal{B}} g(w_t) - g(w^*) \le \sqrt{\frac{2D^2G^2\Gamma}{ET}}$$

Thus, let us set $\epsilon = (N+1)\sqrt{\frac{2D^2G^2\Gamma}{ET}}$ for some $N \ge 0$. With this choice, $\mathcal{A}$ is guaranteed to be nonempty.

Now, we consider two cases. Either $\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) \le 0$ which implies by the convexity of f and g for $\bar{w} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t$ we have

$$f(\bar{w}) - f(w^*) \le 0 < \epsilon, \qquad g(\bar{w}) \le \epsilon. \tag{13}$$

Otherwise, if $\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) > 0$, then

$$\begin{aligned}\sqrt{\frac{2D^2G^2\Gamma}{ET}} &\geq \frac{1}{T}\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) + \frac{1}{T}\sum_{t\in\mathcal{B}} g(w_t) - g(w^*) \\ &> \frac{1}{T}\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) + \frac{1}{T}\sum_{t\in\mathcal{B}} \epsilon \\ &= \frac{|\mathcal{A}|}{T}\frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} f(w_t) - f(w^*) + (1 - \frac{|\mathcal{A}|}{T})\epsilon \\ &\geq \frac{|\mathcal{A}|}{T}\big(f(\bar{w}) - f(w^*)\big) + (1 - \frac{|\mathcal{A}|}{T})\epsilon.\end{aligned} \tag{14}$$

By rearranging

$$\frac{|\mathcal{A}|}{T}\Big(f(\bar{w}) - f(w^*) - \epsilon\Big) < \sqrt{\frac{2D^2G^2\Gamma}{ET}} - \epsilon \leq -N\sqrt{\frac{2D^2G^2\Gamma}{ET}}, \tag{15}$$

Implying $f(\bar{w}) - f(w^*) < \epsilon$ and further by convexity of g for $\bar{w} = \frac{1}{|\mathcal{A}|}\sum_{t\in\mathcal{A}} w_t$, we also have $g(\bar{w}) \leq \epsilon$. ■

A.3.2. SOFT SWITCHING — FULL PARTICIPATION

Theorem 9 (*FedSSGM*) *Consider the problem in (1) and Algorithm 1, under Assumption 1. Define* $D := \|w_0 - w^*\|$ *and*

$$\mathcal{A} = \{t \in [T] \mid g(w_t) < \epsilon\}, \qquad \bar{w} = \sum_{t\in\mathcal{A}} \alpha_t w_t,$$

where

$$\alpha_t = \frac{1 - \sigma_\beta(g(w_t) - \epsilon)}{\sum_{t\in\mathcal{A}}[1 - \sigma_\beta(g(w_t) - \epsilon)]}. \tag{16}$$

Then, if

$$\epsilon = \sqrt{\frac{4D^2G^2E}{T}}, \qquad \eta = \sqrt{\frac{D^2}{4G^2E^3T}}, \quad \textit{and} \quad \beta = \frac{2}{\epsilon}$$

it holds that $\mathcal{A}$ is nonempty, $\bar{w}$ is well-defined, and $\bar{w}$ is an ϵ-solution for P.

Proof Using Algorithm 1, the update rule for the global model is

$$\begin{aligned}&w_{t+1} = w_t - \eta \cdot \frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1} \nu_{j,\tau}^t, \\ &\text{where, } \nu_{j,\tau}^t = \sigma_\beta(g(w_t) - \epsilon)\nabla g_j(w_{j,\tau}^t) + (1 - \sigma_\beta(g(w_t) - \epsilon))\nabla f_j(w_{j,\tau}^t)\end{aligned}$$

We analyze the squared distance to the optimal point w^* as follows,

$$\begin{aligned}\|w_{t+1}-w^*\|^2 &= \left\|w_t-\eta\cdot\frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t-w^*\right\|^2\\ &= \|w_t-w^*\|^2+\underbrace{\eta^2\left\|\frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t\right\|^2}_{Term-A}\underbrace{-2\eta\left\langle w_t-w^*,\frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t\right\rangle}_{Term-B}\end{aligned}\tag{17}$$

Firstly, we start with upper-bounding $Term-A$.

$$\begin{aligned}Term-A=\eta^2\left\|\frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t\right\|^2 &\overset{Jensen's}{\leq} \eta^2\frac{1}{n}\sum_{j=1}^{n}E\sum_{\tau=0}^{E-1}\left\|\nu_{j,\tau}^t\right\|^2\\ &\overset{G-Lip\ \&\ \sigma_\beta(\cdot)\in[0,1]}{\leq} \eta^2E^2G^2\end{aligned}$$

Now we aim to upper bound $Term-B$.

$$\begin{aligned}Term-B&=-2\eta\left\langle w_t-w^*,\frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\nu_{j,\tau}^t\right\rangle\\ &=\frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}-2\eta\left\langle w_t-w^*,\nu_{j,\tau}^t\right\rangle\\ &=\frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\left[\underbrace{-2\eta\left\langle w_t-w_{j,\tau}^t,\nu_{j,\tau}^t\right\rangle}_{Term-B_1}\underbrace{-2\eta\left\langle w_{j,\tau}^t-w^*,\nu_{j,\tau}^t\right\rangle}_{Term-B_2}\right]\end{aligned}$$

Now we define $\sigma_\beta^t=\sigma_\beta(g(w_t)-\epsilon)$ before we start upper bounding $Term-B_1$ for any $\alpha>0$.

$$\begin{aligned}&Term-B_1=-2\eta\left\langle w_t-w_{j,\tau}^t,\nu_{j,\tau}^t\right\rangle\\ &=-2\eta\sigma_\beta^t\left\langle w_t-w_{j,\tau}^t,g_j(w_{j,\tau}^t)\right\rangle-2\eta(1-\sigma_\beta^t)\left\langle w_t-w_{j,\tau}^t,f_j(w_{j,\tau}^t)\right\rangle\\ &\overset{Young's}{\leq}\sigma_\beta^t\left[\frac{\eta}{\alpha}\|w_t-w_{j,\tau}^t\|^2+\eta\alpha\|\nabla g_j(w_{j,\tau}^t)\|^2\right]+(1-\sigma_\beta^t)\left[\frac{\eta}{\alpha}\|w_t-w_{j,\tau}^t\|^2+\eta\alpha\|\nabla f_j(w_{j,\tau}^t)\|^2\right]\\ &\overset{G-Lip}{\leq}\frac{\eta}{\alpha}\|w_t-w_{j,\tau}^t\|^2+\eta\alpha G^2\end{aligned}$$

Similarly, we start to upper bound $Term-B_2$ as well.

$$\begin{aligned}Term-B_2&=-2\eta\left\langle w_{j,\tau}^t-w^*,\nu_{j,\tau}^t\right\rangle\\ &=-2\eta\sigma_\beta^t\left\langle w_{j,\tau}^t-w^*,g_j(w_{j,\tau}^t)\right\rangle-2\eta(1-\sigma_\beta^t)\left\langle w_{j,\tau}^t-w^*,f_j(w_{j,\tau}^t)\right\rangle\\ &\overset{Cvx.}{\leq}2\eta\sigma_\beta^t\left(g_j(w^*)-g_j(w_{j,\tau}^t)\right)+2\eta(1-\sigma_\beta^t)\left(f_j(w^*)-f_j(w_{j,\tau}^t)\right)\\ &\leq 2\eta\sigma_\beta^t\left[g_j(w^*)-g_j(w_t)+\frac{1}{2\alpha}\|w_t-w_{j,\tau}^t\|^2+\frac{\alpha}{2}G^2\right]\qquad\text{(by Cvx., Young's, \& G-Lip)}\\ &\quad+2\eta(1-\sigma_\beta^t)\left[f_j(w^*)-f_j(w_t)+\frac{1}{2\alpha}\|w_t-w_{j,\tau}^t\|^2+\frac{\alpha}{2}G^2\right]\end{aligned}$$

Putting $Term - B_1$ and $Term - B_2$ back in $Term - B$, we get

$$\begin{aligned} Term - B \leq \frac{1}{n}\sum_{j=1}^{n}\sum_{\tau=0}^{E-1}\Big[\frac{2\eta}{\alpha}\|w_t - w_{j,\tau}^t\|^2 + 2\eta\alpha G^2 + 2\eta\sigma_\beta^t\left(g_j(w^*) - g_j(w_t)\right) \\ +2\eta(1-\sigma_\beta^t)\left(f_j(w^*) - f_j(w_t)\right)\Big] \\ \overset{Lemma\ 5}{\leq} \frac{2}{3\alpha}\eta^3E^3G^2 + 2\eta\alpha EG^2 + 2\eta E\sigma_\beta^t\left(g(w^*) - g(w_t)\right) \\ + 2\eta E(1-\sigma_\beta^t)\left(f(w^*) - f(w_t)\right) \end{aligned}$$

Substituting $Term - A$ and $Term - B$ back in eq. (17), we get for $\alpha = \eta$

$$\begin{aligned} \|w_{t+1} - w^*\|^2 \leq \|w_t - w^*\|^2 + \eta^2E^2G^2 + 2\eta^2EG^2 + \frac{2}{3}\eta^2E^3G^2 + 2\eta E\sigma_\beta^t\left(g(w^*) - g(w_t)\right) \\ + 2\eta E(1-\sigma_\beta^t)\left(f(w^*) - f(w_t)\right) \end{aligned}$$

Let $\mathcal{A} = \{t \in [T] | g(w_t) < \epsilon\}$ and $\mathcal{B} = [T]\backslash\mathcal{A} = \{t \in [T] | g(w_t) \geq \epsilon\}$. Note that for all $t \in \mathcal{B}$ it holds that $\sigma_\beta(g(w_t) - \epsilon) = 1$ and $g(w_t) - g(w^*) \geq \epsilon$. Further, for all $t \in \mathcal{A}$ if $\sigma_\beta(g(w_t) - \epsilon) \geq 0$ it holds that $g(w_t) - g(w^*) \geq g(w_t) \geq \epsilon - 1/\beta$. With these observations, using convexity of f and g and decomposing the sum over t according to the definitions of $\mathcal{A}$ and $\mathcal{B}$ and division by T yields,

$$\begin{aligned} \frac{D^2}{2\eta E} + \frac{1}{2}\eta EG^2T + \eta G^2T + \frac{1}{3}\eta E^2G^2T \geq \sum_{t\in\mathcal{A}}\sigma_\beta^t\left(g(w_t) - g(w^*)\right) + \sum_{t\in\mathcal{A}}(1-\sigma_\beta^t)\left(f(w_t) - f(w^*)\right) \\ + \sum_{t\in\mathcal{B}}\left(g(w_t) - g(w^*)\right). \end{aligned}$$

Now choosing $\eta = \sqrt{\frac{D^2}{2G^2ET\Gamma}}$, where $\Gamma = \frac{1}{2}E + 1 + \frac{1}{3}E^2$, we get

$$\begin{aligned} \sqrt{\frac{2D^2G^2T\Gamma}{E}} &\geq \sum_{t\in\mathcal{A}}\sigma_\beta^t\left(g(w_t) - g(w^*)\right) + \sum_{t\in\mathcal{A}}(1-\sigma_\beta^t)\left(f(w_t) - f(w^*)\right) + \sum_{t\in\mathcal{B}}\left(g(w_t) - g(w^*)\right) \\ &\geq \sum_{t\in\mathcal{A}}(1-\sigma_\beta^t)\left(f(w_t) - f(w^*)\right) + \epsilon|\mathcal{B}| + \left(\epsilon - \frac{1}{\beta}\right)\sum_{t\in\mathcal{A}}\sigma_\beta^t. \end{aligned} \tag{18}$$

Similar to the previous proofs, we first need to find the smallest ϵ to ensure $\mathcal{A}$ is non-empty. So, to find a lower bound on ϵ, assume $\mathcal{A}$ is empty in eq. (18) and observe that as long as condition $\sqrt{\frac{2D^2G^2\Gamma}{ET}} < \epsilon$ is met, $\mathcal{A}$ is non-empty. We choose to set $\epsilon = 2\sqrt{\frac{2D^2G^2\Gamma}{ET}}$.

Now, like before, we consider two cases based on the sign of $\sum_{t\in\mathcal{A}}\left(1-\sigma_\beta^t\right)\left(f(w_t) - f(w^*)\right)$. As before, when the sum is non-positive we are done by the definition of $\mathcal{A}$, which implies $0 < 1 - \sigma_\beta(g(w_t) - \epsilon) \leq 1$ and the convexity of f and g.

Assuming the sum is positive, dividing eq. (18) by $\sum_{t\in\mathcal{A}}\left(1-\sigma_\beta^t\right)$ (which by the definition of $\mathcal{A}$ is strictly positive), using convexity of f, and the definition of $\bar{w}$, we have

$$\begin{aligned} f(\bar{w}) - f(w^*) &\leq \frac{0.5\epsilon T - \epsilon|\mathcal{B}| - (\epsilon - \frac{1}{\beta})\sum_{t\in\mathcal{A}}\sigma_\beta^t}{|\mathcal{A}| - \sum_{t\in\mathcal{A}}\sigma_\beta(g(w_t) - \epsilon)} \\ &= \epsilon + \frac{-0.5\epsilon T + \beta^{-1}\sum_{t\in\mathcal{A}}\sigma_\beta^t}{|\mathcal{A}| - \sum_{t\in\mathcal{A}}\sigma_\beta^t}, \end{aligned}$$

where we used $|\mathcal{B}| = T - |\mathcal{A}|$.

Let us now find a lower bound on β to ensure the second term in the bound is non-positive. Note this is done for simplicity, and as long as the second term is $\mathcal{O}(\epsilon)$, an ϵ-solution can be found. Immediate calculations show the second term in the bound is non-positive when

$$\beta \geq \frac{2\sum_{t\in\mathcal{A}} \sigma_\beta^t}{\epsilon T}.$$

Since $\sum_{t\in\mathcal{A}} \sigma_\beta^t < T$, a sufficient (and highly conservative) condition for all $T \geq 1$ is to set $\beta \geq 2/\epsilon$. Thus, we proved the suboptimality gap result. The feasibility result is immediate given the definition of $\mathcal{A}$ and the convexity of g. ■