Multi-Timescale Gradient Sliding for Distributed Optimization

Junhui Zhang JUNHUIZ@MIT.EDU
Patrick Jaillet JAILLET@MIT.EDU
MIT, USA

Abstract

We propose two first-order methods for convex, non-smooth, distributed optimization problems, hereafter called Multi-Timescale Gradient Sliding (MT-GS) and its accelerated variant (AMT-GS). Our MT-GS and AMT-GS can take advantage of similarities between (local) objectives to reduce the communication rounds, are flexible so that different subsets (of agents) can communicate at different, user-picked rates, and are fully deterministic. These three desirable features are achieved through a block-decomposable primal-dual formulation, and a multi-timescale variant of the sliding method introduced in [23, 24], where different dual blocks are updated at potentially different rates.

To find an ϵ-suboptimal solution, the complexities of our algorithms achieve optimal dependency on ϵ: MT-GS needs $O(\overline{r}A/\epsilon)$ communication rounds and $O(\overline{r}/\epsilon^2)$ subgradient steps for Lipchitz objectives, and AMT-GS needs $O(\overline{r}A/\sqrt{\epsilon\mu})$ communication rounds and $O(\overline{r}/(\epsilon\mu))$ subgradient steps if the objectives are also μ-strongly convex. Here, $\overline{r}$ measures the "average rate of updates" for dual blocks, and A measures similarities between (subgradients of) local functions. In addition, the linear dependency of communication rounds on A is optimal [3], thereby providing a positive answer to the open question whether such dependency is achievable for non-smooth objectives [3].

1. Introduction

Distributed optimization is a branch of optimization, where multiple agents, each having access to only partial information about the (global) objective, work together to solve the global problem. As an example, in distributed empirical risk minimization for machine learning, the global objective function is the sum of local loss functions, each depending on the local dataset which is only available to one agent [2–4, 9, 21]. Examples of other application include power system control [27, 29], multi-robot system control [10, 17, 29, 35], and signal processing [8, 25, 32], to name a few.

In this work, we study the following convex, non-smooth, distributed optimization problems:

$$\min_{x\in\mathcal{X}} \sum_{v\in V} f_v(x) \tag{$\mathcal{P}$}$$

where $V=[m]$ represents m agents, $\mathcal{X}\subset\mathbb{R}^d$ is a nonempty, closed convex set, and $f_v:\mathcal{X}\to\mathbb{R}$ is a convex and possibly non-smooth objective function such that for some $M,\mu\geq 0$, for all $v\in V$,

$$\frac{\mu}{2}\|x-y\|^2 \leq f_v(x)-f_v(y)-\langle f_v'(y), x-y\rangle \leq M\|x-y\|, \quad \forall x,y\in\mathcal{X}, \tag{1}$$

where $f_v':\mathcal{X}\to\mathbb{R}^d$ is a subgradient oracle, i.e. $f_v'(x)\in\partial f_v(x)$ for all $x\in\mathcal{X}$, and f_v' is only available to agent v. For instance, when $\|f_v'\|_*\leq M_f$, $M=2M_f$ holds.

We assume that agent v maintains and updates x_v, a local version of the decision variable x. In this setting, due to the lack of "global views" from agents' (local) perspectives, information aggregation – such as communication and averaging – is necessary for agents to reach consensus and approximately solve ($\mathcal{P}$).

We target settings where within and between different subsets of the local objectives, the scales of the *function similarities* and the *costs of information aggregation* could be (vastly) different. For instance, in distributed empirical risk minimization, local loss functions could inherit potential similarities in local datasets, which helps reduce the communication round needed [2, 3, 22, 34, 38, 46]; the costs of communication could depend on factors such as the distance between agents, methods of communication, and amounts of data sent [6, 8, 33, 40, 41].

The heterogeneity in the function similarities and the communication costs makes it desirable to have more refined control over the numbers of communication rounds among different subsets of agents. In addition, sometimes stochastic algorithms are impractical or inefficient due to factors such as unpredictability, random memory access [37], and sampling overhead [14]. For these reasons, we aim at designing *deterministic* algorithms which allow users to pick *different numbers of communication rounds among different subsets of agents*.

Toward this end, we make contributions to both the *problem formulation* (Section 2, Appendix B) and the *algorithm design* (Section 3, Appendices C and D). Moreover, our algorithms achieve linear, and thus optimal dependence on *similarities* between (subgradients of) local functions (Section 4). Numerical experiments for the support vector machine problem with regularized hinge losses confirm the effectiveness of our algorithms and demonstrate the above dependence (Section 5, Appendix E).

2. Formulation: generalized, block-decomposable penalties for consensus constraints.

We propose relaxing the consensus constraints through *generic convex penalty functions*, generalizing the *characteristic function penalty* used in previous works:

$$\min_{X=(x_v)_{v\in V}\in\overline{\mathcal{X}}} F(X)+\sum_{s=1}^{S} R_s(K_sX), \quad F(X):=\sum_{v\in V} f_v(x_v). \tag{$\mathcal{P}_r$}$$

Above, $\overline{\mathcal{X}}=\mathcal{X}^V\subset\mathbb{R}^{md}$, $K_s:\mathbb{R}^{md}\to\mathbb{R}^{n_s}$ is a linear operator which imposes "consensus constraints" within a subset of the agents, and $\cap_{s=1}^{S}\ker(K_s)$ is the subspace in $\mathbb{R}^{md}$ where $\{x_v\}_{v\in V}$ does not violate the consensus constraints. $R_s:\mathbb{R}^{n_s}\to\overline{\mathbb{R}}$ is a proper, convex, and lower-semi-continuous regularization term, penalizing the deviation of K_sX from $\mathbf{0}$. For instance, R_s could be the characteristic function of $\{\mathbf{0}\}$, or any scaled norm $R_s(y_s)=\lambda\|y_s\|$.

We further consider the saddle point reformulation of ($\mathcal{P}_r$):

$$\min_{X\in\overline{\mathcal{X}}}\max_{Y\in\mathbb{R}^n}\sum_{s=1}^{S}\langle K_sX,y_s\rangle+F(X)-\sum_{s=1}^{S}R_s^*(y_s), \tag{$\mathcal{P}_s$}$$

where R_s^* is the Fenchel conjugate of R_s.

Requirements on the penalties. We propose a set of conditions on the growth rates of penalties (Lemma 3 and Corollary 5), under which the duality gap for ($\mathcal{P}_s$) provides upper bounds on

objective value suboptimality and consensus constraint violation. This relates solution qualities for the penalized problems (($\mathcal{P}_r$) and ($\mathcal{P}_s$)) back to the original distributed optimization problem ($\mathcal{P}$).

Communication protocol. We assume that the objective functions $\{f_v\}_{v\in V}$ are distributed among m *primal agents*: for each $v \in V$, $\mathsf{Agent}(x_v)$ has access to f_v', the first order oracle for f_v, and is responsible for updating the variable x_v. In addition, we assume that there are S *dual agents*: for each $s \in [S]$, $\mathsf{Agent}(y_s)$ is responsible for updating the variable y_s.

We assume that for any pair $(s, v) \in [S] \times V$ such that $K_{s,v} \neq \mathbf{0}$, $\mathsf{Agent}(x_v)$ and $\mathsf{Agent}(y_s)$ can communicate (in both directions). For instance, all agents might be nodes in a connected graph with vertices $[S] \cup V$ (representing S dual agents and m primal agents), and communication can be realized through edges (directly) or through paths (i.e. with the help of intermediate agents). In particular, since the graph is connected, any pair can communicate, but the resources consumed and/or time taken by communication between different pairs could be (significantly) different. In Figure 1, we present examples of a decentralized setting and a hierarchical setting.

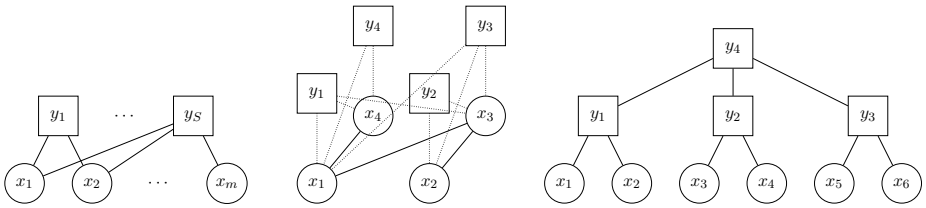

Figure 1: Left: abstract setting with m primal agents and S dual agents. Middle: realization in the decentralized setting, where $S = m = 4$, $\mathsf{Agent}(x_s) = \mathsf{Agent}(y_s)$, and the underlying graph is $(V, E = \{\{1, 3\}, \{1, 4\}, \{2, 3\}\})$. Right: realization in the hierarchical setting.

3. Algorithm: (accelerated) multi-timescale gradient sliding.

When there is only one block ($S = 1$) and R is the characteristic function of $\{\mathbf{0}\}$, ($\mathcal{P}_s$) can be solved through the decentralized communication sliding (DCS) [24], a communication efficient variant of the classical primal-dual hybrid gradient algorithm. As a recap, at iteration k, DCS performs the following updates:

$$\begin{aligned}
\widetilde{X}^k &= X^k + \alpha_k(X^k - X^{k-1}) \\
Y^k &= \operatorname*{argmin}_{Y\in\mathbb{R}^n} R^*(Y) + \langle -K\widetilde{X}^k, Y\rangle + \tau_k D_{w^y}(Y, Y^{k-1}) \\
X^k &\approx \operatorname*{argmin}_{X\in\overline{\mathcal{X}}} F(X) + \langle K^*Y^k, X\rangle + \eta_k D_{w^X}(X, X^{k-1})
\end{aligned}$$

where D_{w^y} and D_{w^X} are the Bregman divergences generated by the distance generating functions w^y and w^X, respectively. Further, the proximal update for X^k is solved inexactly through the Communication-Sliding (CS) procedure (i.e. T_k steps of mirror descent) locally by each primal agent. DCS is communication efficient, since only computing $K\widetilde{X}^k$ and K^*Y^k requires communication between primal and dual agents.

For the formulation ($\mathcal{P}_s$), taking advantage of the block-decomposable structure, we propose MT-GS and AMT-GS, where the dual blocks are updated at different rates. Below, we provide an overview of the updating rules.

Dual updates. The dual blocks are updated at potentially different, user-chosen frequencies: we associate each dual y_s with a rate $r_s \in \mathbb{N}$ and a local time $i_s = 0, 1, \ldots, N_s - 1$, such that $N + 1 =$

	condition (1)	communication round	subgradient oracle
MT-GS (Corollary 9, 10)	$\mu \geq 0$	$O(\frac{\overline{r}A\sqrt{D^X}}{\epsilon})$	$O(\frac{\overline{r}mM^2D^X}{\epsilon^2})$
AMT-GS (Corollary 15, 16)	$\mu > 0$	$O(\frac{\overline{r}A}{\sqrt{\mu\epsilon}})$	$O(\frac{\overline{r}\sqrt{m}M^2}{\mu\epsilon})$

Table 1: Communication rounds and subgradient oracles needed to find an ϵ suboptimal solution to ($\mathcal{P}$). $\overline{r} = \sum_{s=1}^{S} r_s\rho_s$, A measures function similarities, $D^X = D_{w^X}(X^*, X^{init})$. Subgradient oracle for AMT-GS assumes $S = 1$; for general S, see Section D.2.

r_sN_s. For $k = 0, 1, \ldots, N$, y_s is updated only for $k = r_si_s$ for some $i_s \in \{0, 1, \ldots, N_s - 1\}$, using the following rules:

$$\widetilde{x}_{s,v}^{i_s} = \alpha_{s,i_s}\left(\sum_{k'=r_si_s-r_s}^{r_si_s-1}\theta_{k'}(\widehat{x}_v^{k'} - x_v^{k'-r_s})\right) + \sum_{k'=r_si_s-r_s}^{r_si_s-1}\theta_{k'+r_s}x_v^{k'}, \quad v \in V,$$

$$y_s^{i_s} = \operatorname*{argmin}_{y_s\in\mathbb{R}^{n_s}}\langle -\frac{1}{\sum_{k'=r_si_s}^{r_si_s+r_s-1}\theta_{k'}}\sum_{v\in V}K_{s,v}\widetilde{x}_{s,v}^{i_s}, y_s\rangle + R_s^*(y_s) + \tau_{s,i_s}D_{w_s^y}(y_s, y_s^{i_s-1}).$$

Primal updates. At each iteration, our algorithms apply a generalized communication sliding procedure [24] – consisting of multiple subgradient (mirror descent) steps – to approximate the proximal updates of the primal variables:

$$x_v^k, \widehat{x}_v^k \approx \operatorname*{argmin}_{x\in\mathcal{X}} f_v(x) + \langle\sum_{s=1}^{S}K_{s,v}^*\overline{y}_s^k, x\rangle + \sum_{s=1}^{S}\eta_{k,s}D_{w^x}(x, x_v^{k-r_s}), \quad \forall v \in V.$$

Final outputs. Denoting $\widehat{Z}^k = (\widehat{X}^k, \overline{Y}^k)$, then the output is $Z^N = (\sum_{k=0}^{N}\theta_k)^{-1}\sum_{k=0}^{N}\theta_k\widehat{Z}^k$.

Performance. Denoting the (weighted) average of these frequencies as $\overline{r}$, we show that the duality gap of MT-GS converges at the rate of $O(\overline{r}/k)$ for Lipschitz convex objectives (Lemma 7, Theorem 8), and the duality gap of AMT-GS converges at the rate of $O((\overline{r})^2/\mu k^2)$ for μ-strongly convex objectives (Lemma 11, Theorem 12). When specialized in the setting of distributed optimization, to find an ϵ-suboptimal solution, the communication round and subgradient oracle complexities of our algorithms (Table 1) have optimal dependency on ϵ.

To the best of our knowledge, MT-GS and AMT-GS are the first *deterministic* algorithms for saddle point problems with block-decomposable duals that allow *different number of updates for different dual blocks*. This extra flexibility in choosing the update frequencies allows one to design more communication-efficient algorithms, especially when the cost of updating different duals and/or the domain sizes of different duals are different (Section C.5).

4. Beyond Lipschitz constants: function similarity based complexity.

We formalize the notion of function similarity for non-smooth objectives (Definition 4) studied in [3]. For example, in the simplest setting where there is only one block ($S = 1$), our measure of similarity is

$$\sup_{x\in\mathcal{X}}\sum_{v\in V}\|f_v'(x) - \frac{1}{m}\sum_{v'\in V}f_{v'}'(x)\|^2,$$

where f_v' is a subgradient of f_v. We show that with proper choices of the penalties, the communication round complexities of MT-GS and AMT-GS have linear, and thus *optimal*, dependency on

function similarities (Sections C.5 and D.2). This provides positive answers to the open question whether the theoretical communication round lower bounds proposed in [3] can be attained.

5. Numerical experiments: support vector machine

We consider the Support Vector Machine (SVM) problem with hinge loss and additional regularization. More precisely, each primal $\mathsf{Agent}(x_v)$ is given m_s pairs $\{(b_v^l, y_v^l)\}_{l\in[m_s]}$ such that $b_v^l \in \mathbb{R}^d$ is a feature vector satisfying $\|b_v^l\| = 1$, and $y_v^l \in \{\pm 1\}$ is the label. The goal of SVM is to find a weight vector $x \in \mathbb{R}^d$ such that the linear classifier $b \to \mathsf{sign}(\langle b, x\rangle)$ agrees with most pairs (b_v^l, y_v^l) in the dataset. To achieve this, one common approach is to solve the following (regularized) hinge loss minimization problem (in a distributed fashion)[15, 24]:

$$\min_{x\in\mathcal{X}} \sum_{v\in V} f_v(x), \quad f_v(x) = \frac{1}{m_s}\sum_{l=1}^{m_s}[1 - y_v^l\langle b_v^l, x\rangle]_+ + \frac{\mu}{2}\|x\|^2,\ v \in V, \tag{2}$$

where we take $d = 50$, $\mathcal{X} = \{x \in \mathbb{R}^d | \|x\| \le 5\}$, and $[c]_+ = \max(0, c)$ for all $c \in \mathbb{R}$. When $\mu = 0$, (2) is the classical hinge loss minimization problem, and when $\mu > 0$, the local objectives are μ-strongly convex.

Below, we show how the objective values depend on the iteration k, the mean updating rate $\overline{r}$, and the function similarities. For detailed setup, see Appendix E.

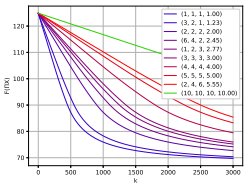

Figure 2: MT-GS ($\mu = 0$). $F(\Pi\underline{X}^k)$ and iteration k; legends: $(r_1, r_2, r_3, \overline{r})$, line colors: $\overline{r}$.

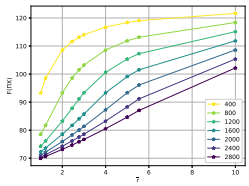

Figure 3: MT-GS ($\mu = 0$). $F(\Pi\underline{X}^k)$ and the mean updating rate $\overline{r}$; legends and line colors: k.

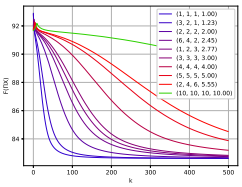

Figure 4: AMT-GS ($\mu = 0.01$). $F(\Pi\underline{X}^k)$ and iteration k; legends: $(r_1, r_2, r_3, \overline{r})$, line colors: $\overline{r}$.

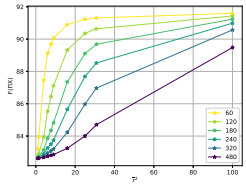

Figure 5: AMT-GS ($\mu = 0.01$). $F(\Pi\underline{X}^k)$ and the mean updating rate $\overline{r}^2$; legends and line colors represent k.

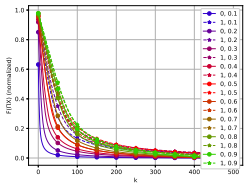

Figure 6: MT-GS ($\mu = 0$). Normalized $F(\Pi\underline{X}^k)$ and iteration k; legends: (setup type, γ), line colors: γ.

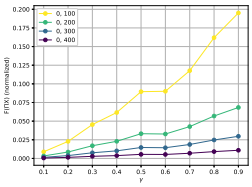

Figure 7: MT-GS ($\mu = 0$). Normalized $F(\Pi\underline{X}^k)$ and function similarities γ; legends: (setup type, k), line colors: k.

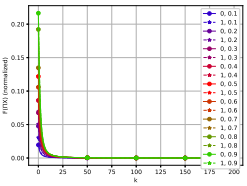

Figure 8: AMT-GS ($\mu = 0.01$). Normalized $F(\Pi\underline{X}^k)$ and iteration k; legends: (setup type, γ), line colors: γ.

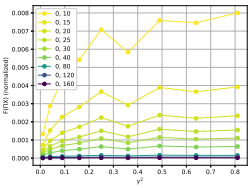

Figure 9: AMT-GS ($\mu = 0.01$). Normalized $F(\Pi\underline{X}^k)$ and function similarities γ^2; legends: (setup type, k), line colors: k.

References

[1] Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. On the convergence of stochastic primal-dual hybrid gradient. *SIAM Journal on Optimization*, 32(2):1288–1318, 2022.

[2] Zeyuan Allen-Zhu, Yang Yuan, and Karthik Sridharan. Exploiting the structure: stochastic gradient methods using raw clusters. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 1650–1658, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[3] Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1756–1764, Cambridge, MA, USA, 2015. MIT Press.

[4] By Mahmoud Assran, Arda Aytekin, Hamid Reza Feyzmahdavian, Mikael Johansson, and Michael G. Rabbat. Advances in asynchronous parallel and distributed optimization. *Proceedings of the IEEE*, 108(11):2013–2031, 2020.

[5] N. S. Aybat, Z. Wang, T. Lin, and S. Ma. Distributed linearized alternating direction method of multipliers for composite convex consensus optimization. *IEEE Transactions on Automatic Control*, 63(1):5–20, 2018.

[6] Albert S. Berahas, Raghu Bollapragada, Nitish Shirish Keskar, and Ermin Wei. Balancing communication and computation in distributed optimization. *IEEE Transactions on Automatic Control*, 64(8):3141–3155, 2019.

[7] Dimitri P. Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, October 2011. ISSN 1436-4646.

[8] Dimitri P. Bertsekas and John N. Tsitsiklis. *Parallel and distributed computation: numerical methods*. Prentice-Hall, Inc., USA, 1989. ISBN 0136487009.

[9] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. 2011.

[10] Francesco Bullo, Jorge Cortés, and Sonia Martínez. *Distributed Control of Robotic Networks: A Mathematical Approach to Motion Coordination Algorithms*. 2009.

[11] 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, May 2011. ISSN 1573-7683.

[12] Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, September 2016. ISSN 1436-4646.

[13] Antonin Chambolle, Matthias J. Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[14] Dominik Csiba and Peter Richtárik. Importance sampling for minibatches. *Journal of Machine Learning Research*, 19(27):1–21, 2018.

[15] John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2012.

[16] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[17] Hassan Jaleel and Jeff S. Shamma. Distributed optimization for robot networks: From real-time convex optimization to game-theoretic self-organization. *Proceedings of the IEEE*, 108 (11):1953–1967, 2020.

[18] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143, virtual conference, 13–18 Jul 2020. PMLR.

[19] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5381–5393, virtual conference, 13–18 Jul 2020. PMLR.

[20] Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence, 2016.

[21] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[22] Dmitry Kovalev, Aleksandr Beznosikov, Ekaterina Borodich, Alexander Gasnikov, and Gesualdo Scutari. Optimal gradient sliding and its application to distributed optimization under similarity. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

[23] Guanghui Lan. Gradient sliding for composite optimization. *Mathematical Programming*, 159 (1):201–235, September 2016. ISSN 1436-4646.

[24] Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 180(1):237–284, March 2020. ISSN 1436-4646.

[25] Dan Li, K.D. Wong, Yu Hen Hu, and A.M. Sayeed. Detection, classification, and tracking of targets. *IEEE Signal Processing Magazine*, 19(2):17–29, 2002.

[26] Carl D. Meyer. Generalized inversion of modified matrices. *SIAM Journal on Applied Mathematics*, 24(3):315–323, 1973. ISSN 00361399.

[27] Daniel K. Molzahn, Florian Dörfler, Henrik Sandberg, Steven H. Low, Sambuddha Chakrabarti, Ross Baldick, and Javad Lavaei. A survey of distributed optimization and control algorithms for electric power systems. *IEEE Transactions on Smart Grid*, 8(6):2941–2962, 2017.

[28] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[29] Angelia Nedić and Ji Liu. Distributed optimization for control. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(Volume 1, 2018):77–103, 2018. ISSN 2573-5144.

[30] Angelia Nedić, Alex Olshevsky, and Michael G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

[31] Roger J. B. Wets R. Tyrrell Rockafellar. *Variational Analysis*. Springer Science & Business Media, Berlin, Germany, 2009.

[32] M. Rabbat and R. Nowak. Distributed optimization in sensor networks. In *Third International Symposium on Information Processing in Sensor Networks, 2004. IPSN 2004*, pages 20–27, 2004.

[33] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20(159):1–31, 2019.

[34] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1000–1008, Bejing, China, 22–24 Jun 2014. PMLR.

[35] Ola Shorinwa, Trevor Halsted, Javier Yu, and Mac Schwager. Distributed optimization methods for multi-robot systems: Part 1—a tutorial [tutorial]. *IEEE Robotics & Automation Magazine*, 31(3):121–138, 2024.

[36] Sebastian Urban Stich. Local sgd converges fast and communicates little. In *ICLR 2019-International Conference on Learning Representations*, 2019.

[37] Tao Sun, Robert Hannah, and Wotao Yin. Asynchronous coordinate descent under more realistic assumption. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6183–6191, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[38] Ye Tian, Gesualdo Scutari, Tianyu Cao, and Alexander Gasnikov. Acceleration in distributed optimization under similarity. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5721–5756, virtual conference, 28–30 Mar 2022. PMLR.

[39] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9):803–812, 1986.

[40] Alexander Tyurin and Peter Richtárik. On the optimal time complexities in decentralized stochastic asynchronous optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 122652–122705, Red Hook,NY, 2024. Curran Associates, Inc.

[41] Santosh S. Vempala, Ruosong Wang, and David P. Woodruff. *The Communication Complexity of Optimization*, pages 1733–1752. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2020.

[42] Ermin Wei and Asuman Ozdaglar. On the o(1/k) convergence of asynchronous distributed alternating direction method of multipliers, 2013.

[43] Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6281–6292, Red Hook, NY, 2020. Curran Associates, Inc.

[44] Stephen J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, June 2015. ISSN 1436-4646.

[45] Yuchen Zhang and Xiao Lin. Stochastic primal-dual coordinate method for regularized empirical risk minimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 353–361, Lille, France, 07–09 Jul 2015. PMLR.

[46] Yuchen Zhang and Xiao Lin. Disco: Distributed optimization for self-concordant empirical loss. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 362–370, Lille, France, 07–09 Jul 2015. PMLR.

[47] Martin A. Zinkevich, Markus Weimer, Alex Smola, and Lihong Li. Parallelized stochastic gradient descent. In *Proceedings of the 24th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, page 2595–2603, Red Hook, NY, USA, 2010. Curran Associates Inc.

Appendix A. Related works

Non-smooth distributed optimization. Since the seminal works [8, 39], numerous algorithms have been proposed for non-smooth distributed optimization under various settings, and we refer readers to surveys such as [4, 20, 30]. For the function class of Lipschitz, non-smooth, convex objectives, most of these algorithms fall into the following two categories: subgradient based and dual based [24]. Subgradient based algorithms such as the incremental gradient method [7], decentralized subgradient method [28], and the dual averaging [15] usually require $O(1/\epsilon^2)$ rounds of communication, each followed by one gradient step. Within the function class, this achieves the optimal subgradient oracle complexity, but is suboptimal with respect to the communication rounds: as proven by [3, 33], the communication rounds needed is $O(1/\epsilon)$.

As a comparison, dual based algorithms, which dualize the consensus constraints, usually have better communication complexity: $O(1/\epsilon)$ rounds are needed for distributed ADMM [5, 42] and the decentralized communication sliding (DCS) [24], as examples. However, each round of communication is followed by optimization of Lagrangians or proximal updates, performed locally by each agent. To make the overall algorithm first-order, [24] proposes the Communication Sliding (CS) procedure, which approximates the proximal updates through $O(1/\epsilon)$ steps of (local) mirror descent, thereby achieving the $O(1/\epsilon^2)$ subgradient oracle complexity. The CS procedure has roots in the gradient sliding technique [23], which can save gradient computation for the smooth component when the objective involves a smooth and a non-smooth component.

For the class of strongly convex objectives, DCS can be accelerated, needing $O(1/\sqrt{\epsilon})$ rounds of communication and $O(1/\epsilon)$ gradient steps in total, both achieving the theoretical optimal [24]. In this work, due to the different time scales, we generalize the CS procedure for problems involving a mixture of Bregman divergences. In addition, we point out that for problems with *smooth* objectives, Local SGD – which applies gradient steps locally but communicates only once in a while – has been studied under various settings [36, 43, 47].

Primal-Dual Hybrid Gradient and its block variant. DCS [24] is inspired by the Primal-Dual Hybrid Gradient (PDHG) algorithm [11, 12]. In this work, motivated by real-life settings where the costs – e.g. time and/or resources – needed for communication between different agents are different [33], we further decompose the dual variables into blocks, and propose updating them at different frequencies. Thus, our algorithms can be viewed as *multi-timescale* variants of the PDHG. As a comparison, existing block-coordinate descent type of algorithms for the saddle point problem of interest, such as the Stochastic Primal-Dual Coordinate (SPDC) [45] and the Stochastic-PDHG (S-PDHG)[1, 13], update a random subset of blocks in each iteration k, where all blocks have strictly positive probability of being selected. The $O(1/k)$ rate of convergence, due to the randomness, is only shown for the *expected* objective value suboptimality (for SPDC) or *expected* duality gap (for S-PDHG). Nevertheless, in real applications, stochastic algorithms could potentially be less efficient, due to reasons such as random memory access [37] and potential overhead in computing sampling distributions [14]. Although deterministic block coordinate descent for convex optimization has been shown to converge, such as under the cyclic updating rule [37, 44], to the best of our knowledge, the *multi-timescale* updating rule we propose is the first *deterministic* block updating rule for PDHG with separable duals, such that different blocks could be updated different numbers of times, and the duality gap converges deterministically at the optimal rate.

Lower bounds on communication. In [3], it is shown that for distributed convex optimization, $O(1/\epsilon)$ rounds of communication are needed for 1-Lipschitz objectives, and $O(1/\sqrt{\mu\epsilon})$ rounds

are needed when the objectives are also μ-strongly convex. These lower bounds are achieved by splitting a "chain like" objective into two, each given to one agent. [33] extends these results to a decentralized, network setting and shows the dependence of the lower bounds on the network diameter and communication delay. [40] provides lower bound and (nearly) optimal algorithm for a different setup, where distributed agents have stochastic first order oracles to the same smooth nonconvex objective, but computation and communication speeds are bounded and different for different edges and agents. Apart from the round complexity, [41] shows a dimension-dependent lower bound on the bit-complexity of communication.

In addition, motivated by distributed training in machine learning, communication lower and upper bounds have been established using function similarities [2, 3, 18, 19, 22, 34, 38]: for instance, in (distributed) empirical risk minimization, the local loss functions have the same functional form but use different subsets of data, thereby inheriting the similarity in data. In [3], function similarities are measured using norms of the differences in (sub)gradients (and Hessians if exist), and a communication round lower bound linear in this measure is shown for convex Lipschitz objectives and strongly convex objectives. Known algorithms that take advantage of function similarities usually require additional assumptions such as strong convexity and smoothness [2, 18, 19, 22, 34, 38, 46]. As pointed out in [3], there is no known algorithm which achieve these communication round lower bounds for non-smooth convex objectives. In this work, we formalize the notion of function similarity for non-smooth convex objectives (Definition 4), and show that the communication round complexity for our (A)MT-GS indeed achieve these lower bounds, thereby answering [3]'s open question positively.

Appendix B. Setup and formulations

In (1) above and in the rest of this work, $\langle \cdot, \cdot \rangle$ is the standard Euclidean inner product. In addition, we assume that $\mathbb{R}^d$ is equipped with a norm $\|\cdot\|$ not necessarily generated by the inner product, and denote its dual norm as $\|\cdot\|_{x,*}$. We extend this norm to $\mathbb{R}^{md}$ through $\|X'\|^2 := \sum_{v\in V} \|x'_v\|^2$ for any $X' \in \mathbb{R}^{md}$, and its dual norm satisfies $\|X'\|^2_{X,*} = \sum_{v\in V} \|x'_v\|^2_{x,*}$. When there is no confusion, we drop the subscript x and X. We make the following assumption.

Assumption B.1 *($\mathcal{P}$) has an optimal solution $x^* \in \mathcal{X}$.*

When Assumption B.1 holds, we denote $X^* \in \mathbb{R}^{md}$ as $(X^*)_v = x^*$ for all $v \in V$.

B.1. Saddle point problem formulations and assumptions

Consensus constraints. In ($\mathcal{P}_r$), $K_s : \mathbb{R}^{md} \to \mathbb{R}^{n_s}$ is a linear operator such that $\cap_{s\in S} \ker(K_s)$ is the subspace in $\mathbb{R}^{md}$ where $\{x_v\}_{v\in V}$ does not violate the consensus constraints. That is, denoting $n = \sum_{s=1}^S n_s$ and $K : \mathbb{R}^{md} \to \mathbb{R}^n$ as $(KX)_s = K_s X$ for $s \in [S]$, we make the following requirement.

Assumption B.2 *$KX = \mathbf{0}$ if and only if $x_v = x_{v'}$ for all $v, v' \in V$.*

Denoting $\Pi : \mathbb{R}^{md} \to \mathbb{R}^{md}$ as the projection such that for any $X \in \mathbb{R}^{md}$, $\Pi(X)_v = \frac{1}{m}\sum_{v'\in V} x_{v'}$, we have $KX = \mathbf{0}$ if and only if $\Pi X = X$. Thus, Assumption B.2 implies that

$$K^*(KK^*)^\dagger K = I - \Pi.$$

For convenience, we further decompose $K_s X = \sum_{v \in V} K_{s,v} x_v$ for $K_{s,v} : \mathbb{R}^d \to \mathbb{R}^{n_s}$.

Penalty. In ($\mathcal{P}_r$), $R_s : \mathbb{R}^{n_s} \to \overline{\mathbb{R}}$ is a regularization term, penalizing the deviation of $K_s X$ from $\mathbf{0}$. We further define $R : \mathbb{R}^n \to \overline{\mathbb{R}}$ as $R(y_1, \ldots, y_S) = \sum_{s=1}^S R_s(y_s)$. Recall that the Fenchel conjugate of R_s is defined as

$$R_s^*(y_s) = \sup_{y_s' \in \mathbb{R}^{n_s}} \langle y_s, y_s' \rangle - R_s(y_s'),$$

and it is easy to see $R^*(Y) = \sum_{s=1}^S R_s^*(y_s)$ for $Y = (y_s)_{s \in [S]}$. We make the following assumption.

Assumption B.3 *For each $s \in [S]$, $R_s : \mathbb{R}^{n_s} \to \overline{\mathbb{R}}$,*

1. *R_s is proper, convex, and lower-semicontinuous;*
2. *$R_s(y_s) \geq 0$ for all $y_s \in \mathbb{R}^{n_s}$, and $R_s(y_s) = 0$ if and only if $y_s = \mathbf{0}$.*

As an example, if for all $s \in [S]$, R_s is the characteristic function of the set $\{\mathbf{0}\}$, i.e. $R_s(\mathbf{0}) = 0$ and $R_s(y_s) = \infty$ for $y_s \neq \mathbf{0}$, then ($\mathcal{P}_r$) is equivalent to ($\mathcal{P}$), and $R_s^*(y_s) = 0$ for all $y_s \in \mathbb{R}^{n_s}$. As another example, R_s can be any scaled norm, for instance $R_s(y_s) = \lambda \|y_s\|_p$ for some $p \geq 1$ and $\lambda > 0$, then $R_s^*(y_s) = 0$ for $\|y_s\|_q \leq \lambda$ and $R_s^*(y_s) = \infty$ otherwise, where $\|\cdot\|_q$ is the dual norm of $\|\cdot\|_p$ (i.e. $p^{-1} + q^{-1} = 1$). That is R_s^* is the characteristic function of the dual-norm-ball of size λ. We would like to point out that the first condition in Assumption B.3 is standard in the PDHG literature [11, 12]. We discuss further requirements on the choices of R_s in Section B.3.

Under the first part of Assumption B.3, we have $R_s = R_s^{**}$ (Theorem 11.1 [31]). Thus, ($\mathcal{P}_r$) can be equivalently formulated as the saddle point problem ($\mathcal{P}_s$).

Performance measure. To measure the performance of $X \in \overline{\mathcal{X}}$, following [24], we consider the (ϵ, δ)-solution, satisfying the following conditions

$$F(X) \leq F(X^*) + \epsilon, \quad \|(I - \Pi)X\| \leq \delta. \tag{3}$$

That is, X is ϵ-suboptimal in terms of the objective value, while violating the consensus constraints by at most δ.[1]

To solve ($\mathcal{P}$), we resort to the primal dual formulation ($\mathcal{P}_s$), where the common measure of performance of (X, Y) is the duality gap, defined as $G : \mathcal{Z} \times \mathcal{Z} \to \overline{\mathbb{R}}$ where $\mathcal{Z} = \overline{\mathcal{X}} \times \mathbb{R}^n$ as

$$\begin{aligned} G(X, Y; X', Y') :=& \langle KX, Y' \rangle + F(X) - R^*(Y') \\ & - \left\{ \langle KX', Y \rangle + F(X') - R^*(Y) \right\}. \end{aligned} \tag{4}$$

Our algorithms are guaranteed to find a pair $Z = (X, Y)$ such that $\sup_{Z' \in \mathcal{Z}} G(Z; Z')$ is small (Theorems 8 and 12). To transfer such duality gap guarantee back to (ϵ, δ)-solution guarantee, in Section B.3, we propose additional requirements for the regularization R.

1. In [24], $\|KX\| \leq \delta$ is used instead of $\|(I - \Pi)X\|$, and K is assumed to be the Laplacian matrix for the underlying graph of communication. However, we use a generic K satisfying condition B.2. In particular, for any K that is a valid choice, λK is also valid for any $\lambda \neq 0$. Thus, it makes sense to "normalize" K, and we use $I - \Pi = K^*(KK^*)^\dagger K$.

Distance generating functions and mirror maps. We equip $\mathcal{X}$ with a distance generating function[2] $w^x : \mathcal{X} \to \mathbb{R}$ with modulus 1, and extend it to $\overline{\mathcal{X}}$ through $w^X(X) := \sum_{v\in V} w^x(x_v)$. Similarly, for each $s \in [S]$, we assume that $\mathbb{R}^{n_s}$ is equipped with a norm $\|\cdot\|$ not necessarily generated by the inner product, and a distance generating function $w_s^y : \mathsf{dom}(R_s^*) \to \mathbb{R}$ with modulus 1. Recall that for any distance generating function w, the Bregman divergence is defined as $D_w(x,z) := w(x) - w(z) - \langle \nabla w(z), x - z\rangle$.

Assumption B.4 *For any $\overline{y}_s \in \mathsf{dom}(R_s^*)$, $g \in \mathbb{R}^{n_s}$ the following problem can be solved exactly:*

$$\min_{y_s \in \mathsf{dom}(R_s^*)} R_s^*(y_s) + \langle g, y_s\rangle + D_{w_s^y}(y_s, \overline{y}_s).$$

For any $g \in \mathbb{R}^d$, the following problem can be solved exactly:

$$\min_{x\in\mathcal{X}} \langle g, x\rangle + w^x(x).$$

B.2. Agents, communication, and additional requirements on K

By distributed optimization, we mean that the objective functions $\{f_v\}_{v\in V}$ are distributed among m *primal agents*: for each $v \in V$, $\mathsf{Agent}(x_v)$ has access to f_v', the first order oracle for f_v, and is responsible for updating the variable x_v. In addition, we assume that there are S *dual agents*: for each $s \in [S]$, $\mathsf{Agent}(y_s)$ is responsible for updating the variable y_s.

We assume that for any pair $(s,v) \in [S] \times V$ such that $K_{s,v} \neq \mathbf{0}$, $\mathsf{Agent}(x_v)$ and $\mathsf{Agent}(y_s)$ can communicate (in both directions). For instance, all agents might be nodes in a connected graph with vertices $[S] \cup V$ (representing S dual agents and m primal agents), and communication can be realized through edges (directly) or through paths (i.e. with the help of intermediate agents). In particular, since the graph is connected, any pair can communicate, but the resources consumed and/or time taken by communication between different pairs could be (significantly) different.

At this point, we abstract away from how such communication is realized, and leave the discussion of the costs of communication to Sections C.5.1 and C.5.2. Below, we provide two such realizations: *decentralized* and *hierarchical*, and provide examples in Figure 1.

Decentralized setting. In this setup, we assume that the dual variables are kept and updated by primal agents, respecting a graph based communication constraints. Precisely, let $\mathcal{G} = (V, E)$ denote an undirected, connected graph, and for each $s \in [S]$, we assign all tasks of $\mathsf{Agent}(y_s)$ to $\mathsf{Agent}(x_{v_s})$ for some $v_s \in V$, such that $\{v_s, v'\} \in E$ for each $K_{s,v'} \neq \mathbf{0}$.

As an example, let $W \in \mathbb{R}^{V\times V}$ be a doubly stochastic matrix such that $W_{v,v'} \neq 0$ only if $\{v,v'\} \in E$ or $v = v'$, and $\ker(I - W) = \mathsf{Span}(\mathbf{1})$ (and so $K := (I - W) \otimes I_d$ satisfies Assumption B.2). We can choose $S = m$, $n_s = d$, and decompose K as $K_s := (I - W)_s \otimes I_d$,

$$K_s X = \sum_{v\in V} (I - W)_{s,v} x_v = x_s - \sum_{\{v,s\}\in E} W_{s,v} x_v, \quad s = 1, \ldots, m.$$

2. For a convex closed set $\mathcal{S}$, a function $w : \mathcal{S} \to \mathbb{R}$ is a distance generating function [16] with modulus $\nu > 0$ w.r.t. $\|\cdot\|$ if w is continuously differentiable and

$$\langle x - z, \nabla w(x) - \nabla w(z)\rangle \geq \nu \|x - z\|^2, \quad \forall x, z \in \mathcal{S}.$$

Thus, $\mathsf{Agent}(y_s)$'s tasks can be assigned to $\mathsf{Agent}(x_s)$.

Hierarchical setting. In this setup, we assume that there is an underlying tree with nodes $[S] \cup V$, where all non-leaf nodes ($[S]$) correspond to dual agents and all leaf nodes (V) correspond to primal agents. Each non-leaf node can communicate with its child nodes directly. Precisely, for $s \in [S]$, we use $\mathsf{Chi}(s) \subset [S] \cup V$ to denote the child nodes of $\mathsf{Agent}(y_s)$, and $\mathsf{Des}(s) \subset V$ to denote all *primal agents* in the subtree rooted at $\mathsf{Agent}(y_s)$.

For convenience, for each $s \in [S]$, we denote the "mean" of all descendants of $\mathsf{Agent}(y_s)$ as $\overline{x}_s = |\mathsf{Des}(s)|^{-1} \sum_{j \in \mathsf{Des}(s)} x_j$. Then, consider $K_s : \mathbb{R}^{md} \to \mathbb{R}^{|\mathsf{Chi}(s)|d}$ defined as

$$(K_s X)_i = \overline{x}_i - \overline{x}_s = \overline{x}_i - \sum_{j \in \mathsf{Chi}(s)} \frac{|\mathsf{Des}(j)|}{|\mathsf{Des}(s)|} \overline{x}_j, \quad i \in \mathsf{Chi}(s). \tag{5}$$

Then, it is easy to see that K satisfies Assumption B.2, and since $\overline{x}_j$ can be computed in a bottom up manner, $\{K_s\}_{s \in [S]}$ can be realized through this tree. In addition, the set of $\{K_s\}_{s \in [S]}$ admits the following orthogonality properties which will be useful in choosing R. We defer the proof to Appendix F.1.

Lemma 1 *Let $\{K_s\}_{s \in [S]}$ be as defined in (5). Then for $s \neq s' \in [S]$, $K_s K_{s'}^* = \mathbf{0}$. In addition, denoting $\Pi_s := K_s^*(K_s K_s^*)^\dagger K_s$, we have for any $\widetilde{X}, \widehat{X} \in \mathbb{R}^{md}$*

$$\langle \widehat{X}, \Pi_s \widetilde{X} \rangle = \langle \Pi_s \widehat{X}, \Pi_s \widetilde{X} \rangle = \sum_{i \in \mathsf{Chi}(s)} |\mathsf{Des}(i)| \cdot \langle (K_s \widehat{X})_i, (K_s \widetilde{X})_i \rangle.$$

B.3. Requirements for R

Recall that when R is the characteristic function of $\{\mathbf{0}\}$, the penalized formulation ($\mathcal{P}_r$) and its primal-dual version ($\mathcal{P}_s$) are equivalent to ($\mathcal{P}$). In this section, we discuss the requirements for R such that the duality gap provides upper bounds on the suboptimality of the objective value and the violation of the consensus constraints.

Lemma 2 *Under Assumption B.3, we have for any $\widehat{X} \in \overline{\mathcal{X}}$ such that $K\widehat{X} = \mathbf{0}$,*

$$F(X) \leq F(\widehat{X}) + \sup_{Y' \in \mathsf{dom}(R^*)} G(X, Y; \widehat{X}, Y').$$

In particular, under Assumption (B.1), if $\sup_{Y' \in \mathsf{dom}(R^*)} G(X, Y; X^*, Y') \leq \epsilon$, then $F(X) \leq F(X^*) + \epsilon$.

Proof [Proof of Lemma 2] Notice that by Assumption (B.3),

$$\sup_{Y \in \mathbb{R}^n} \langle KX, Y \rangle + F(X) - R^*(Y) = F(X) + R(KX).$$

In addition, since $R(\mathbf{0}) = 0$, we have $R^*(Y) = \sup_{Y' \in \mathbb{R}^n} \langle Y', Y \rangle - R(Y') \geq \langle \mathbf{0}, Y \rangle - R(\mathbf{0}) = 0$, and so

$$\langle K\widehat{X}, Y \rangle + F(\widehat{X}) - R^*(Y) \leq F(\widehat{X}), \quad \forall\, \widehat{X} \in \overline{\mathcal{X}},\ K\widehat{X} = \mathbf{0}.$$

The second claim follows directly from the first since $KX^* = \mathbf{0}$. ■

To connect the duality gap with the constraint violation $\|(I-\Pi)X\|$ in (3), or with the objective value suboptimality of $\frac{1}{m}\sum_{v\in V} x_v$, it turns out additional requirements are needed for R.

For convenience, we denote $\sigma^+_{\min}(K_s) = \min_{X\in\mathbb{R}^{md},\ \Pi_s X\neq \mathbf{0}} \frac{\|K_s X\|_*}{\|\Pi_s X\|}$, where the numerator uses the dual norm to the norm in $\mathbb{R}^{n_s}$ and the denominator uses the norm in $\mathbb{R}^{md}$. As an example, when all norms are l_2 norms, $\sigma^+_{\min}(K_s)$ is the smallest non-zero singular value of K_s.

B.3.1. REQUIREMENTS ON R UNDER ORTHOGONALTIY

Below, we show that if K_s measures the constraint violation in *orthogonal* subspaces, then as long as R_s grows fast enough, the duality gap provides an upper bound on the constraint violation $\|(I-\Pi)X\|$ and the suboptimality of ΠX.

Lemma 3 *Under Assumption (B.3), further assuming that for any $s\neq s'\in[S]$, $K_sK_{s'}^* = \mathbf{0}$, and for each $s\in[S]$, denoting*

$$\Pi_s = K_s^*(K_sK_s^*)^\dagger K_s, \quad a_s \geq \sup_{X'\in\overline{\mathcal{X}},\ KX'=\mathbf{0}} \|\Pi_s \nabla F(X')\|_*,$$

where $\nabla F:\overline{\mathcal{X}}\to\mathbb{R}^{md}$ is an arbitrary subgradient oracle, i.e. $(\nabla F(X))_v\in\partial f_v(x_v)$. If Assumption B.1 also holds and $\sup_{Y'\in\mathsf{dom}(R^*)} G(X,Y;X^*,Y')\leq\epsilon$,

1. *X is an $(\epsilon,\epsilon/\xi)$-solution if for each $s\in[S]$,*

$$R_s(y_s)\geq R_s^{ccv}(y_s) := \frac{\xi+a_s}{\sigma^+_{\min}(K_s)}\|y_s\|_*. \tag{6}$$

2. *the projected solution ΠX is an $(\epsilon(1+1/\xi),0)$-solution if for each $s\in[S]$,*

$$a_s>0, \quad R_s(y_s)\geq R_s^{prj}(y_s) := \frac{(1+\xi)a_s}{\sigma^+_{\min}(K_s)}\|y_s\|_*. \tag{7}$$

In Lemma 3 (Corollary 5 below), the superscript ccv means $\{R_s^{ccv}\}_{s\in[S]}$ ($\{\widehat{R}_s^{ccv}\}_{s\in[S]}$) are designed to provide guarantees on the consensus constraint violation, and the superscript prj means $\{R_s^{prj}\}_{s\in[S]}$ ($\{\widehat{R}_s^{prj}\}_{s\in[S]}$) are designed to provide guarantees the projected solution ΠX.

Proof [Proof of Lemma 3] First, notice that by the orthogonality of $\{K_s\}_{s\in[S]}$, for any $Y\in\mathbb{R}^n$

$$(KK^*Y)_s = K_s(\sum_{s'=1}^S K_{s'}^* y_{s'}) = K_sK_s^*y_s,\ \forall s\in[S].$$

That is, KK^* is diagonal, and so

$$((KK^*)^\dagger Y)_s = (K_sK_s^*)^\dagger y_s,\ \forall s\in[S].$$

Thus, we can make the following decomposition

$$K^*(KK^*)^\dagger KX = \sum_{s=1}^S K_s^*(K_sK_s^*)^\dagger K_sX = \sum_{s=1}^S \Pi_s X.$$

For convenience, we denote $\widetilde{X} := \Pi X$, and by Lemma 3,

$$\sup_{Y' \in \mathsf{dom}(R^*)} G(X, Y; X^*, Y') \le \epsilon \implies F(X) + R(KX) \le F(X^*) + \epsilon \le F(\widetilde{X}) + \epsilon. \quad (8)$$

In addition, using the convexity of F,

$$\begin{aligned} F(\widetilde{X}) - F(X) &\le -\langle \nabla F(\widetilde{X}), (I - \Pi)X \rangle = -\sum_{s=1}^{S} \langle \nabla F(\widetilde{X}), \Pi_s X \rangle \\ &\le \sum_{s=1}^{S} \|\Pi_s \nabla F(\widetilde{X})\|_* \cdot \|\Pi_s X\| \le \sum_{s=1}^{S} a_s \cdot \|\Pi_s X\|. \end{aligned} \quad (9)$$

For the first claim, since $\|K_s X\|_* \ge \|\Pi_s X\| \sigma^+_{\min}(K_s)$, with the first condition (6) on R_s, we have

$$R_s(K_s X) \ge (\xi + a_s) \cdot \|\Pi_s X\| \quad (10)$$

Combining the (8), (9), and (10), we get

$$\xi \cdot \sum_{s=1}^{S} \|\Pi_s X\| \le \epsilon \implies \|(I - \Pi)X\| = \|\sum_{s=1}^{S} \Pi_s X\| \le \sum_{s=1}^{S} \|\Pi_s X\| \le \epsilon/\xi.$$

For the second claim, following a similar argument as above but with the second condition (7) on R_s, we get

$$\sum_{s=1}^{S} a_s \cdot \|\Pi_s X\| \le \epsilon/\xi. \quad (11)$$

Thus, using (8), (9), and (11), we have

$$F(\widetilde{X}) \le F(X) + \epsilon/\xi \le F(X^*) + \epsilon/\xi + \epsilon.$$

■

We would like to point out that in (9), $\langle \nabla F(\widetilde{X}), \Pi_s X \rangle$ is upper bounded using $\|\nabla F(\widetilde{X})\|_* \cdot \|\Pi_s X\|$. A tighter upper bound could be obtained if one has more information about the set $\mathcal{G}_s := \{\Pi_s \nabla F(X'),\ X' \in \overline{\mathcal{X}},\ KX' = \mathbf{0}\}$. Indeed, $\langle \nabla F(\widetilde{X}), \Pi_s X \rangle \le \sup_{G_s \in \mathcal{G}_s} \langle G_s, \Pi_s X \rangle$, and so the inner product can be bounded using the support function of the set $\mathcal{G}_s$.

B.3.2. FUNCTION SIMILARITY FOR GENERAL CONVEX FUNCTIONS

The terms $\{a_s\}_{s \in [S]}$ in Lemma 3 can be viewed as a "decomposition" of the function variation into different subspaces spanned by (the row spaces of) $\{K_s\}_{s \in [S]}$. To be more concrete, consider the hierarchical setting presented in Section B.2, which satisfies exactly the conditions in Lemma 3 due to Lemma 1. Defining $\mu_s(i) = \frac{|\mathsf{Des}(i)|}{|\mathsf{Des}(s)|}$ for $i \in \mathsf{Chi}(s)$ as a probability measure, and assuming that all norms are l_2 norms, then by Lemma 1,

$$\|\Pi_s \nabla F\|_*^2 = |\mathsf{Des}(s)| \cdot \mathsf{Var}_{i \sim \mu_s}(\overline{f'}_i), \quad \overline{f'}_i = \frac{\sum_{j \in \mathsf{Des}(i)} f'_j}{|\mathsf{Des}(i)|},\ i \in \mathsf{Chi}(s), \quad (12)$$

where for a random vector V, we denote $\mathsf{Var}(V) := \mathbb{E}[\|V - \mathbb{E}[V']\|_*^2]$. Thus, $\|\Pi_s \nabla F\|_*$ measures the function variation among the *descendants of different child nodes* of $\mathsf{Agent}(y_s)$, i.e. among $\left\{\sum_{j\in\mathsf{Des}(i)} f_j'\right\}_{i\in\mathsf{Chi}(s)}$. As a result, the agents closer to the root of the tree, with more descendants, take care of function variation at *larger scales*, but at *lower resolution*, since for all $i \in \mathsf{Chi}(s)$, the variation inside $\left\{f_j'(x)\right\}_{j\in\mathsf{Des}(i)}$ has been taken care of by the dual agents in each sub-tree rooted at i.

For general but still orthogonal $\{K_s\}_{s\in[S]}$, a_s measures the function variation along the span of K_s. This interpretation in mind, we make the following definition regarding function similarity.

Definition 4 *Assume that for all $s \neq s' \in [S]$, $K_s K_{s'}^* = \mathbf{0}$. We say that the set of functions $\{f_v\}_{v\in V}$ is $\{(a_s, K_s)\}_{s\in[S]}$-similar if there exists a subgradient oracle $\nabla F : \overline{\mathcal{X}} \to \mathbb{R}^{md}$, i.e. $(\nabla F(X))_v \in \partial f_v(x_v))$, such that for each $s \in [S]$,*

$$\Pi_s = K_s^*(K_s K_s^*)^\dagger K_s, \quad a_s \geq \sup_{X'\in\overline{\mathcal{X}},\, KX'=\mathbf{0}} \|\Pi_s \nabla F(X')\|_*.$$

If $S = 1$ and $\Pi_1 = I - \Pi$, we abbreviate $\{(a_1, K_1)\}$-similar as a_1-similar.

For instance, if $S = 1$ and all norms are l_2 norms, then Assumption B.2 requires that $\Pi_1 = I - \Pi$, and one can take a_1 as

$$a_1^2 \geq \sup_{x\in\mathcal{X}} \sum_{v\in V} \|f_v'(x) - \frac{1}{m}\sum_{v'\in V} f_{v'}'(x)\|^2.$$

Thus, if $\|f_v'(x)\| \leq M_f$ for all $v \in V, x \in \mathcal{X}$, we can also take $a_1 = 2\sqrt{m}M_f$.

Comparisons with existing notions of function similarity. [18] proposes the *bounded gradient dissimilarity* for differentiable convex objectives, which coincides with our Definition 4 when $S = 1$ and when the objectives are differentiable. For twice differentiable objectives, function similarity is also defined in terms of differences in Hessians, i.e. $\|\nabla^2 f_v - \nabla^2 f_{v'}\|$ [3, 18, 20, 38]. For general convex functions which could be non-differentiable, [3] informally defines it (δ-*relatedness* in their terminology) as the condition that "subgradients of local functions are at most δ-different from each other". Our Definition 4 formalizes this idea, and extend it to the case where $S > 1$.

B.3.3. REQUIREMENTS ON R_s WITHOUT ORTHOGONALITY

The above Lemma 3 imposes orthogonality assumptions on $\{K_s\}_{s\in[S]}$. In the more general case where such assumptions do not hold, one can always view ($\mathcal{P}_s$) as a problem with only 1 block, with K and R as the corresponding operator and regularization. Applying Lemma 3, we get the following corollary.

Corollary 5 *Under Assumption (B.3), denoting $\widehat{a}_1 \geq \sup_{X'\in\overline{\mathcal{X}},\, KX'=\mathbf{0}} \|(I-\Pi)\nabla F(X')\|_*$ where $\nabla F : \overline{\mathcal{X}} \to \mathbb{R}^{md}$ is an arbitrary subgradient oracle, i.e. $(\nabla F(X))_v \in \partial f_v(x_v)$. If Assumption B.1 also holds and $\sup_{Y'\in\mathsf{dom}(R^*)} G(X, Y; X^*, Y') \leq \epsilon$,*

1. *X is an $(\epsilon, \epsilon/\xi)$-solution if for each $s \in [S]$,*

$$R_s(y_s) \geq \widehat{R}_s^{ccv}(y_s) := \frac{\xi + \widehat{a}_1}{\sigma_{\min}^+(K)}\|y_s\|_*, \tag{13}$$

2. *assume that* $\widehat{a}_1 > 0$*, then the projected solution* ΠX *is an* $(\epsilon(1+1/\xi), 0)$*-solution if for each* $s \in [S]$*,*

$$R_s(y_s) \geq \widehat{R}_s^{prj}(y_s) := \frac{(1+\xi)\widehat{a}_1}{\sigma_{\min}^+(K)}\|y_s\|_*. \tag{14}$$

Assume that all norms are l_2 norms, then with orthogonality, KK^* is "diagonal" and so $\sigma_{\min}^+(K) = \min_{s\in[S]}\sigma_{\min}^+(K_s)$. In addition, since $\|(I-\Pi)\nabla F(X)\| \geq \|\Pi_s \nabla F(X)\|$ for any $X \in \overline{\mathcal{X}}$ and $s \in [S]$, one can always take $a_s \leq \widehat{a}_1$ for all $s \in [S]$. Thus $\frac{\xi+\widehat{a}_1}{\sigma_{\min}^+(K)} \geq \frac{\xi+a_s}{\sigma_{\min}^+(K_s)}$. Since for the function $h(x) = \lambda\|x\|$ defined on $\mathbb{R}^{n_0}$ for some $\lambda > 0$, the conjugate h^* is the characteristic function of $\{x \in \mathbb{R}^{n_0} | \|x\| \leq \lambda\}$, i.e. $\mathsf{dom}(h^*) = \{x \in \mathbb{R}^{n_0} | \|x\| \leq \lambda\}$. As will be seen in Theorem 8, the convergence is faster with smaller domains, suggesting that one should use the more refined decomposition when orthogonality holds.

Comparisons with [24] when $S = 1$**.** Assume that $\widehat{R}_1^{prj}$ in (14) is used for some constant $\xi > 0$ and $\widehat{a}_1 = 2\sqrt{m}M_f$, where M_f (defined below) is an upper bound on the norm of *the subgradient oracle* $f_v' \in \partial f_v$ (i.e. only one subgradient in the subdifferential for each $x \in \mathcal{X}$, $v \in V$). Then, the diameter of $\mathsf{dom}(R^*)$ is $O(\frac{\sqrt{m}M_f}{\sigma_{\min}^+(K)})$. In [24], it is shown that for ($\mathcal{P}_s$) with R being the characteristic function of $\{\mathbf{0}\}$, there exists an optimal dual solution $\|Y^*\| \leq \frac{\sqrt{m}\widehat{M}_f}{\sigma_{\min}^+(K)}$, where $\widehat{M}_f$ is an upper bound on the norms of *all subgradients* $g \in \partial f_v$:

$$\widehat{M}_f := \sup_{x\in\mathcal{X},\, v\in V,\, g\in\partial f_v(x)} \|g\|_* \geq M_f := \sup_{x\in\mathcal{X},\, v\in V} \|f_v'(x)\|_*.$$

Thus, even without function similarity, our $\widehat{R}_1^{prj}$ provides better control over the dual variables, leading to faster convergence. Moreover, when $\mathcal{X} \neq \mathbb{R}^d$, due to the normal cones at the boundary of $\mathcal{X}$, $\widehat{M}_f = \infty$.

Appendix C. Multi-timescale gradient sliding

To solve ($\mathcal{P}_s$) with the costs of information aggregation in mind, we resort to the decentralized communication sliding (DCS)[24], a communication efficient variant of the classical primal-dual hybrid gradient algorithm. As a recap, at iteration k, DCS performs the following updates: (R is the characteristic function of $\{\mathbf{0}\}$, and $w^y(Y) = \frac{1}{2}\langle Y, Y\rangle$)

$$\widetilde{X}^k = X^k + \alpha_k(X^k - X^{k-1}) \tag{15}$$

$$Y^k = \operatorname*{argmin}_{Y\in\mathbb{R}^n} R^*(Y) + \langle -K\widetilde{X}^k, Y\rangle + \tau_k D_{w^y}(Y, Y^{k-1}) \tag{16}$$

$$X^k \approx \operatorname*{argmin}_{X\in\overline{\mathcal{X}}} F(X) + \langle K^*Y^k, X\rangle + \eta_k D_{w^X}(X, X^{k-1}) \tag{17}$$

where the proximal update (17) is solved inexactly through the Communication-Sliding (CS) procedure (i.e. T_k steps of mirror descent) locally by each primal agent. Since only computing $K\widetilde{X}^k$ and K^*Y^k requires the communication between primal and dual agents, to reach ϵ suboptimality in terms of the gap, $O(1/\epsilon)$ rounds of communication/matrix-vector multiplication are needed. This is desirable, especially when the bottleneck of the entire algorithm is communication and/or matrix-vector multiplication.

We take a step further: with separable duals, we consider potential heterogeneity in the costs of information aggregation (e.g. sending messages and/or computing matrix-vector products). Since the dual (16) can be separated into S blocks, one can apply block-coordinate descent type of algorithms such as S-PDHG [13], which updates only a random subset of the dual blocks at each iteration. The flexibility in choosing the sampling distribution allows one to control the frequency of updates of different blocks. However, due to the randomness, the $O(1/k)$ rate of convergence is usually shown only for the expected gap.

To maintain the *deterministic convergence guarantee* as well as the *flexibility in choosing the number of updates applied to each dual block*, we propose the Multi-Timescale Gradient Sliding (MT-GS) for ($\mathcal{P}_s$), where different dual agents live in different timescales, and update at different rates: $\mathsf{Agent}(y_s)$ only updates y_s at iteration $0, r_s, 2r_s, \ldots$. Our convergence results indicate that to reach ϵ duality gap, the number of communication rounds needed is $N = O(\frac{\overline{r}}{\epsilon})$, where $\overline{r}$ is the weighted average of $\{r_s\}_{s\in[S]}$, with weights depending on "dual domain sizes", and thus on the function similarities if the penalties R_s's are chosen as suggested by Section B.3. In addition, when the costs of updating different duals and/or the function variation along different K_s's are different (significantly), one can take advantage of the flexibility in choosing $\{r_s\}_{s\in[S]}$ to design more efficient algorithms.

C.1. Updating rules of multi-timescale gradient sliding

We assume that there is a global time $k = 0, 1, 2, \ldots, N$.

Initialization. We assume that each $\mathsf{Agent}(x_v)$ is given some $x_v^{init} \in \mathcal{X}$, and initialize $x_v^{k'} = \widehat{x}_v^{k'} = x_v^{init}$ for all $k' < 0$. Similarly, each $\mathsf{Agent}(y_s)$ is given some $y_s^{init} \in \mathsf{dom}(R_s^*)$, and initialize $y_s^{k'} = y_s^{init}$ for all $k' < 0$.

Dual updates. We associate each dual y_s with a rate $r_s \in \mathbb{N}$ and a local time $i_s = 0, 1, \ldots, N_s - 1$, such that $N + 1 = r_s N_s$. For $k = 0, 1, \ldots, N$, $\mathsf{Agent}(y_s)$ remains dormant (no computation or communication) unless $k = r_s i_s$ for some $i_s \in \{0, 1, \ldots, N_s - 1\}$, where $\mathsf{Agent}(y_s)$ computes the updated $y_s^{i_s}$ using the following rules.

$$\widetilde{x}_{s,v}^{i_s} = \alpha_{s,i_s}\left(\sum_{k'=r_s i_s - r_s}^{r_s i_s - 1} \theta_{k'}(\widehat{x}_v^{k'} - x_v^{k'-r_s})\right) + \sum_{k'=r_s i_s - r_s}^{r_s i_s - 1} \theta_{k'+r_s} x_v^{k'}, \quad v \in V, \tag{18}$$

$$y_s^{i_s} = \operatorname*{argmin}_{y_s \in \mathbb{R}^{n_s}} \langle -\frac{1}{\sum_{k'=r_s i_s}^{r_s i_s + r_s - 1} \theta_{k'}} \sum_{v\in V} K_{s,v}\widetilde{x}_{s,v}^{i_s}, y_s\rangle + R_s^*(y_s) + \tau_{s,i_s} D_{w_s^y}(y_s, y_s^{i_s-1}). \tag{19}$$

Further, we denote $\overline{y}_s^k = y_s^{\lfloor k/r_s \rfloor}$, i.e. the corresponding dual y_s at the global time k, and $\overline{Y}^k = (\overline{y}_s^k)_{s\in[S]}$

Primal updates. All primal variables are updated at each global time, through a generalized CS procedure, which we provide details in Section C.2. For $k = 0, 1, \ldots, N$,

$$(x_v^k, \widehat{x}_v^k) = CS(f_v, \mathcal{X}, D_{w^x}, T_k, (\eta_{k,s})_{s\in[S]}, (x_v^{k-r_s})_{s\in[S]}, \sum_{s=1}^S K_{s,v}^* \overline{y}_s^k, x_v^{k-1}), \quad \forall\, v \in V. \tag{20}$$

With correct choices of the parameters, (20) provides approximate solutions to the following problem

$$\min_{x\in\mathcal{X}} f_v(x) + \langle \sum_{s=1}^S K_{s,v}^* \overline{y}_s^k, x\rangle + \sum_{s=1}^S \eta_{k,s} D_{w^x}(x, x_v^{k-r_s}). \tag{21}$$

Final outputs. Denoting $\widehat{Z}^k = (\widehat{X}^k, \overline{Y}^k)$, then the output is

$$Z^N = (\sum_{k=0}^{N} \theta_k)^{-1} \sum_{k=0}^{N} \theta_k \widehat{Z}^k. \tag{22}$$

Algorithm 1: (Accelerated) Multi-timescale gradient sliding

Data: $\{\alpha_{s,i_s}\}, \{\theta_k\}, \{T_k\}, \{\eta_{k,s}\}, \{\tau_{s,i_s}\}, \{r_s\}, X^{init}, Y^{init}$
$(X^{k'}, \widehat{X}^{k'}, Y^{k'}) \leftarrow (X^{init}, X^{init}, Y^{init})$ for all $k' < 0$;
for $k = 0, 1, \ldots, N$ **do**
  ▷ implicitly $i_s = \lfloor k/r_s \rfloor$ for all $s \in [S]$ **for** $s \in [S]$ *such that* $k = 0$ (*mod* r_s) **do**
    for $v \in V$ *such that* $K_{s,v} \neq \mathbf{0}$ **do**
      Agent(x_v) computes $\widetilde{x}_{s,v}^{i_s}$ using (18), and sends to Agent(y_s);
    end
    Agent(y_s) computes $y_s^{i_s}$ using (19), and sends $y_s^{i_s} - y_s^{i_s-1}$ (y_s^0 if $i_s = 0$) to Agent(x_v) for all $v \in V$ such that $K_{s,v} \neq \mathbf{0}$;
  end
  for $v \in V$ **do**
    Agent(x_v) computes $\sum_{s=1}^{S} K_{s,v}^* \overline{y}_s^k$, and updates $(x_v^k, \widehat{x}_v^k)$ using (20);
  end
end
Output Z^N in (22);

Notice Agent(x_v) needs to be able to compute all the $\widetilde{x}_{s,v}^{i_s}$, which could require extra space to store past $(x_v, \widehat{x}_v)$'s. One approach is to keep in memory all $x_v^{(k-2r_{\max}):(k-1)}$ and $\widehat{x}_v^{(k-r_{\max}):(k-1)}$, where $r_{\max} = \max_{s\in[S]} r_s$. This requires storing $3r_{\max}$ vectors in $\mathbb{R}^d$. Another approach is to keep in memory the accumulative $(x_v, \widehat{x}_v)$: for each $s \in [S]$, for $r_s(i_s - 1) \leq k \leq r_s i_s$, keep in memory (previously computed) $\sum_{k'=r_s i_s - 2r_s}^{r_s i_s - r_s - 1} \theta_{k'+r_s} x_v^{k'}$, while computing the sum $\sum_{k'=r_s i_s - r_s}^{r_s i_s - 1} \theta_{k'} \widehat{x}_v^{k'}$ and $\sum_{k'=r_s i_s - r_s}^{r_s i_s - 1} \theta_{k'+r_s} x_v^{k'}$ as k increases. This requires storing $3S$ vectors in $\mathbb{R}^d$. Also, due to the proximal centers, $x_v^{(k-r_{\max}):k}$ should be kept in memory, which requires storing $r_{\max}$ vectors.

In addition, notice that in Algorithm 1, Agent(y_s) calculates $K_{s,v}\widetilde{x}_{s,v}^{i_s}$ and sends a vector in $\mathbb{R}^{n_s}$, Agent(x_v) calculates $K_{s,v}^*(\overline{y}_s^k - \overline{y}_s^{k-1})$ and sends a vector in $\mathbb{R}^d$. In fact, there are many task assignment strategies: for instance, $K_{s,v}\widetilde{x}_{s,v}^{i_s}$ can also be computed by Agent(x_v), and the message from Agent(x_v) to Agent(y_s) will be $K_{s,v}\widetilde{x}_{s,v}^{i_s}$. This is preferable if Agent(x_v) can compute matrix-vector products faster/at lower cost than Agent(y_s). Due to this variability, in the cost analysis in Section C.5, we take a "modular" perspective and assume that the cost of updating y_s (including all matrix-vector multiplication and communication) is c_s.

C.2. Generalized communication sliding

In [24], the CS procedure is used to approximately solve the primal proximal updates. However, due to differences in dual time scales, we need extra control on the variation of the primal sequence. Intuitively, when computing X^k, the $K_s^* y_s$ term is evaluated using $\overline{y}_s^k = y_s^{i_s}$, where $i_s = \lfloor k/r_s \rfloor$. However, $y_s^{i_s}$ is computed at the global iteration $r_s\lfloor k/r_s \rfloor \leq k$, using the (outdated)

$X_s^{(r_s i_s - 2r_s):(r_s i_s - 1)}$ and $\widehat{X}_s^{(r_s i_s - r_s):(r_s i_s - 1)}$. Thus, for the multi-timescale updates to converge, we need to control the variation of the primal sequence. To achieve that, we take the proximal term in the primal updates as a mixture of proximal terms centered at $\{X^{k-r_s}\}_{s\in[S]}$, which gives the formulation (21).

Motivated by this, we propose the following generalization of the CS procedure in [24] (which is a special case where $|\mathcal{I}| = 1$).

Algorithm 2: Generalized communication sliding procedure

Data: The sequences $\{\beta_t\}$ and $\{\lambda_t\}$, $\phi' : U \to \mathbb{R}^{d_0}$ a subgradient oracle for ϕ.
Result: $(u^T, \widehat{u}^T) = CS(\phi, U, D_w, T, (\eta_i)_{i\in\mathcal{I}}, v, (x_i)_{i\in\mathcal{I}}, x^{init})$, an approximate solution to

$$\min_{u\in U} \Phi(u) := \langle v, u\rangle + \phi(u) + \sum_{i\in\mathcal{I}} \eta_i D_w(u, x_i)$$

$(u^0, \widehat{u}^0) \leftarrow (x^{init}, x^{init})$, $\eta \leftarrow \sum_{i\in\mathcal{I}} \eta_i$;
for $t = 1, \ldots, T$ **do**

$$u^t = \operatorname*{argmin}_{u\in U} \langle v + \phi'(u^{t-1}), u\rangle + \sum_{i\in\mathcal{I}} \eta_i D_w(u, x_i) + \eta\beta_t D_w(u, u^{t-1})$$

end

$$\widehat{u}^T = (\sum_{t=1}^T \lambda_t)^{-1} \sum_{t=1}^T \lambda_t u^t.$$

As a corollary to Lemma 18, we have the following performance guarantee.

Corollary 6 *Assume that $U \subset \mathbb{R}^{d_0}$ is a convex set, and $\phi : U \to \mathbb{R}$ is a convex function such that*

$$\frac{\mu}{2}\|x - y\|^2 \le \phi(x) - \phi(y) - \langle \phi'(y), x - y\rangle \le M\|x - y\|, \quad \forall x, y \in U,$$

where $\phi' : U \to \mathbb{R}^{d_0}$ is a subgradient oracle, i.e. for each $y \in U$, $\phi'(y) \in \partial\phi(y)$ is a subgradient. With $\lambda_t = t + 1$ and $\beta_t = \frac{t}{2}$ for $t \ge 1$, we have for any $u \in U$

$$\begin{aligned}\langle v, \widehat{u}^T - u\rangle + \phi(\widehat{u}^T) - \phi(u) \le{}& \frac{2\eta}{T(T+3)} D_w(u, x^{init}) + \sum_{i\in\mathcal{I}} \eta_i D_w(u, x_i) \\ & - \frac{(T+1)(T+2)}{T(T+3)} \eta D_w(u, u^T) - \sum_{i\in\mathcal{I}} \eta_i D_w(\widehat{u}^T, x_i) + \frac{4M^2}{\eta(T+3)}.\end{aligned}$$

Further, if $\mu > 0$, and $D_{w^x}(x, x') \le \frac{C}{2}\|x - x'\|^2$ for some $C < \infty$, then denoting $\eta = \sum_{i\in\mathcal{I}} \eta_i$, setting $\lambda_t = t$ and $\beta_t = \frac{(t+1)\mu}{2\eta C} + \frac{t-1}{2}$, we have for any $u \in U$,

$$\begin{aligned}\langle v, \widehat{u}^T - u\rangle + \phi(\widehat{u}^T) - \phi(u) \le{}& \sum_{i\in\mathcal{I}} \eta_i D_w(u, x_i) - \sum_{i\in\mathcal{I}} \eta_i D_w(\widehat{u}^T, x_i) \\ & - (\frac{\mu}{C} + \eta) D_w(u, u^T) + \frac{2M^2/\eta}{T(T+1)} \sum_{t=1}^T \frac{\lambda_t}{\beta_t},\end{aligned}$$

and $\frac{2M^2/\eta}{T(T+1)}\sum_{t=1}^{T}\frac{\lambda_t}{\beta_t} \le \frac{4CM^2}{\mu(T+1)}$.

C.3. Convergence of multi-timescale gradient sliding

The proof of convergence of Algorithm 1 follows a similar type of argument as the proof of convergence of PDHG and DCS: the primal updates (20) and dual updates control the following two terms ((25), (26)):

$$\left\{\sum_{k=0}^{N}\langle K^*\overline{Y}^k, \widehat{X}^k - X\rangle + F(\widehat{X}^k) - F(X)\right\} + \left\{\sum_{i_s=0}^{N_s-1}\langle -K_s\widetilde{X}_s^{i_s}, y_s^{i_s} - y_s\rangle + r_s(R_s^*(y_s^{i_s}) - R_s^*(y_s))\right\}.$$

The above sum (approximately) matches the gap $\sum_{k=0}^{N} Q(\widehat{Z}^k, Z)$ up to an additive term ((27))

$$\sum_{s=1}^{S}\sum_{k=0}^{N}\langle \widehat{X}^k - \widetilde{X}_s^{\lfloor k/r_s\rfloor}, K_{s,v}^*(y_s - \overline{y}_s^k)\rangle = \sum_{i_s=0}^{N_s-1}\langle\sum_{i=0}^{r_s-1}\widehat{x}_v^{r_s i_s+i} - \widetilde{x}_{s,v}^{i_s}, K_{s,v}^*(y_s - y_s^{i_s})\rangle.$$

Notice that as compared to the proof of PDHG and DCS, due to the different timescales for the duals, we bound the above terms at dual time scales: instead of controlling $\langle \widehat{x}_v^k - \widetilde{x}_{s,v}^{\lfloor k/r_s\rfloor}, K_{s,v}^*(y_s - y_s^{\lfloor k/r_s\rfloor})\rangle$ for each k, we control the cumulative term (sum from $k = r_s i_s$ to $k = r_s(i_1+1)-1$). With our choice of the $\widetilde{X}_s^{i_s}$ and the mixture terms used in primal proximal updates, the result follows. We defer the proof to Appendix F.3.

Lemma 7 *Under Assumption B.4 and the first part of Assumption B.3, with the following choice of parameters:*

- $\alpha_{s,i_s} = \alpha = 1$, $\theta_k = 1$, $T_k = T \ge 1$;
- $\eta_{k,s} = \eta\rho_s$ *where* $\rho_s \ge 0$ *and* $\sum_{s=1}^{S}\rho_s = 1$;
- $\tau_s = \frac{2\widetilde{\kappa}_s^2}{\rho_s\eta}$ *where* $\widetilde{\kappa}_s := \sup_{\|y_s\|\le 1}\|K_s^* y_s\|_*$;
- $\lambda_t = t+1$ *and* $\beta_t = t/2$ *for the CS procedure for all iteration* k.

Then for any $Z \in \overline{\mathcal{X}} \times \mathbb{R}^n$,

$$\begin{aligned}&(N+1)\cdot Q(Z^N; Z)\\ \le\ &\eta\left\{\frac{3}{2}(\sum_{s=1}^{S} r_s\rho_s)D_{w^X}(X, X^{init}) - D_{w^X}(X, X^N)\right\}\\ &+\frac{1}{\eta}\left\{\sum_{s=1}^{S}\frac{\widetilde{\kappa}_s^2 r_s}{\rho_s}\left\{3D_{w_s^y}(y_s, y_s^{init}) - D_{w_s^y}(y_s, y_s^{N_s-1})\right\} + \frac{4mM^2(N+1)}{T+3}\right\}\end{aligned}$$

Thus, with proper choices of η, ρ_s, T, we get the following bound.

Theorem 8 *For $\widehat{X} \in \overline{\mathcal{X}}$, assume that the following are finite:*

$$D_{w^X}(\widehat{X}, X^{init}) \leq D^X < \infty, \qquad \sup_{y_s \in \mathsf{dom}(R_s^*)} D_{w_s^y}(y_s, y_s^{init}) \leq D_s^y < \infty.$$

Under the conditions in Lemma 7, taking $\eta = (\sum_{s'=1}^S \widetilde{\kappa}_{s'} \sqrt{D_{s'}^y}) \sqrt{\frac{8}{3D^X}}$, $\rho_s = \frac{\widetilde{\kappa}_s \sqrt{D_s^y}}{\sum_{s'=1}^S \widetilde{\kappa}_{s'} \sqrt{D_{s'}^y}}$, *and* $T \geq \lfloor \frac{4mM^2(N+1)}{\overline{r}(\sum_{s=1}^S \widetilde{\kappa}_s \sqrt{D_s^y})^2} \rfloor$ *where* $\overline{r} := \sum_{s=1}^S r_s \rho_s$, *we have*

$$\sup_{Y' \in \mathbb{R}^n} Q(Z^N; \widehat{X}, Y') \leq \frac{2\sqrt{6\overline{r}} \cdot (\sum_{s=1}^S \widetilde{\kappa}_s \sqrt{D_s^y}) \cdot \sqrt{D^X}}{N+1}. \tag{23}$$

Proof [Proof of Theorem 8] From Lemma 7, we first notice that with $\rho_s = \frac{\widetilde{\kappa}_s \sqrt{D_s^y}}{\sum_{s'=1}^S \widetilde{\kappa}_{s'} \sqrt{D_{s'}^y}}$

$$\sum_{s=1}^S \frac{\widetilde{\kappa}_s^2 r_s D_s^y}{\rho_s} = \overline{r} (\sum_{s'=1}^S \widetilde{\kappa}_{s'} \sqrt{D_{s'}^y})^2.$$

Thus, we have

$$T + 3 \geq \frac{4mM^2(N+1)}{\overline{r}(\sum_{s=1}^S \widetilde{\kappa}_s \sqrt{D_s^y})^2} \implies \frac{4mM^2(N+1)}{T+3} \leq \sum_{s=1}^S \frac{\widetilde{\kappa}_s^2 r_s D_s^y}{\rho_s}$$

Thus, with the additional assumptions, we get

$$\begin{aligned}
\sup_{Y' \in \mathbb{R}^n} Q(Z^N; \widehat{X}, Y') &\leq (N+1)^{-1} \left\{ \frac{3\eta D^X}{2} (\sum_{s=1}^S r_s \rho_s) + \frac{4}{\eta} (\sum_{s=1}^S \frac{\widetilde{\kappa}_s^2 r_s D_s^y}{\rho_s}) \right\} \\
&= (N+1)^{-1} \left\{ \frac{3\eta D^X}{2} \cdot \overline{r} + \frac{4}{\eta} \overline{r} (\sum_{s'=1}^S \widetilde{\kappa}_{s'} \sqrt{D_{s'}^y})^2 \right\} \\
&= \frac{2\sqrt{6\overline{r}} (\sum_{s'=1}^S \widetilde{\kappa}_{s'} \sqrt{D_{s'}^y}) \cdot \sqrt{D^X}}{N+1}.
\end{aligned}$$

■

We would like to point out that Algorithm 1 and the above guarantees Lemma 7 and Theorem 8 (as well as Lemma 11 and Theorem 12 below) hold for *any* saddle point problem of the structure ($\mathcal{P}_s$), where the duals are block-separable. In particular, the convergence holds without the assumptions specific to distributed optimization problems, such as the assumption that $\ker(K) = \mathsf{Span}(\mathbf{1})$ (Assumption B.2), or R_s has to be nonnegative and $R_s^{-1}(\mathbf{0}) = \{\mathbf{0}\}$ (second part of Assumption B.3). In addition, Lemma 7 (and Lemma 11 below) can be used to show the convergence of $\sup_{Z \in \overline{\mathcal{Z}}} Q(Z^N; Z)$, with $D^X \geq \sup_{X \in \overline{\mathcal{X}}} D_{w^X}(X, X^{init})$. In Theorem 8 (and Theorem 12), we give the "weaker" convergence for a fixed $\widehat{X}$. This is due to our Lemmas 2, 3, and Corollary 5, which only require an upper bound on $\sup_{Y' \in \mathbb{R}^n} Q(Z; X^*, Y')$. In addition, as will be seen in

Section D.1, the dependence on $D_{w^X}(X^*, X^{init})$ rather than $\sup_{X\in\overline{\mathcal{X}}} D_{w^X}(X, X^{init})$ is crucial to obtain complexities which depend on function similarities for AMT-GS.

In addition, we point out that according to Theorme 8, $T = O(\frac{mM^2N}{\overline{r}(\sum_{s'=1}^S \widetilde{\kappa}_{s'}\sqrt{D_{s'}^y})^2})$, and so the total subgradient needed to find an ϵ suboptimal solution is

$$NT = O(\frac{mM^2N^2}{\overline{r}(\sum_{s'=1}^S \widetilde{\kappa}_{s'}\sqrt{D_{s'}^y})^2}) = O(\frac{\overline{r}mM^2D^X}{\epsilon^2}).$$

C.4. Communication round complexities for ($\mathcal{P}$)

With additional assumptions specific to distributed optimization, and with proper choices of R_s's, the duality gap for ($\mathcal{P}_s$) can be related to the suboptimality in terms of objective values F and/or violation of the consensus constraint for the original problem ($\mathcal{P}$). Next, we establish such connection.

We point out that with $w_s^y = \frac{1}{2}\|y_s\|_2^2$ and $y_s^0 = \mathbf{0}$, we can take $\sqrt{2D_s^y}$ as $\frac{\xi+a_s}{\sigma_{\min}^+(K_s)}$ for R_s^{ccv} in (6), as $\frac{(1+\xi)a_s}{\sigma_{\min}^+(K_s)}$ for R_s^{prj} in (7), as $\frac{\xi+\widehat{a}_1}{\sigma_{\min}^+(K)}$ for $\widehat{R}_s^{ccv}$ in (13), and as $\frac{(1+\xi)\widehat{a}_1}{\sigma_{\min}^+(K)}$ for $\widehat{R}_s^{prj}$ in (14).

Corollary 9 *Assume that all norms are the l_2 norm, and take $y_s^0 = \mathbf{0}$, $w_s^y(y_s) = \frac{1}{2}\|y_s\|^2$ and $w^x(x) = \frac{1}{2}\|x\|^2$. Assume that Assumptions B.1, B.2 and B.3 and the conditions of Theorem 8 hold, and $\{f_v\}_{v\in V}$ is $\widehat{a}_1$-similar. Take $\rho_s = \frac{\|K_s\|}{\sum_{s'=1}^S \|K_{s'}\|}$, $\overline{r} = \sum_{s=1}^S r_s\rho_s$, and assume $D_{w^X}(X^*, X^{init}) \le D^X$, then for*

$$N \ge \frac{2\sqrt{3}\overline{r}A\sqrt{D^X}}{\epsilon},$$

1. $\frac{1}{N+1}\sum_{k=0}^N \widehat{X}^k$ *is an $(\epsilon, \epsilon/\xi)$-solution if $R_s = \widehat{R}_s^{ccv}$ as defined in (13) and $A = \frac{(\sum_{s=1}^S \|K_s\|)}{\sigma_{\min}^+(K)} \cdot (\xi + \widehat{a}_1)$;*
2. $\Pi(\frac{1}{N+1}\sum_{k=0}^N \widehat{X}^k)$ *is an $(\epsilon(1+1/\xi), 0)$-solution if $R_s = \widehat{R}_s^{prj}$ as defined in (14) and $A = \frac{(1+\xi)(\sum_{s=1}^S \|K_s\|)\cdot\widehat{a}_1}{\sigma_{\min}^+(K)}$.*

Corollary 10 *Assume that all norms are the l_2 norm, and take $y_s^0 = \mathbf{0}$, $w_s^y(y_s) = \frac{1}{2}\|y_s\|^2$ and $w^x(x) = \frac{1}{2}\|x\|^2$. Assume that $K_sK_{s'}^* = \mathbf{0}$ for all $s, s' \in [S]$, that Assumptions B.1, B.2 and B.3 and the conditions of Theorem 8 hold, and that $\{f_v\}_{v\in V}$ is $\{(a_s, K_s)\}_{s\in[S]}$-similar. Take $\overline{r} = \sum_{s=1}^S r_s\rho_s$, and assume $D_{w^X}(X^*, X^{init}) \le D^X$, then for*

$$N \ge \frac{2\sqrt{3}\overline{r}A\sqrt{D^X}}{\epsilon},$$

1. $\frac{1}{N+1}\sum_{k=0}^N \widehat{X}^k$ *is an $(\epsilon, \epsilon/\xi)$-solution if $R_s = R_s^{ccv}$ as defined in (6), $\rho_s = (\frac{\xi+a_s}{\sigma_{\min}^+(K_s)})/(\sum_{s'=1}^S \frac{\xi+a_{s'}}{\sigma_{\min}^+(K_{s'})})$, and $A = \sum_{s=1}^S(\xi + a_s)\cdot\frac{\|K_s\|}{\sigma_{\min}^+(K_s)}$;*
2. $\Pi(\frac{1}{N+1}\sum_{k=0}^N \widehat{X}^k)$ *is an $(\epsilon(1+1/\xi), 0)$-solution if $R_s = R_s^{prj}$ satisfies (7), $\rho_s = (\frac{a_s}{\sigma_{\min}^+(K_s)})/(\sum_{s'=1}^S \frac{a_{s'}}{\sigma_{\min}^+(K_{s'})})$, and $A = (1+\xi)(\sum_{s=1}^S a_s\cdot\frac{\|K_s\|}{\sigma_{\min}^+(K_s)})$.*

Thus, the communication round N depends on $\overline{r}$, the weighted average of the rates at which the duals are updated, as well as A, which measures the function similarities.

C.5. Discussions

Below, we look at the communication rounds (N) and the subgradient oracle complexities (T and NT) of Algorithm 1 in order to find an ϵ suboptimal solution. In Section C.5.1, we focus on the saddle point formulation ($\mathcal{P}_s$), assuming that the regularization R is given. In Section C.5.2, we look at the original distributed optimization problem ($\mathcal{P}$), and choose R based on the discussion in Section B.3.

C.5.1. SADDLE POINT PROBLEM ($\mathcal{P}_s$)

Theorem 8 implies that to find an ϵ-suboptimal solution, one can take

$$N = O(\frac{\overline{r} \cdot (\sum_{s=1}^S \widetilde{\kappa}_s \sqrt{D_s^y}) \cdot \sqrt{D^X}}{\epsilon}), \quad T = O(\frac{M^2 m \sqrt{D^X}}{\epsilon \sum_{s=1}^S \widetilde{\kappa}_s \sqrt{D_s^y}}),$$

where $\overline{r} := \sum_{s=1}^S r_s \rho_s = \frac{\sum_{s=1}^S r_s \widetilde{\kappa}_s \sqrt{D_s^y}}{\sum_{s=1}^S \widetilde{\kappa}_s \sqrt{D_s^y}}$. In particular, the total rounds of communication is $O(\frac{\overline{r}}{\epsilon})$. For the subgradient, notice that $NT = O(\frac{\overline{r} M^2 m D^X}{\epsilon^2})$, which agrees with [24] for when $r_s = 1$ for all s.

To further illustrate the benefits of having different update frequencies for different duals, we analyze the "cost" of Algorithm 1. Precisely, we assume that every time y_s is updated, the cost, including all the communication between $\mathsf{Agent}(y_s)$ and $\mathsf{Agent}(x_v)$ for $K_{s,v} \neq \mathbf{0}$, together with the matrix-vector products involving $\{K_{s,v}\}_{v \in V}$, is $c_s \in [0, \infty]$. When multiple duals $\mathcal{S} \subset [S]$ are updated at the same time (in parallel), we assume that the total cost is additive, i.e. $\sum_{s \in \mathcal{S}} c_s$.

Then, with the above N, the dual variable y_s is updated $O(\frac{\overline{r}}{r_s \epsilon})$ times, which is different for duals with different r_s. Thus, suppose one is allowed to choose the update frequencies $\{r_s\}_{s \in [S]}$, to minimize the total cost to find an ϵ suboptimal solution, the following should be (approximately) minimized

$$O(\overline{r} \sum_{s=1}^S \frac{c_s}{r_s}) = O((\sum_{s=1}^S \rho_s r_s) \cdot (\sum_{s=1}^S \frac{c_s}{r_s})).$$

With $r_s \propto \sqrt{c_s/\rho_s}$[3], the above becomes $O((\sum_{s=1}^S \sqrt{c_s \rho_s})^2)$. As a comparison, the strategy where all $r_s = r_0'$ are the same has the cost $O(\sum_{s=1}^S c_s)$. By Cauchy–Schwarz inequality, $(\sum_{s=1}^S \sqrt{c_s \rho_s})^2 \leq \sum_{s=1}^S c_s$, and the difference can be very large when $\{c_s \rho_s\}_{s \in [S]}$ are very different, thereby showing the benefit of optimizing the updating rates $\{r_s\}_{s \in [S]}$ when $\{c_s / (\widetilde{\kappa}_s \sqrt{D_s^y})\}_{s \in [S]}$ are heterogeneous.

The additive cost is motivated by resources consumption when sending messages along each edge. In general, the total cost can be an arbitrary set function of the set of duals updated. For instance, to model time required to send messages (in parallel) where total time depends on the largest time, the cost could be $\max_{s \in \mathcal{S}} c_s$. In fact, when the costs $\{c_s\}_{s \in [S]}$ are differently significantly (e.g.

3. Here and below, $\propto$ means (approximately) proportional to, i.e. there exists $r_0 \in \mathbb{R}$ such that $r_s \approx r_0 \sqrt{c_s/\rho_s}$ for all $s \in [S]$.

by an order of magnitude), $\max_{s\in\mathcal{S}} c_s \approx \sum_{s\in\mathcal{S}} c_s$. Precisely, assume that there exists $\xi > 1$, such that for any $s \neq s' \in [S]$, either $c_s \leq \xi c_{s'}$ or $c_{s'} \leq \xi c_s$. In this case, for any $\mathcal{S} \subset [S]$,

$$\sum_{s\in\mathcal{S}} c_s \leq \Big(\sum_{l=0}^{|\mathcal{S}|-1} \xi^{-l}\Big) \cdot \max_{s\in\mathcal{S}} c_s \leq \frac{\xi}{\xi-1} \cdot \max_{s\in\mathcal{S}} c_s \leq \frac{\xi}{\xi-1} \cdot \sum_{s\in\mathcal{S}} c_s,$$

and so the additive cost is a constant approximation to the maximum.

C.5.2. DISTRIBUTED OPTIMIZATION PROBLEM ($\mathcal{P}$)

For the distributed optimization problem ($\mathcal{P}$), the more natural measure of performance is the objective value suboptimality and the consensus constraint violation, as defined in (3). As indicated by Corollaries 9 and 10, with good choices of the regularization R, $\frac{1}{N+1}\sum_{k=0}^{N} \widehat{X}^k$ or its projection using Π satisfies (3). In the rest of Section C.5.2, we assume that all norms are l_2 norms, and take $w^x(x) = \frac{1}{2}\|x\|^2$ and $w_s^y(x) = \frac{1}{2}\|y_s\|^2$.

Bounds using the Lipschitz constants. Consider the case where $\|f_v'\| \leq M_f$, and to guarantee that $\frac{1}{N+1}\sum_{k=0}^{N} \widehat{X}^k$ is an $(\epsilon, \epsilon/\xi)$-solution, in Corollary 9 we can take $\xi = \widehat{a}_1 = 2\sqrt{m}M_f$ which gives the following N_1, and in Corollary 10, we can take $\xi = a_s = 2\sqrt{m}M_f$ for all $s \in [S]$, which gives the following N_2:

$$N_1 = O\Big(\frac{\overline{r}M_f\sqrt{mD^X}}{\epsilon} \cdot \frac{\sum_{s=1}^{S}\|K_s\|}{\sigma_{\min}^+(K)}\Big), \quad N_2 = O\Big(\frac{\overline{r}M_f\sqrt{mD^X}}{\epsilon} \cdot \Big(\sum_{s=1}^{S} \frac{\|K_s\|}{\sigma_{\min}^+(K_s)}\Big)\Big).$$

Both N_1 and N_2 depend linearly in $\overline{r}$. However, when $\{K_s\}_{s\in[S]}$ are orthogonal, as discussed in Section B.3, $\sigma_{\min}^+(K) \leq \sigma_{\min}^+(K_s)$ for all s, and so in terms of the rounds of communication N, it appears that orthogonality allows a more refined (i.e. s-dependent) control over the decomposition of the function variation and thus the dual domain size, thereby achieving better convergence. In addition, similar to the argument in Section C.5.1, when the cost of updating y_s is c_s and total cost is additive, one should choose $r_s \propto \sqrt{c_s/\|K_s\|}$ when Corollary 9 holds, and $r_s \propto \sqrt{c_s/(\|K_s\|/\sigma_{\min}^+(K_s))}$ if Corollary 10 holds. Similar results hold for $\Pi(\frac{1}{N+1}\sum_{k=0}^{N} \widehat{X}^k)$ to be an $(\epsilon, 0)$-solution.

Bounds using the function similarity. In reality, sometimes the functions $\{f_v\}_{v\in V}$ exhibit similarity. For instance, in the extreme case $f_v = f_{v'}$ for all $v, v' \in V$, and thus communication is not needed at all! In that case, the bound on $\sup_{\widetilde{X}\in\overline{\mathcal{X}},\, K\widetilde{X}=\mathbf{0}} \|(I-\Pi)\nabla F(\widetilde{X})\|$ (and other terms using Π_s) using the Lipschitz constant M_f is too loose: in fact, one can choose $\widehat{a}_1 = a_s = \epsilon_0$ for all s for arbitrarily small $\epsilon_0 > 0$, then when R_s are set according to (14) or (7) with constant ξ, one only needs $N = O(\frac{\epsilon_0}{\epsilon}\sum_{s=1}^{S} r_s)$, which can be arbitrarily small.

More generally, choosing $\xi = 1$ and setting R_s according to (14) or (7), we obtain the following bound on the rounds of communication following Corollary 9 (N_3) and 10 (N_4)

$$N_3 = O\Big(\frac{\overline{r}\sqrt{D^X}}{\epsilon} \cdot \frac{\sum_{s=1}^{S}\|K_s\|}{\sigma_{\min}^+(K)} \cdot \widehat{a}_1\Big), \quad N_4 = O\Big(\frac{\overline{r}\cdot\sqrt{D^X}}{\epsilon} \cdot \Big(\sum_{s=1}^{S} \frac{a_s\|K_s\|}{\sigma_{\min}^+(K_s)}\Big)\Big).$$

Importantly, the number of rounds needed now depends on the *function similarity* instead of crude quantities such as Lipschitz constants.

In fact, when $S = 1$, such dependency is optimal. Indeed, [3] designs a pair of "chain like" functions $\{F_1, F_2\}$, such that for any $\gamma \geq 0$, $\{\gamma F_1, \gamma F_2\}$ is $\sqrt{1.5}\gamma$-similar. In addition, when $m/2$ agents are given γF_1 and the rest are give γF_2, finding an ϵ suboptimal x (in terms of the objective value) in the l_2 unit ball requires $\Omega(\frac{\gamma}{\epsilon/m})$ rounds of communication (see Theorem 2 and the discussions after it in [3]). For our algorithm, with $\widehat{a}_1 = O(\sqrt{m}\gamma)$, $D^x = 1/2$, and $K = I - \Pi$ (and so $\|K\| = \sigma^+_{\min}(K)$), we have $N_3 = O(\frac{r_1\gamma}{\epsilon/m})$. Thus, y_1 is updated only $N_3/r_1 = O(\frac{\gamma}{\epsilon/m})$ times, which is also the number of actual communication rounds needed. This achieves the theoretical lower bound, and so is optimal.

However, to achieve the above function-variation-dependent bounds, the parameters $\widehat{a}_1$ and $\{a_s\}_{s\in[S]}$ need to be set correctly. It is an interesting open question how one can achieve such dependence without additional prior knowledge (such as function similarity) about $\{f_v\}_{v\in V}$.

The hierarchical setting and function similarity at different scales. In addition, we provide results when function variations could be different along the span of K_s for different $s \in [S]$. As an example, consider the hierarchical setting discussed in Section B.2, with the additional assumption that for each non-leaf layer of the tree, all dual variables in that layer have the same number of child nodes. Then it can be shown that $\|K_s\| = \sigma^+_{\min}(K_s) = \sqrt{|\mathsf{Chi}(s)|/|\mathsf{Des}(s)|}$ (by (24) in the proof of Lemma 1), so the above bound N_4 can be simplified as

$$N_4' = O(\frac{\overline{r}\cdot(\sum_{s=1}^S a_s)\cdot\sqrt{D^X}}{\epsilon}), \quad \rho_s \propto a_s\sqrt{|\mathsf{Des}(s)|}.$$

As discussed in Section B.3.2, a_s measures the function variation along the span of K_s, i.e. variation in $\{f_v\}_{v\in\mathsf{Des}(s)}$ *not taken care of* by $\mathsf{Agent}(y_{s'})$ in the subtree rooted at $\mathsf{Agent}(y_s)$. In addition, (12) shows that $a_s^2 = |\mathsf{Des}(s)|\cdot\sup_{x\in\mathcal{X}}\mathsf{Var}_{i\sim\mu_s}(\overline{f'}_i(x))$, and so $\rho_s \propto |\mathsf{Des}(s)|\cdot\sqrt{\sup_{x\in\mathcal{X}}\mathsf{Var}_{i\sim\mu_s}(\overline{f'}_i(x))}$.

Thus, from the cost-minimization perspective in Section C.5.1, denoting the cost of updating y_s as c_s, one should choose $r_s \propto \sqrt{\frac{c_s/|\mathsf{Des}(s)|}{\sqrt{\sup_{x\in\mathcal{X}}\mathsf{Var}_{i\sim\mu_s}(\overline{f'}_i(x))}}}$. This corroborates the intuition that if along some K_s the function does not vary by too much ($\mathsf{Var}_{i\sim\mu_s}(\overline{f'}_i)$ is small), then $\mathsf{Agent}(y_s)$ does not need to update y_s very frequently (can use larger r_s).

Appendix D. Accelerated convergence under strong convexity

The convergence result in Section C.3 can be applied to objectives $\{f_v\}_{v\in V}$ where $\mu = 0$. In case strong convexity holds, i.e. $\mu > 0$, Algorithm 1 can achieve the accelerated convergence rate $O(1/N^2)$. We defer the proof of Lemma 11 to Appendix F.4.

Lemma 11 *Under Assumption B.4, further assume that $D_{w^x}(x, x') \leq \frac{C}{2}\|x-x'\|^2$ for all $x, x' \in \mathcal{X}$ for some $1 \leq C < \infty$, and $\mu > 0$. Let $\{\rho_s\}_{s\in[S]}$ be a distribution over $[S]$, $\overline{r} = \sum_{s=1}^S r_s\rho_s$ and similarly define $\overline{r^2}$ and $\overline{r^3}$.*

With $\alpha_{s,i_s} = 1$, $\theta_k = k + 2\overline{r^2}/\overline{r}$, $\eta_k = \frac{\mu}{2\overline{r}C}(k + \overline{r^2}/\overline{r})$, $\eta_{k,s} = \eta_k\rho_s$, $\tau_{s,i_s}(\sum_{k'=r_si_s}^{r_si_s+r_s-1}\theta_{k'}) = \tau_s = \frac{\widetilde{\kappa}_s^2}{\rho_s}\cdot\frac{4r_s\overline{r^2}C}{\mu}$, $T_k/N = T/N \geq \max(\frac{5}{\sqrt{D_1}}, \frac{64\overline{r}}{D_1})$ where $D_1 = \frac{\mu^2(\overline{r^2}/\overline{r})^2}{2M^2C^2}D_0$, assume that the CS procedure at iteration k is run with $\lambda_t = t$ and $\beta_t^k = \frac{(t+1)\mu}{2\eta_kC} + \frac{t-1}{2}$ for $t = 1, \ldots, T_k$, then for any

$Z \in \overline{\mathcal{Z}}$

$$Q(Z^N; Z) \leq \frac{2}{N(N+1)} \Bigg\{ \frac{\mu(\overline{r^3}/\overline{r} + 5(\overline{r^2}/\overline{r})^2)}{2C} D_{w^X}(X, X^{init}) + \frac{m\mu(\overline{r^2}/\overline{r})^2}{C} D_0 + \frac{4\overline{r}C}{\mu} \sum_{s=1}^{S} \frac{\widetilde{\kappa}_s^2 r_s}{\rho_s} D_{w_s^y}(y_s, y_s^{init}) \Bigg\}.$$

As a direct consequence, we have the following theorem.

Theorem 12 *For $\widehat{X} \in \overline{\mathcal{X}}$, assume that the following are finite:*

$$D_{w^X}(\widehat{X}, X^{init}) \leq D^X < \infty, \qquad \sup_{y_s \in \mathsf{dom}(R_s^*)} D_{w_s^y}(y_s, y_s^{init}) \leq D_s^y < \infty.$$

Under the conditions in Lemma 11, taking $\rho_s = \frac{\widetilde{\kappa}_s \sqrt{D_s^y}}{\sum_{s'=1}^{S} \widetilde{\kappa}_{s'} \sqrt{D_{s'}^y}}$ and $D_0 = D^X/m$,

$$\sup_{Y' \in \mathbb{R}^n} Q(Z^N; \widehat{X}, Y') \leq \frac{2}{N(N+1)} \left\{ \frac{\mu(\overline{r^3}/\overline{r} + 7(\overline{r^2}/\overline{r})^2)}{2C} D^X + \frac{4C(\overline{r})^2}{\mu} (\sum_{s'=1}^{S} \widetilde{\kappa}_{s'} \sqrt{D_{s'}^y})^2 \right\}.$$

Notice that (1) implies that $\|x - x'\| \leq \frac{M}{\mu}$ for all $x, x' \in \mathcal{X}$. Thus, one can take $D^X = O(\frac{mCM^2}{\mu^2})$. The resulting upper bound, when $\overline{r^3} = O((\overline{r})^3)$ and $\overline{r^2} = O((\overline{r})^2)$, becomes

$$\sup_{Y' \in \mathbb{R}^n} Q(Z^N; \widehat{X}, Y') = O(\frac{\overline{r}^2}{\mu N^2} \left\{ mM^2 + C(\sum_{s'=1}^{S} \widetilde{\kappa}_{s'} \sqrt{D_{s'}^y})^2 \right\}).$$

D.1. Good initialization for ($\mathcal{P}$)

In Theorem 12, assuming that $\mathcal{X}$ is compact, then one can always use $D^X \geq \sup_{X \in \overline{\mathcal{X}}} D_{w^X}(X, X^{init})$, suggesting that X^{init} should be chosen as the "center" of $\overline{\mathcal{X}}$, and D^X measures the (squared) radius of $\overline{\mathcal{X}}$. The resulting N, then, depends on D^X. However, such dependence on the size of $\overline{\mathcal{X}}$ could be suboptimal, especially when local objectives are similar. Indeed, in the extreme case where all local functions are the same, then primal agents can optimize their local objectives without communication at all.

To take advantage of potential similarities in the local functions, we propose initializing the primal variables at (approximate) local optimal solutions, which has the following guarantee on $D_{w^X}(X^*, X^{init})$.

Lemma 13 *Assume that all norms are the l_2 norm, and for some $\epsilon_0 \geq 0$, $\widehat{X} = (\widehat{x}_v)_{v \in V} \in \overline{\mathcal{X}}$ satisfies the following condition*

$$F(\widehat{X}) \leq \min_{X \in \overline{\mathcal{X}}} F(X) + \epsilon_0.$$

Assume that (1) holds for some $\mu > 0$ and Assumption B.1 holds, and $\{f_v\}_{v \in V}$ is $\{(a_s, K_s)\}_{s \in [S]}$-similar, then

$$\|\widehat{X} - X^*\| \leq \frac{(\sum_{s=1}^{S} a_s^2)^{1/2}}{\mu} + \sqrt{\frac{\sum_{s=1}^{S} a_s^2}{\mu^2} + \frac{2\epsilon_0}{\mu}}.$$

Proof [Proof of Lemma 13] By the suboptimality condition for $\widehat{X}$ and (1), we get

$$\frac{\mu}{2}\|\widehat{X}-X^*\|^2 \le F(\widehat{X})-F(X^*)-\langle \nabla F(X^*),\widehat{X}-X^*\rangle \le -\langle \nabla F(X^*),\widehat{X}-X^*\rangle+\epsilon_0.$$

Notice that by the first-order optimality condition of X^*, we get

$$\langle \nabla F(X^*),\Pi(\widehat{X}-X^*)\rangle \ge 0.$$

Combining the above two results, we get

$$\begin{aligned}
\frac{\mu}{2}\|\widehat{X}-X^*\|^2 &\le -\langle \nabla F(X^*),(I-\Pi)(\widehat{X}-X^*)\rangle+\epsilon_0\\
&= -\sum_{s=1}^{S}\langle \Pi_s\nabla F(X^*),\Pi_s(\widehat{X}-X^*)\rangle+\epsilon_0\\
&\le \sum_{s=1}^{S}\|\Pi_s\nabla F(X^*)\|_*\cdot\|\Pi_s(\widehat{X}-X^*)\|+\epsilon_0\\
&\le (\sum_{s=1}^{S}\|\Pi_s\nabla F(X^*)\|_*^2)^{1/2}\cdot(\sum_{s=1}^{S}\|\Pi_s(\widehat{X}-X^*)\|^2)^{1/2}+\epsilon_0\\
&\le (\sum_{s=1}^{S}a_s^2)^{1/2}\cdot\|\widehat{X}-X^*\|+\epsilon_0,
\end{aligned}$$

where the last $\le$ is because of the assumption that $\{f_v\}_{v\in V}$ are $\{(a_s,K_s)\}_{s\in[S]}$-similar, and all norms are l_2 norm. The above inequality is quadratic in $\|\widehat{X}-X^*\|$, and the result follows. ■

The above Lemma 13 shows that if $\{\widehat{x}_v\}_{v\in V}$ are all approximately optimal to local objectives, then $D_{w^X}(X^*,\widehat{X})\sim\frac{\sum_{s=1}^{S}a_s^2}{\mu^2}$. To find such initialization, one can apply the CS procedure.

Corollary 14 *Assume that all norms are l_2 norms and $w^x(x)=\frac{1}{2}\|x\|^2$, that (1) holds with some $\mu>0$, and that Assumption B.1 holds. For each $v\in V$, assume that* $\mathsf{Agent}(x_v)$ *is given some $\underline{x}_v^0\in\mathcal{X}$ such that $\sup_{x\in\mathcal{X}}D_{w^x}(x,\underline{x}_v^0)\le\underline{D}^x<\infty$*

$$(\_,x_v^{init})=CS(f_v,\mathcal{X},D_{w^x},\underline{T},\underline{\eta},\mathbf{0},\underline{x}_v^0,\underline{x}_v^0),$$

where the CS procedure uses λ_t,β_t according to Corollary 6 (for $\mu>0$), then with $\epsilon_0=\widetilde{a}^2/\mu$, $\underline{T}\ge\frac{8CM^2m}{\epsilon_0\mu}$ and $\underline{\eta}=\frac{\epsilon_0/2}{m\underline{D}^x}$

$$D_{w^X}(X^*,X^{init})\le\frac{4\widetilde{a}^2}{\mu^2},$$

where $\widetilde{a}=\widehat{a}_1$ if $\{f_v\}_{v\in V}$ is $\widehat{a}_1$-similar for some $\widehat{a}_1>0$, and $\widetilde{a}=(\sum_{s=1}^{S}a_s^2)^{1/2}$ if $\{f_v\}_{v\in V}$ is $\{(a_s,K_s)\}_{s\in[S]}$-similar such that $(\sum_{s=1}^{S}a_s^2)^{1/2}>0$.

D.2. Complexities for ($\mathcal{P}$) and discussions

Recall that with $w_s^y = \frac{1}{2}\|y_s\|_2^2$ and $y_s^0 = \mathbf{0}$, we can take $\sqrt{2D_s^y}$ as $\frac{\xi+a_s}{\sigma_{\min}^+(K_s)}$ for R_s^{ccv} in (6), as $\frac{(1+\xi)a_s}{\sigma_{\min}^+(K_s)}$ for R_s^{prj} in (7), as $\frac{\xi+\widehat{a}_1}{\sigma_{\min}^+(K)}$ for $\widehat{R}_s^{ccv}$ in (13), and as $\frac{(1+\xi)\widehat{a}_1}{\sigma_{\min}^+(K)}$ for $\widehat{R}_s^{prj}$ in (14). Now, combining Theorem 12, Corollary 14, Lemma 3, and Corollary 5, we get the following results.

Corollary 15 *Assume that all norms are the l_2 norm, and take $y_s^0 = \mathbf{0}$, $w_s^y(y_s) = \frac{1}{2}\|y_s\|^2$ and $w^x(x) = \frac{1}{2}\|x\|^2$. Assume that Assumptions B.1, B.2 and B.3 and the conditions of Theorem 12 and Corollary 14 hold, and $\{f_v\}_{v\in V}$ is $\widehat{a}_1$-similar for some $\widehat{a}_1 > 0$, and take $\rho_s = \frac{\|K_s\|}{\sum_{s'=1}^S \|K_{s'}\|}$. In addition, assume that X^{init} is initialized according to Corollary 14, then we have*

$$\sup_{Y'\in\mathbb{R}^n} Q(Z^N; X^*, Y') \le \frac{4(\overline{r})^2}{\mu N^2}\left\{(\overline{r^3}/(\overline{r})^3 + 7(\overline{r^2}/(\overline{r})^2)^2)\cdot\widehat{a}_1^2 + A_0^2\right\},$$

and so for

$$N \ge \frac{2\overline{r}A}{\sqrt{\mu\epsilon}}, \quad A = \sqrt{\overline{r^3}/(\overline{r})^3 + 7(\overline{r^2}/(\overline{r})^2)^2}\cdot\widehat{a}_1 + A_0,$$

1. $\frac{\sum_{k=0}^N \theta_k \widehat{X}^k}{\sum_{k=0}^N \theta_k}$ *is an $(\epsilon, \epsilon/\xi)$-solution if $R_s = \widehat{R}_s^{ccv}$ as defined in (13) and $A_0 = \frac{(\sum_{s=1}^S \|K_s\|)}{\sigma_{\min}^+(K)}\cdot(\xi+\widehat{a}_1)$;*
2. $\Pi(\frac{\sum_{k=0}^N \theta_k \widehat{X}^k}{\sum_{k=0}^N \theta_k})$ *is an $(\epsilon(1+1/\xi), 0)$-solution if $R_s = \widehat{R}_s^{prj}$ as defined in (14) and $A_0 = \frac{(1+\xi)(\sum_{s=1}^S \|K_s\|)\cdot\widehat{a}_1}{\sigma_{\min}^+(K)}$.*

Corollary 16 *Assume that all norms are the l_2 norm, and take $y_s^0 = \mathbf{0}$, $w_s^y(y_s) = \frac{1}{2}\|y_s\|^2$ and $w^x(x) = \frac{1}{2}\|x\|^2$. Assume that $K_s K_{s'}^* = \mathbf{0}$ for all $s, s' \in [S]$, that Assumptions B.1, B.2 and B.3 and the conditions of Theorem 12 and Corollary 14 hold, and that $\{f_v\}_{v\in V}$ is $\{(a_s, K_s)\}_{s\in[S]}$-similar where $a_s > 0$ for all s. In addition, assume that X^{init} is initialized according to Corollary 14, then we have*

$$\sup_{Y'\in\mathbb{R}^n} Q(Z^N; X^*, Y') \le \frac{4(\overline{r})^2}{\mu N^2}\left\{(\overline{r^3}/(\overline{r})^3 + 7(\overline{r^2}/(\overline{r})^2)^2)\cdot(\sum_{s=1}^S a_s^2) + A_0^2\right\},$$

and so for

$$N \ge \frac{2\overline{r}A}{\sqrt{\mu\epsilon}}, \quad A = \sqrt{\overline{r^3}/(\overline{r})^3 + 7(\overline{r^2}/(\overline{r})^2)^2}\cdot(\sum_{s=1}^S a_s^2)^{1/2} + A_0,$$

1. $\frac{\sum_{k=0}^N \theta_k \widehat{X}^k}{\sum_{k=0}^N \theta_k}$ *is an $(\epsilon, \epsilon/\xi)$-solution if $R_s = R_s^{ccv}$ as defined in (6), $\rho_s = (\frac{\xi+a_s}{\sigma_{\min}^+(K_s)})/(\sum_{s'=1}^S \frac{\xi+a_{s'}}{\sigma_{\min}^+(K_{s'})^+})$, and $A_0 = \sum_{s=1}^S(\xi+a_s)\cdot\frac{\|K_s\|}{\sigma_{\min}^+(K_s)}$;*
2. $\Pi(\frac{\sum_{k=0}^N \theta_k \widehat{X}^k}{\sum_{k=0}^N \theta_k})$ *is an $(\epsilon(1+1/\xi), 0)$-solution if $R_s = R_s^{prj}$ satisfies (7), $\rho_s = (\frac{a_s}{\sigma_{\min}^+(K_s)})/(\sum_{s'=1}^S \frac{a_{s'}}{\sigma_{\min}^+(K_{s'})^+})$, and $A_0 = (1+\xi)(\sum_{s=1}^S a_s\cdot\frac{\|K_s\|}{\sigma_{\min}^+(K_s)})$.*

Subgradient oracle complexities. With the initialization in Corollary 14 and $C = 1$, the number of subgradient steps needed to find X^{init} is $\underline{T} \geq \frac{8mM^2}{\widetilde{a}^2}$, constant in ϵ.

In addition, in Theorem 12, we can take $D_0 = \frac{4\widetilde{a}^2}{\mu^2}$, and so $D_1 = \frac{2\widetilde{a}^2(\overline{r^2}/\overline{r})^2}{M^2}$. Thus, the requirement on T becomes $T/N \geq \max(\frac{5}{\sqrt{D_1}}, \frac{64\overline{r}}{D_1})$, i.e. $T/N = \Omega(\max(\frac{M/\widetilde{a}}{\overline{r^2}/\overline{r}}, (\frac{M/\widetilde{a}}{\overline{r^2}/\overline{r}})^2 \cdot \overline{r}))$, and so the total subgradient steps needed (for each agent) is

$$N^2 \cdot O(\max(\frac{M/\widetilde{a}}{\overline{r^2}/\overline{r}}, (\frac{M/\widetilde{a}}{\overline{r^2}/\overline{r}})^2 \cdot \overline{r})) = O(\frac{\overline{r}^2 A^2}{\mu\epsilon} \cdot \max(\frac{M/\widetilde{a}}{\overline{r^2}/\overline{r}}, (\frac{M/\widetilde{a}}{\overline{r^2}/\overline{r}})^2 \cdot \overline{r}))$$

In the special case where $S = 1$ and $\|K\| = O(\sigma^+_{\min}(K))$, we have $\widetilde{a} = \Omega(A)$. Further assuming that $A = O(\sqrt{m}M)$ (which holds when $\|f'_v\| \leq M_f$ for all v and $M = 2M_f$), the above can be simplified as $\frac{\overline{r}\sqrt{m}M^2}{\mu\epsilon}$.

Communication rounds complexities. For both Corollaries 15 and 16, the communication rounds needed is $N = O(\frac{\overline{r}A}{\sqrt{\mu\epsilon}})$, where A depends on function similarities and higher moments of $\{r_s\}_{s\in[S]}$. In terms of N and T's dependency on ϵ, μ, $O(1/\sqrt{\mu\epsilon})$ communication rounds and $O(1/(\mu\epsilon))$ gradient steps are needed.

Now, consider the special case where $\overline{r^2} = O((\overline{r})^2)$ and $\overline{r^3} = O((\overline{r})^3)$, i.e. r_s has small variation, then $\sqrt{\overline{r^3}/(\overline{r})^3 + 7(\overline{r^2}/(\overline{r})^2)^2} = O(1)$, and so $A = O(\widetilde{a} + A_0)$, then $\xi = 1$ with $\widehat{R}^{prj}_s$ and R^{prj}_s give the following N_1 and N_2 respectively

$$N_1 = O\left(\frac{\overline{r}\widehat{a}_1}{\sqrt{\epsilon\mu}} \cdot \frac{\sum_{s=1}^S \|K_s\|}{\sigma^+_{\min}(K)}\right), \quad N_2 = O\left(\frac{\overline{r}}{\sqrt{\epsilon\mu}} \cdot \left(\sum_{s=1}^S a_s \cdot \frac{\|K_s\|}{\sigma^+_{\min}(K_s)}\right)\right),$$

both have linear dependence on $\overline{r}$ and function similarities. (For N_2, we use $(\sum_{s=1}^S a_s^2)^{1/2} \leq \sum_{s=1}^S a_s$ and $\|K_s\| \geq \sigma^+_{\min}(K_s)$.)

Comparison with communication lower bounds. When $S = 1$ (and so $\overline{r^2} = (\overline{r})^2$), $K = I - \Pi$ (and so $\|K\| = \sigma^+_{\min}(K)$), with $w^x(x) = \frac{1}{2}\|x\|^2$ and $w^y_s(y_s) = \frac{1}{2}\|y_s\|^2$, assuming that $\|f'_v\| \leq M_f$ (and take $M = 2M_f$), $a_1^2 \leq m\gamma^2$ for some $\gamma \leq M_f$, we have $N = O(\frac{\overline{r}\cdot\sqrt{m}\gamma}{\sqrt{\epsilon\mu}})$. Since communication is only needed when $\mathsf{Agent}(y_1)$ updates, the total number of communication rounds is $N/r_s = O(\frac{\sqrt{m}\gamma}{\sqrt{\epsilon\mu}})$, which achieves the theoretical lower bound (Theorem 2 and discussion after in [3]) on the communication round complexity for μ-strongly convex, $\sqrt{m}\gamma$-similar functions, and so is optimal[4].

Appendix E. Numerical experiments: support vector machine

Setup. We take $w^x(x) = \frac{1}{2}\|x\|^2$ and $w^y_s(y_s) = \frac{1}{2}\|y_s\|^2$, and

$$f'_v(x) = -\frac{1}{m_s}\sum_{l=1}^{m_s} y^l_v b^l_v \cdot \mathbf{1}[1 > y^l_v\langle b^l_v, x\rangle] + \mu x.$$

4. [3] constructs a pair of "chain like" functions $\{\gamma F_1, \gamma F_2\}$ which are $\Theta(\gamma)$-similar and μ-strongly convex. In addition, when $m/2$ agents are given γF_1 and the rest are given γF_2 (and so this set of m functions is $\Theta(\sqrt{m}\gamma)$-similar), the number of rounds of communication needed is $\Omega(\gamma\sqrt{\frac{1}{\mu\epsilon/m}})$.

Since $\|b_v^l\| \leq 1$ and $\|x\| \leq 5$ (since $x \in \mathcal{X}$), we have $\|f_v'\| \leq 1 + 5\mu$ and so we can take $M = 2(1 + 5\mu)$.

Below, we look at the suboptimality of $F(\Pi \underline{X}^k)$ where $\underline{X}^k := \frac{\sum_{k'=0}^k \theta_{k'} \widehat{X}^{k'}}{\sum_{k'=0}^k \theta_{k'}}$. In Section E.1, we focus on how $F(\Pi \underline{X}^k)$ depends on the iteration k and the mean updating rates $\overline{r}$. In Section E.2, we focus on how $F(\Pi \underline{X}^k)$ depends on the iteration k and function similarities A.

E.1. Suboptimality and $k, \overline{r}$

In this experiment, we investigate the suboptimality of $F(\Pi \underline{X}^k)$ as a function of the iteration number k and the mean updating rates for the dual $\overline{r}$.

Data simulation. We first sample $x^* \in \mathbb{R}^d$ on the unit sphere uniformly at random as the "true solution". For each agent $\mathsf{Agent}(x_v)$, we generate the feature vectors b_v^l i.i.d. uniformly on the unit sphere and take $y_v^l = \mathsf{sign}(\langle b_v^l, x^* \rangle)$ w.p. $1 - p$, and $y_v^l = -\mathsf{sign}(\langle b_v^l, x^* \rangle)$ otherwise, independent of $\{b_v^l\}_{v \in V, l \in [m_s]}$. We take $m_s = 50$ and $p = 0.05$.

Communication setup. We consider the hierarchical setup, with 3 layers of dual agents and 1 layer of primal agents, where each non-leaf node has 5 child nodes. Thus, there are $m = 125$ primal agents and $S = 31$ dual agents, and $|\mathsf{Chi}(s)| = 5$ for each non-leaf node. We assume that the dual agents at layer i are updated with rate r_i for $i = 1, 2, 3$, and test with various (r_1, r_2, r_3)

Algorithm setup. Notice that the μx part in f_v' is common to all v so we can take $a_s = 2\sqrt{|\mathsf{Des}(s)|}$ for all s. We use R_s^{prj} as defined in (7) with $\xi = 1$, and set $x_v^{init} = \mathbf{0}$ and $y_s^{init} = \mathbf{0}$. We test MT-GS for $N + 1 = 3000$ and $\mu = 0$. All parameters are set according to Theorem 8. We test AMT-GS for $N + 1 = 500$ and $\mu = 0.01$. We set $T = N + 1$[5] and set all other parameters according to Theorem 12.

Results: $F(\Pi \underline{X}^k)$ and k. In Figures 2 and 4, we present $F(\Pi \underline{X}^k)$ as a function of k for MT-GS and AMT-GS respectively. Different lines correspond to different $(r_1, r_2, r_3, \overline{r})$, with the line colors indicating $\overline{r}$. As can be seen, for both MT-GS and AMT-GS, all settings of $(r_1, r_2, r_3, \overline{r})$ converge or show trend of convergence, and the convergence is faster for smaller $\overline{r}$. In addition, comparing Figures 2 and 4, we see that strong convexity (with AMT-GS) indeed accelerates the convergence.

Results: $F(\Pi \underline{X}^k)$ and $\overline{r}$. In Figure 3, we present $F(\Pi \underline{X}^k)$ as a function of $\overline{r}$, taken at $k = 400, 800, \ldots, 2800$ for MT-GS. In Figure 5, we present $F(\Pi \underline{X}^k)$ as a function of $\overline{r}^2$, taken at $k = 60, 120, \ldots, 480$ for AMT-GS. In both figures, the line colors indicate k. As can be seen, as k increases, the suboptimaltiy of $F(\Pi \underline{X}^k)$ is approximately linear in $\overline{r}$ for MT-GS and is approximately linear in $\overline{r}^2$ for AMT-GS, agreeing with our Theorem 8 and 12.

E.2. Suboptimality and k, A

In this experiment, there is only one dual agent (thus $A = \Theta(a_1)$) who updates at rate $r_1 = 1$. We focus on (normalized) $F(\Pi \underline{X}^k)$ as a function of k and function similarities a_1.

Dataset induced function similarities. In this experiment, similarities between $\{f_v'\}_{v \in V}$ are inherited from similarities in the local datasets. More precisely, we generate a global dataset $\{(b_{global}^l, y_{global}^l)\}_{l \in [m_{global}]}$ consisting of m_{global} pairs of data. In addition, for each agent, we generate m_{local} pairs of private data $\{(b_{local,v}^l, y_{local,v}^l)\}_{l \in [m_{local}]}$. Thus, $\mathsf{Agent}(x_v)$ has access to $\{(b_{global}^l, y_{global}^l)\}_{l \in [m_{global}]} \cup \{(b_{local,v}^l, y_{local,v}^l)\}_{l \in [m_{local}]}$, a total of $m_s = m_{global} + m_{local}$ pairs of data.

5. According to Lemma 11, T should be linear in N, and in this experiment, we set it as $N + 1$ for simplicity.

Data simulation. We simulate the x^* and all pairs of data the same as in Section E.1, with $m_s = 50$, $m_{local} = \lfloor \gamma m_s \rfloor$ and $m_{global} = m_s - m_{local}$, for $\gamma = 0.1, 0.2, \ldots, 0.9$. We assume that there are $m = 500$ primal agents. Due to the global dataset, we have $\|f'_v - f'_{v'}\| \leq \frac{m_{local}}{m_s}$ for any $v, v' \in V$, and so we can take $a_1 = 2\gamma\sqrt{m}$.

Algorithm setup. We consider two setups.

1. Type-0 setup. We use $R_s = R_s^{prj}$ as defined in (7) with $\xi = 1$, and $y_s^{init} = \mathbf{0}$. We set the parameter $T = N + 1$[6] and all other parameters are set according to Theorems 8 and 12. For the initialization, for MT-GS, we use $x_v^{init} = \mathbf{0}$ and for AMT-GS, we use $\underline{x}_v^0 = \mathbf{0}$ and construct x_v^{init} according to Corollary 14.

2. Type-1 setup. In addition to the above γ-aware setup, we also test our algorithms for $R_s = 100R_s^{prj}$, $a_1 = 2\sqrt{m}$, and $x_v^{init} = \mathbf{0}$, which we denote as type-1 setup. Compared to type-0, type-1 has larger dual domain size and ignores the function similarities, making it a closer approximation to the DCS algorithms in [24].

For both types of setups, we test MT-GS for $N+1 = 500$ and $\mu = 0$ and AMT-GS for $N+1 = 200$ and $\mu = 0.01$. Since the datasets are different for different γ, below, for each γ, we normalize $F(\Pi \underline{X}^k)$ such that $F(\mathbf{0}) = m = 500$ is normalized to 1, and the minimum (over k and two types) of $F(\Pi \underline{X}^k)$ is normalized to 0.

Results: normalized $F(\Pi \underline{X}^k)$ and k. In Figures 6 and 8, we present the normalized $F(\Pi \underline{X}^k)$ as a function of k for MT-GS and AMT-GS respectively. Different lines correspond to different types of setup and different γ, with the line colors indicating γ. As can be seen, our MT-GS and AMT-GS converge in all the tested settings, and strong convexity (with AMT-GS) accelerates the convergence.

Moreover, from Figure 6, solid curves are always below dotted curves of the same γ (color), indicating that type-0 setups converge faster than type-1 setups. In addition, solid curves corresponding to smaller γ (functions are more similar) are dominated by curves corresponding to larger ones, while dotted curves converge at roughly the same rates. These suggest that type-0 setups, with the γ-aware domain sizes and parameters, indeed take advantage of the function similarities to speed up the convergence.

Results: normalized $F(\Pi \underline{X}^k)$ and γ. In Figure 7, we present normalized $F(\Pi \underline{X}^k)$ as a function of $\overline{\gamma}$, taken at $k = 100, 200, 300, 400$ for MT-GS. The line colors indicate k. As can be seen, the normalized $F(\Pi \underline{X}^k)$ is approximately linear in γ for MT-GS, which agrees with our Theorem 8.

In Figure 9, we present normalized $F(\Pi \underline{X}^k)$ as a function of γ^2, taken at $k = 10, 15, 20, 25, 30, 40, 80, 120, 160$ for AMT-GS. The line colors indicate k. It appears that the normalized $F(\Pi \underline{X}^k)$ is increasing in γ^2 for most γ (the dip when $\gamma^2 = 0.36$ could be due to the randomness in the simulated dataset). This confirms that our AMT-GS can take advantage of function similarities. However, the normalized $F(\Pi \underline{X}^k)$ does not appear to be linear in γ^2: for small γ^2, it does not converge to 0, which could be because setting $T = N + 1$ (smaller than suggested) introduces additional suboptimality; in addition, it is possible that the normalization process introduces extra γ-dependent factors.

6. According to Theorems 8 and 12, T should be linear in N, and we set $T = N + 1$ for simplicity.

Appendix F. Additional proof

F.1. Proof for Section B

Proof [Proof of Lemma 1] For $d = 1$, the matrix representation of $K_s \in \mathbb{R}^{|\mathsf{Chi}(s)|\times m}$ is $K_s = (I - \mathbf{1}(\frac{|\mathsf{Des}(j)|}{|\mathsf{Des}(s)|})^T_{j\in\mathsf{Chi}(s)})P_s$ where $P_s \in \mathbb{R}^{|\mathsf{Chi}(s)|\times m}$, and $P_s(i,j) = |\mathsf{Des}(i)|^{-1}$ if $j \in \mathsf{Des}(i)$ and $P_s(i,j) = 0$ otherwise. Notice that

$$K_s K_{s'}^* = (I - \mathbf{1}(\frac{|\mathsf{Des}(j)|}{|\mathsf{Des}(s)|})^T_{j\in\mathsf{Chi}(s)})P_s P_{s'}^T (I - \mathbf{1}(\frac{|\mathsf{Des}(j)|}{|\mathsf{Des}(s')|})^T_{j\in\mathsf{Chi}(s')})^T.$$

If s is not in the subtree rooted at s' and s' is not in the subtree rooted at s, then $\mathsf{Des}(s)\cap\mathsf{Des}(s') = \emptyset$, and so $P_s P_{s'}^T = \mathbf{0}$. If s is in the subtree rooted at s', then s is in the subtree rooted at some $\widehat{s} \in \mathsf{Chi}(s')$. In particular, $(P_s P_{s'}^T)(i,j) = 0$ for all $j \neq \widehat{s}$ and $(P_s P_{s'}^T)(i,\widehat{s}) = |\mathsf{Des}(\widehat{s})|^{-1}$, and thus $(I - \mathbf{1}(\frac{|\mathsf{Des}(j)|}{|\mathsf{Des}(s)|})^T_{j\in\mathsf{Chi}(s)})P_s P_{s'}^T = \mathbf{0}$. Similarly for the case when s' is in the subtree rooted at s. The case when $d > 1$ follows by applying the above argument coordinate-wise.

For the second claim, consider the case $d = 1$, denoting

$$D = \mathsf{diag}((|\mathsf{Des}(j)|)_{j\in\mathsf{Chi}(s)}), \quad v = (|\mathsf{Des}(j)|)_{j\in\mathsf{Chi}(s)}/\|(|\mathsf{Des}(j)|)_{j\in\mathsf{Chi}(s)}\|_2,$$

where the norm in the denominator in the definition of v is the l_2 norm. Then when $d = 1$, we have (applying Theorem 6 in [26])

$$K_s K_s^* = D^{-1} - \frac{1}{|\mathsf{Des}(s)|}\mathbf{1}\mathbf{1}^T, \quad (K_s K_s^*)^\dagger = D - vv^T D - Dvv^T + (v^T Dv)vv^T. \tag{24}$$

Thus, noticing that $v^T K_s = \mathbf{0}$, we have

$$\Pi_s = K_s^*(K_s K_s^*)^\dagger K_s = K_s^T D K_s.$$

Thus, for any $\widetilde{X}, \widehat{X} \in \mathbb{R}^m$

$$\langle \widehat{X}, \Pi_s \widetilde{X}\rangle = \langle \Pi_s \widehat{X}, \Pi_s \widetilde{X}\rangle = (K_s\widehat{X})^T D(K_s\widetilde{X}) = \sum_{i\in\mathsf{Chi}(s)} |\mathsf{Des}(i)| \cdot \langle (K_s\widehat{X})_i, (K_s\widetilde{X})_i\rangle.$$

The above argument can be applied coordinate-wise, and so extend to $d \geq 1$. ■

F.2. Proofs for Section C.2

Lemma 17 (generalized lemma 5 in [24]) *Let the convex function $q : U \to \mathbb{R}$, and $\mathcal{I}$ an arbitrary finite index set. Assume that the points $x_i \in U$ and the numbers $\eta_i \geq 0$ for $i \in \mathcal{I}$. Let $w : U \to \mathbb{R}$ be a distance generating function and*

$$u^* \in \operatorname*{argmin}_{u\in U} q(u) + \sum_{i\in\mathcal{I}} \eta_i D_w(u, x_i).$$

Then for any $u \in U$, we have

$$q(u^*) + \sum_{i\in\mathcal{I}} \eta_i D_w(u^*, x_i) \leq q(u) + \sum_{i\in\mathcal{I}} \eta_i D_w(u, x_i) - \sum_{i\in\mathcal{I}} \eta_i D_w(u, u^*).$$

Proof [Proof of Lemma 17] First, by the optimality condition for u^*, there exists $q'(u^*) \in \partial q(u^*)$ such that

$$\langle q'(u^*) + \sum_{i\in\mathcal{I}} \eta_i \nabla D_w(u^*, x_i), u - u^* \rangle \geq 0, \quad \forall u \in U.$$

By definition, we have for each $i \in \mathcal{I}$ that

$$D_w(u, x_i) - D_w(u^*, x_i) - D_w(u, u^*) = \langle \nabla w(x_i) - \nabla w(u^*), u - u^* \rangle = -\langle \nabla D_w(u^*, x_i), u - u^* \rangle$$

Thus, we have for any $u \in U$,

$$\begin{aligned}
&q(u) + \sum_{i\in\mathcal{I}} \eta_i D_w(u, x_i) \\
&\geq q(u^*) + \langle q'(u^*), u - u^* \rangle + \sum_{i\in\mathcal{I}} \eta_i \left(D_w(u^*, x_i) + D_w(u, u^*) - \langle \nabla D_w(u^*, x_i), u - u^* \rangle \right) \\
&\geq q(u^*) + \sum_{i\in\mathcal{I}} \eta_i D_w(u^*, x_i) + \sum_{i\in\mathcal{I}} \eta_i D_w(u, u^*).
\end{aligned}$$

■

Lemma 18 *Assume that $U \subset \mathbb{R}^{d_0}$ is a convex set, and $\phi : U \to \mathbb{R}$ is a convex function such that*

$$\frac{\mu}{2}\|x - y\|^2 \leq \phi(x) - \phi(y) - \langle \phi'(y), x - y \rangle \leq M\|x - y\|, \quad \forall x, y \in U,$$

where $\phi' : U \to \mathbb{R}^{d_0}$ is a subgradient oracle, i.e. for each $y \in U$, $\phi'(y) \in \partial\phi(y)$ is a subgradient. In addition, $D_{w^x}(x, x') \leq \frac{C}{2}\|x - x'\|^2$ for some $C \in [0, \infty]$. If $\{\beta_t\}$ and $\{\lambda_t\}$ in Algorithm 2 satisfies that

$$\lambda_{t+1}(\eta\beta_{t+1} - \mu/C) \leq \lambda_t(1 + \beta_t)\eta, \quad \forall t \geq 1,$$

then for $t \geq 1$ and $u \in U$

$$(\sum_{t=1}^{T} \lambda_t) \cdot (\Phi(\widehat{u}^T) - \Phi(u)) \leq (\eta\beta_1 - \mu/C)\lambda_1 D_w(u, u^0) - \eta(1 + \beta_T)\lambda_T D_w(u, u^T) + \sum_{t=1}^{T} \frac{M^2\lambda_t}{2\eta\beta_t}.$$

Proof [Proof of Lemma 18]

Applying Lemma 17, and using $\sum_{i\in\mathcal{I}} \eta_i = \eta$, we have

$$\begin{aligned}
&\langle v + \phi'(u^{t-1}), u^t - u \rangle + \sum_{i\in\mathcal{I}} \eta_i D_w(u^t, x_i) - \sum_{i\in\mathcal{I}} \eta_i D_w(u, x_i) \\
&\leq \eta\beta_t D_w(u, u^{t-1}) - \eta\beta_t D_w(u^t, u^{t-1}) - (1 + \beta_t)\eta D_w(u, u^t)
\end{aligned}$$

The rest follows a similar argument as in the proof of Proposition 2 [24].

■

F.3. Proofs for Section C.3

Proof [Proof of Lemma 7] For convenience, denote $\eta_k = \sum_{s=1}^S \eta_{k,s}$ and $\overline{r} = \sum_{s=1}^S r_s\rho_s$.

Primal update properties. By Corollary 6, for any $v \in V$ and $x_v \in \mathcal{X}$

$$\begin{aligned}
&\langle \sum_{s=1}^S K_{s,v}^* \overline{y}_s^k, \widehat{x}_v^k - x_v\rangle + f_v(\widehat{x}_v^k) - f_v(x_v)\\
\le\ & \frac{2\eta_k}{T_k(T_k+3)} D_{w^x}(x_v, x_v^{k-1}) + \sum_{s=1}^S \eta_{k,s} D_{w^x}(x_v, x_v^{k-r_s})\\
&- \frac{(T_k+1)(T_k+2)}{T_k(T_k+3)}\eta_k D_{w^x}(x_v, x_v^k) - \sum_{s=1}^S \eta_{k,s} D_{w^x}(\widehat{x}_v^k, x_v^{k-r_s}) + \frac{4M^2}{\eta_k(T_k+3)}.
\end{aligned}$$

Summing over k, and defining $\eta_{k,s} = 0$ for all $k < 0$ and $k \ge N+1$, we have

$$\begin{aligned}
&\sum_{k=0}^N \langle \sum_{s=1}^S K_{s,v}^* \overline{y}_s^k, \widehat{x}_v^k - x_v\rangle + f_v(\widehat{x}_v^k) - f_v(x_v)\\
\le\ & \sum_{k=0-\max\{r_s\}}^N \left(\frac{2\eta_{k+1}}{T_{k+1}(T_{k+1}+3)} + \sum_{s=1}^S \eta_{k+r_s,s}\right) D_{w^x}(x_v, x_v^k)\\
&- \sum_{k=0}^N \frac{(T_k+1)(T_k+2)}{T_k(T_k+3)}\eta_k D_{w^x}(x_v, x_v^k)\\
&- \sum_{k=0}^N \sum_{s=1}^S \eta_{k,s} D_{w^x}(\widehat{x}_v^k, x_v^{k-r_s}) + \sum_{k=0}^N \frac{4M^2}{\eta_k(T_k+3)}.
\end{aligned}$$

Recall that $x_v^k = x_v^{init}$ for all $k < 0$, and so with $T_k = T$ for all k and $\eta_{k,s} = \eta\rho_s$ for $k = 0, 1, \ldots, N$,

$$\sum_{k=0-\max\{r_s\}}^{-1} \left(\frac{2\eta_{k+1}}{T(T+3)} + \sum_{s=1}^S \eta_{k+r_s,s}\right) \le \eta\left(\frac{2}{T(T+3)} + \sum_{s=1}^S r_s\rho_s\right) = \eta\left(\frac{2}{T(T+3)} + \overline{r}\right).$$

For $k = 0, 1, \ldots, N-1$

$$\frac{2\eta_{k+1}}{T_{k+1}(T_{k+1}+3)} + \sum_{s=1}^S \eta_{k+r_s,s} \le \eta\left(\frac{2}{T(T+3)} + \sum_{s=1}^S \rho_s\right) = \frac{\eta\cdot(T+1)(T+2)}{T(T+3)}.$$

Thus we have

$$\begin{aligned}
&\sum_{k=0}^N \langle \sum_{s=1}^S K_{s,v}^* \overline{y}_s^k, \widehat{x}_v^k - x_v\rangle + f_v(\widehat{x}_v^k) - f_v(x_v)\\
\le\ & \eta\left(\frac{2}{T(T+3)} + \overline{r}\right) D_{w^x}(x_v, x_v^{init}) - \frac{(T+1)(T+2)}{T(T+3)}\eta D_{w^x}(x_v, x_v^N)\\
&- \eta\sum_{k=0}^N \sum_{s=1}^S \rho_s D_{w^x}(\widehat{x}_v^k, x_v^{k-r_s}) + \frac{4M^2(N+1)}{\eta(T+3)}
\end{aligned} \tag{25}$$

Dual update properties. In addition, by the updating rule for the dual, we have by Proposition 2 in [24] for any $y_s \in \mathbb{R}^{n_s}$,

$$\begin{aligned}&\langle -\frac{1}{r_s}\sum_{v\in V} K_{s,v}\widetilde{x}_{s,v}^{i_s}, y_s^{i_s} - y_s\rangle + R_s^*(y_s^{i_s}) - R_s^*(y_s)\\ &\leq \tau_{s,i_s}\left(D_{w_s^y}(y_s, y_s^{i_s-1}) - D_{w_s^y}(y_s, y_s^{i_s}) - D_{w_s^y}(y_s^{i_s}, y_s^{i_s-1})\right).\end{aligned}$$

Thus, with $\tau_{s,i_s} = \tau_s$ for all i_s, summing over the above, we get

$$\begin{aligned}&\sum_{i_s=0}^{N_s-1}\langle -\frac{1}{r_s}\sum_{v\in V} K_{s,v}\widetilde{x}_{s,v}^{i_s}, y_s^{i_s} - y_s\rangle + R_s^*(y_s^{i_s}) - R_s^*(y_s)\\ &\leq \tau_s\left(D_{w_s^y}(y_s, y_s^{init}) - D_{w_s^y}(y_s, y_s^{N_s-1})\right) - \tau_s\cdot\sum_{i_s=0}^{N_s-1} D_{w_s^y}(y_s^{i_s}, y_s^{i_s-1}).\end{aligned} \tag{26}$$

Gap properties. Notice that for each $s \in [S]$ and $v \in V$, we have

$$\begin{aligned}&\sum_{k=0}^{N}\left\{\langle \widehat{x}_v^k, K_{s,v}^* y_s\rangle - \langle x_v, K_{s,v}^*\overline{y}_s^k\rangle\right\}\\ &= \sum_{i_s=0}^{N_s-1}\left\{\sum_{i=0}^{r_s-1}\langle \widehat{x}_v^{r_s i_s+i}, K_{s,v}^* y_s\rangle - r_s\langle x_v, K_{s,v}^* y_s^{i_s}\rangle\right\}\\ &= \sum_{i_s=0}^{N_s-1}\langle\sum_{i=0}^{r_s-1}\widehat{x}_v^{r_s i_s+i} - \widetilde{x}_{s,v}^{i_s}, K_{s,v}^*(y_s - y_s^{i_s})\rangle + \sum_{k=0}^{N}\langle \widehat{x}_v^k - x_v, K_{s,v}^*\overline{y}_s^k\rangle + \sum_{i_s=0}^{N_s-1}\langle K_{s,v}\widetilde{x}_{s,v}^{i_s}, y_s - y_s^{i_s}\rangle.\end{aligned} \tag{27}$$

Recall that for $i_s = 0, 1, \ldots, N_s - 1$,

$$\widetilde{x}_{s,v}^{i_s} = \alpha\Big(\sum_{k'=r_s i_s - r_s}^{r_s i_s - 1}\widehat{x}_v^{k'} - \sum_{k'=r_s i_s - 2r_s}^{r_s i_s - r_s - 1} x_v^{k'}\Big) + \sum_{k'=r_s i_s - r_s}^{r_s i_s - 1} x_v^{k'}.$$

We first bound the first term in (27). Notice that for $i_s = 0, 1, \ldots, N_s - 1$, we have

$$\begin{aligned}&\langle\sum_{i=0}^{r_s-1}\widehat{x}_v^{r_s i_s+i} - \widetilde{x}_{s,v}^{i_s}, K_{s,v}^*(y_s - y_s^{i_s})\rangle\\ &= \langle\sum_{i=0}^{r_s-1}(\widehat{x}_v^{r_s i_s+i} - x_v^{r_s(i_s-1)+i}) - \alpha\sum_{i=0}^{r_s-1}(\widehat{x}_v^{r_s(i_s-1)+i} - x_v^{r_s(i_s-2)+i}), K_{s,v}^*(y_s - y_s^{i_s})\rangle\\ &= \langle\sum_{i=0}^{r_s-1}(\widehat{x}_v^{r_s i_s+i} - x_v^{r_s(i_s-1)+i}), K_{s,v}^*(y_s - y_s^{i_s})\rangle\\ &\quad - \alpha\langle\sum_{i=0}^{r_s-1}(\widehat{x}_v^{r_s(i_s-1)+i} - x_v^{r_s(i_s-2)+i}), K_{s,v}^*(y_s - y_s^{i_s-1})\rangle\\ &\quad + \alpha\langle\sum_{i=0}^{r_s-1}(\widehat{x}_v^{r_s(i_s-1)+i} - x_v^{r_s(i_s-2)+i}), K_{s,v}^*(y_s^{i_s} - y_s^{i_s-1})\rangle.\end{aligned}$$

Thus, with $\alpha = 1$, and recall that for $i_s = 0$, $i = 0, \ldots, r_s - 1$, $\widehat{x}_v^{r_s(i_s-1)+i} - x_v^{r_s(i_s-2)+i} = x_v^{init} - x_v^{init} = \mathbf{0}$, we have

$$\begin{aligned}
&\sum_{v\in V}\sum_{i_s=0}^{N_s-1}\langle\sum_{i=0}^{r_s-1}\widehat{x}_v^{r_s i_s+i} - \widetilde{x}_{s,v}^{i_s}, K_{s,v}^*(y_s - y_s^{i_s})\rangle \\
=&\langle\sum_{i=0}^{r_s-1}(\widehat{X}^{N-r_s+i} - X^{N-2r_s+i}), K_s^*(y_s - y_s^{N_s-1})\rangle \\
&+\sum_{i_s=1}^{N_s-1}\langle\sum_{i=0}^{r_s-1}(\widehat{X}^{r_s(i_s-1)+i} - X^{r_s(i_s-2)+i}), K_s^*(y_s^{i_s} - y_s^{i_s-1})\rangle \\
\le&\sum_{i=0}^{r_s-1}\|\widehat{X}^{r_s(N_s-1)+i} - X^{r_s(N_s-2)+i}\|\cdot\|K_s^*(y_s - y_s^{N_s-1})\|_* \\
&+\sum_{i_s=1}^{N_s-1}\sum_{i=0}^{r_s-1}\|\widehat{X}^{r_s(i_s-1)+i} - X^{r_s(i_s-2)+i}\|\cdot\|K_s^*(y_s^{i_s} - y_s^{i_s-1})\|_*
\end{aligned}$$

Thus, for any $\rho > 0$, we have

$$\begin{aligned}
&\sum_{v\in V}\sum_{i_s=0}^{N_s-1}\langle\sum_{i=0}^{r_s-1}\widehat{x}_v^{r_s i_s+i} - \widetilde{x}_{s,v}^{i_s}, K_{s,v}^*(y_s - y_s^{i_s})\rangle \\
\le&\sum_{k=0}^{N}\frac{\rho}{2}\|\widehat{X}^k - X^{k-r_s}\|^2 + \frac{r_s\widetilde{\kappa}_s^2}{2\rho}(\sum_{i_s=1}^{N_s-1}\|y_s^{i_s} - y_s^{i_s-1}\|^2 + \|y_s - y_s^{N_s-1}\|^2).
\end{aligned} \tag{28}$$

Bounding the gap. Thus, with (25), (26), and (28), we have the following upper bound on the gap

$$\begin{aligned}
&\sum_{k=0}^{N}Q(\widehat{X}^k, \overline{Y}^k; Z) \\
=&\sum_{k=0}^{N}\Big\{F(\widehat{X}^k) - R^*(Y) - F(X) + R^*(\overline{Y}^k)\Big\} + \sum_{s=1}^{S}\sum_{k=0}^{N}\Big\{\langle K_s\widehat{X}^k, Y_s\rangle - \langle K_s X, \overline{Y}^k\rangle\Big\} \\
\le&\,\eta(\frac{2}{T(T+3)} + \overline{r})D_{w^X}(X, X^{init}) - \frac{(T+1)(T+2)}{T(T+3)}\eta D_{w^X}(X, X^N) \\
&-\eta\sum_{k=0}^{N}\sum_{s=1}^{S}\rho_s D_{w^X}(\widehat{X}^k, X^{k-r_s}) + \frac{4mM^2(N+1)}{\eta(T+3)} \\
&+\sum_{s=1}^{S}\tau_s r_s\left\{D_{w_s^y}(y_s, y_s^{init}) - D_{w_s^y}(y_s, y_s^{N_s-1}) - \sum_{i_s=0}^{N_s-1}D_{w_s^y}(y_s^{i_s}, y_s^{i_s-1})\right\} \\
&+\sum_{k=0}^{N}\sum_{s=1}^{S}\frac{\eta\rho_s}{2}\|\widehat{X}^k - X^{k-r_s}\|^2 + \sum_{s=1}^{S}\frac{r_s\widetilde{\kappa}_s^2}{2\eta\rho_s}(\sum_{i_s=1}^{N_s-1}\|y_s^{i_s} - y_s^{i_s-1}\|^2 + \|y_s - y_s^{N_s-1}\|^2)
\end{aligned}$$

where we take $\rho = \eta\rho_s$ in (26). Thus, with $\frac{\widetilde{\kappa}_s^2}{\rho_s\tau_s} \leq \frac{\eta}{2}$ for all $s \in [S]$, we have

$$\begin{aligned}
&\sum_{k=0}^{N} Q(\widehat{X}^k, \overline{Y}^k; Z) \\
&\leq \eta(\frac{2}{T(T+3)} + \overline{r}) D_{w^X}(X, X^{init}) - \frac{(T+1)(T+2)}{T(T+3)} \eta D_{w^X}(X, X^N) \\
&\quad + \sum_{s=1}^{S} \tau_s r_s \cdot \left\{ \frac{3}{2} D_{w_s^y}(y_s, y_s^{init}) - \frac{1}{2} D_{w_s^y}(y_s, y_s^{N_s-1}) \right\} + \frac{4mM^2(N+1)}{\eta(T+3)}
\end{aligned}$$

In particular, taking $\tau_s = \frac{2\widetilde{\kappa}_s^2}{\rho_s\eta}$, and using the convexity, we get

$$\begin{aligned}
&(N+1) \cdot Q(Z^N; Z) \\
&\leq \eta \left\{ (\frac{2}{T(T+3)} + \overline{r}) D_{w^X}(X, X^{init}) - \frac{(T+1)(T+2)}{T(T+3)} D_{w^X}(X, X^N) \right\} \\
&\quad + \frac{1}{\eta} \left\{ \sum_{s=1}^{S} \frac{\widetilde{\kappa}_s^2 r_s}{\rho_s} \left\{ 3D_{w_s^y}(y_s, y_s^{init}) - D_{w_s^y}(y_s, y_s^{N_s-1}) \right\} + \frac{4mM^2(N+1)}{T+3} \right\}
\end{aligned}$$

The result follows from noticing that for any $T \geq 1$, $\frac{2}{T(T+3)} \leq \frac{1}{2} \leq \frac{1}{2}(\sum_{s=1}^{S} r_s\rho_s)$, and $\frac{(T+1)(T+2)}{T(T+3)} \geq 1$. ■

F.4. Proofs for Section D

Proof [Proof of Lemma 11] **Primal update properties.** By Corollary 6, for any $v \in V$ and $x_v \in \mathcal{X}$

$$\begin{aligned}
&\langle \sum_{s=1}^{S} K_{s,v}^* \overline{y}_s^k, \widehat{x}_v^k - x_v \rangle + f_v(\widehat{x}_v^k) - f_v(x_v) \\
&\leq \sum_{s=1}^{S} \eta_{k,s} D_{w^x}(x_v, x_v^{k-r_s}) - \sum_{s=1}^{S} \eta_{k,s} D_{w^x}(\widehat{x}_v^k, x_v^{k-r_s}) \\
&\quad - (\frac{\mu}{C} + \eta_k) D_{w^x}(x_v, x_v^k) + \frac{2M^2}{\eta_k T_k(T_k+1)} \sum_{t=1}^{T_k} \frac{\lambda_t}{\beta_t^k}.
\end{aligned}$$

Thus, taking a weighted sum of the above, we get

$$\begin{aligned}
&\sum_{k=0}^{N} \theta_k \left\{ \langle \sum_{s=1}^{S} K_{s,v}^* \overline{y}_s^k, \widehat{x}_v^k - x_v \rangle + f_v(\widehat{x}_v^k) - f_v(x_v) \right\} \\
&\leq \frac{\mu(\overline{r^3}/\overline{r} + 5(\overline{r^2}/\overline{r})^2)}{2C} D_{w^x}(x_v, x_v^{init}) - \sum_{k=0}^{N} \theta_k \sum_{s=1}^{S} \eta_{k,s} D_{w^x}(\widehat{x}_v^k, x_v^{k-r_s}) + \frac{\mu\overline{r^2}}{C\overline{r}} D_0,
\end{aligned}$$

where the bounds on the sum of Bregman divergence terms follows from the following bounds on the coefficients of $D_{w^x}(x_v, x_v^k)$ term in the sum ($\theta_k = \eta_k = 0$ for all $k \geq N+1$) for $k = 0, 1, \ldots, N$

$$\begin{aligned}
&\sum_{s=1}^{S} \theta_{k+r_s}\eta_{k+r_s,s} - (\frac{\mu}{C} + \eta_k)\theta_k \\
&\leq \frac{\mu}{2\overline{r}C}\left\{\sum_{s=1}^{S}(k + r_s + 2\overline{r^2}/\overline{r})(k + r_s + \overline{r^2}/\overline{r})\rho_s - (k + \overline{r^2}/\overline{r} + 2\overline{r})(k + 2\overline{r^2}/\overline{r})\right\} \\
&= \frac{\mu}{2\overline{r}C}\left\{\left(k^2 + (2\overline{r} + \frac{3\overline{r^2}}{\overline{r}})k + (4\overline{r^2} + 2(\overline{r^2}/\overline{r})^2)\right) - \left(k^2 + (2\overline{r} + \frac{3\overline{r^2}}{\overline{r}})k + (4\overline{r^2} + 2(\overline{r^2}/\overline{r})^2)\right)\right\} = 0,
\end{aligned}$$

and for x_v^{init}

$$\sum_{s=1}^{S}\sum_{k=0}^{r_s-1}\eta_{k,s}\theta_k \leq \frac{\mu}{2\overline{r}C}\sum_{s=1}^{S}\rho_s \cdot r_s(r_s + \frac{\overline{r^2}}{\overline{r}})(r_s + 2\frac{\overline{r^2}}{\overline{r}}) = \frac{\mu(\overline{r^3}/\overline{r} + 5(\overline{r^2}/\overline{r})^2)}{2C}.$$

The bound on the rest of the terms is since

$$\frac{1}{\eta_k}\sum_{t=1}^{T_k}\frac{\lambda_t}{\beta_t^k} = \sum_{t=1}^{T_k}\frac{2tC/\mu}{(t+1) + (t-1)(k + \overline{r^2}/\overline{r})/2\overline{r}} \leq \frac{C}{\mu}(1 + \frac{4(T_k - 1)}{1 + (k + \overline{r^2}/\overline{r})/2\overline{r}}),$$

and notice that

$$\sum_{k=0}^{N}\frac{k + 2\overline{r^2}/\overline{r}}{k + \overline{r^2}/\overline{r} + 2\overline{r}} \leq 2(N+1) \leq 4N,$$

and since $N + 1 \geq r_{\max} := \max_{s\in[S]} r_s$, we have

$$\overline{r^2} = \sum_{s=1}^{S}\rho_s r_s^2 \leq \sum_{s=1}^{S}\rho_s r_s(N+1) = (N+1)\overline{r} \implies \overline{r^2}/\overline{r} \leq N + 1 \leq 2N.$$

Thus, for $T_k/N = T/N \geq \max(\frac{5}{\sqrt{D_1}}, \frac{64\overline{r}}{D_1})$ where $D_1 = \frac{\mu^2(\overline{r^2}/\overline{r})^2}{2M^2C^2}D_0$, we have

$$\begin{aligned}
\sum_{k=0}^{N}\frac{2M^2\theta_k}{\eta_k T_k(T_k+1)}\sum_{t=1}^{T_k}\frac{\lambda_t}{\beta_t^k} &\leq \frac{2M^2C}{\mu}\left\{\sum_{k=0}^{N}\frac{k + 2\overline{r^2}/\overline{r}}{T_k(T_k+1)} + \sum_{k=0}^{N}\frac{4(k + 2\overline{r^2}/\overline{r})}{(1 + (k + \overline{r^2}/\overline{r})/2\overline{r})(T_k+1)}\right\} \\
&\leq \frac{2M^2C}{\mu}\left\{\frac{10N^2}{T^2} + \frac{32N\overline{r}}{T}\right\} \leq \frac{\mu}{C}D_0(\overline{r^2}/\overline{r})^2
\end{aligned}$$

Dual update properties. Similar to (26), we get

$$\begin{aligned}
&\sum_{i_s=0}^{N_s-1}\left\{\langle -\sum_{v\in V}K_{s,v}\widetilde{x}_{s,v}^{i_s}, y_s^{i_s} - y_s\rangle + (\sum_{k'=r_s i_s}^{r_s i_s + r_s - 1}\theta_k')(R_s^*(y_s^{i_s}) - R_s^*(y_s))\right\} \\
&\leq \sum_{i_s=0}^{N_s-1}\tau_{s,i_s}(\sum_{k'=r_s i_s}^{r_s i_s + r_s - 1}\theta_k')\left\{D_{w_s^y}(y_s, y_s^{i_s-1}) - D_{w_s^y}(y_s, y_s^{i_s}) - D_{w_s^y}(y_s^{i_s}, y_s^{i_s-1})\right\} \\
&= \tau_s\left\{D_{w_s^y}(y_s, y_s^{init}) - D_{w_s^y}(y_s, y_s^{N_s-1}) - \sum_{i_s=0}^{N_s-1}D_{w_s^y}(y_s^{i_s}, y_s^{i_s-1})\right\}. \qquad (29)
\end{aligned}$$

Gap properties. Notice that for each $s \in [S]$ and $v \in V$, we have

$$\begin{aligned}
&\sum_{k=0}^{N} \theta_k \left\{ \langle \widehat{x}_v^k, K_{s,v}^* y_s \rangle - \langle x_v, K_{s,v}^* \overline{y}_s^k \rangle \right\} \\
&= \sum_{i_s=0}^{N_s-1} \langle \sum_{i=0}^{r_s-1} \theta_{r_s i_s+i} \widehat{x}_v^{r_s i_s+i} - \widetilde{x}_{s,v}^{i_s}, K_{s,v}^*(y_s - y_s^{i_s}) \rangle \\
&\quad + \sum_{k=0}^{N} \theta_k \langle \widehat{x}_v^k - x_v, K_{s,v}^* \overline{y}_s^k \rangle + \sum_{i_s=0}^{N_s-1} \langle K_{s,v} \widetilde{x}_{s,v}^{i_s}, y_s - y_s^{i_s} \rangle. \qquad (30)
\end{aligned}$$

We first bound the first term in (27). With $\alpha_{s,i_s} = 1$

$$\begin{aligned}
&\sum_{v \in V} \sum_{i_s=0}^{N_s-1} \langle \sum_{i=0}^{r_s-1} \theta_{r_s i_s+i} \widehat{x}_v^{r_s i_s+i} - \widetilde{x}_{s,v}^{i_s}, K_{s,v}^*(y_s - y_s^{i_s}) \rangle \\
&= \langle \sum_{i=0}^{r_s-1} (\theta_{N-r_s+i} (\widehat{X}^{N-r_s+i} - X^{N-2r_s+i})), K_s^*(y_s - y_s^{N_s-1}) \rangle \\
&\quad + \sum_{i_s=1}^{N_s-1} \langle \sum_{i=0}^{r_s-1} (\theta_{r_s(i_s-1)+i} (\widehat{X}^{r_s(i_s-1)+i} - X^{r_s(i_s-2)+i})), K_s^*(y_s^{i_s} - y_s^{i_s-1}) \rangle \\
&\leq \sum_{i=0}^{r_s-1} \theta_{r_s(N_s-1)+i} \| \widehat{X}^{r_s(N_s-1)+i} - X^{r_s(N_s-2)+i} \| \cdot \| K_s^*(y_s - y_s^{N_s-1}) \|_* \\
&\quad + \sum_{i_s=1}^{N_s-1} \sum_{i=0}^{r_s-1} \theta_{r_s(i_s-1)+i} \| \widehat{X}^{r_s(i_s-1)+i} - X^{r_s(i_s-2)+i} \| \cdot \| K_s^*(y_s^{i_s} - y_s^{i_s-1}) \|_*
\end{aligned}$$

Bounding the gap. Putting the above together, and for convenience, denoting $y_s^{N_s} = y_s$, we have

$$\begin{aligned}
&\sum_{k=0}^{N} \theta_k Q(\widehat{X}^k, \overline{Y}^k; Z) \\
&\leq \frac{\mu(\overline{r^3}/\overline{r} + 5(\overline{r^2}/\overline{r})^2)}{2C} D_{w^X}(X, X^{init}) - \sum_{k=0}^{N} \theta_k \sum_{s=1}^{S} \eta_{k,s} D_{w^X}(\widehat{X}^k, X^{k-r_s}) + \frac{m\mu(\overline{r^2}/\overline{r})^2}{C} D_0 \\
&\quad + \sum_{s=1}^{S} \tau_s \left\{ D_{w_s^y}(y_s, y_s^{init}) - D_{w_s^y}(y_s, y_s^{N_s-1}) - \sum_{i_s=1}^{N_s-1} D_{w_s^y}(y_s^{i_s}, y_s^{i_s-1}) \right\} \\
&\quad + \sum_{s=1}^{S} \left\{ \sum_{k=0}^{N} \frac{\theta_k \eta_{k,s}}{2} \| \widehat{X}^k - X^{k-r_s} \|^2 + \sum_{i_s=1}^{N_s} \frac{\widetilde{\kappa}_s^2}{2\rho_s} (\sum_{i=0}^{r_s-1} \frac{\theta_{r_s(i_s-1)+i}}{\eta_{r_s(i_s-1)+i}}) \| y_s^{i_s} - y_s^{i_s-1} \|^2 \right\} \\
&\leq \frac{\mu(\overline{r^3}/\overline{r} + 5(\overline{r^2}/\overline{r})^2)}{2C} D_{w^X}(X, X^{init}) + \frac{m\mu(\overline{r^2}/\overline{r})^2}{C} D_0 + \sum_{s=1}^{S} \tau_s D_{w_s^y}(y_s, y_s^{init})
\end{aligned}$$

using

$$\tau_s = \frac{\widetilde{\kappa}_s^2}{\rho_s} \cdot \frac{4r_s\overline{r}C}{\mu} \geq \frac{\widetilde{\kappa}_s^2}{\rho_s} \cdot \max_{i_s \in [N_s]} \Big(\sum_{i=0}^{r_s-1} \frac{\theta_{r_s(i_s-1)+i}}{\eta_{r_s(i_s-1)+i}} \Big).$$

Since $\sum_{k=0}^{N} \theta_k \geq \frac{N(N+1)}{2}$, using convexity, we get

$$Q(Z^N; Z) \leq \frac{2}{N(N+1)} \Bigg\{ \frac{\mu(\overline{r^3}/\overline{r} + 5(\overline{r^2}/\overline{r})^2)}{2C} D_{w^X}(X, X^{init}) + \frac{m\mu(\overline{r^2}/\overline{r})^2}{C} D_0 + \frac{4\overline{r}C}{\mu} \sum_{s=1}^{S} \frac{\widetilde{\kappa}_s^2 r_s}{\rho_s} D_{w_s^y}(y_s, y_s^{init}) \Bigg\}$$

■