A Linearly Convergent Algorithm for Decentralized Optimization: Sending Less Bits for Free!

Dmitry Kovalev KAUST **Anastasia Koloskova** EPFL **Martin Jaggi** EPFL **Peter Richtárik** KAUST **Sebastian U. Stich** EPFL

Abstract

Decentralized optimization methods enable on-device training of machine learning models without a central coordinator. In many scenarios communication between devices is energy demanding and time consuming and forms the bottleneck of the entire system.

We propose a new randomized first-order method which tackles the communication bottleneck by applying randomized compression operators to the communicated messages. By combining our scheme with a new variance reduction technique that progressively throughout the iterations reduces the adverse effect of the injected quantization noise, we obtain a scheme that converges linearly on strongly convex decentralized problems while using compressed communication only. We prove that our method can solve the problems without any increase in the number of communications compared to the baseline which does not perform any communication compression while still allowing for a significant compression factor which depends on the conditioning of the problem and the topology of the network. We confirm our theoretical findings in numerical experiments.

1 Introduction

We consider large-scale convex optimization problems of the form

$$f^\star := \min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \left[f_i(x) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[f_i(x,\xi)\right]\right], \qquad (1)$$

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

with private loss functions $f_i\colon \mathbb{R}^d \to \mathbb{R}$ split among n machines (workers). This problem formulation covers for instance empirical risk minimization over finite datasets with equal loss functions but different data samples available on each device, but more generally also the stochastic setting where the workers have access to unbounded number of independent samples.

We assume that the workers are connected over an arbitrary network and that they can only exchange information with their immediate neighbors in the network. This setting covers the classical parameter-server infrastructures, where all devices are connected to one central server (Dean et al., 2012), the emerging federated learning paradigm (McMahan et al., 2016, 2017; Kairouz et al., 2019), and most generally, arbitrary decentralized communication topologies (Tsitsiklis, 1984; Nedić, 2020; Xin et al., 2020).

Communication is a key bottleneck when the working devices are connected over networks (Seide and Agarwal, 2016; Alistarh et al., 2017). Quantization techniques enable optimization with compressed messages, hereby reducing the number of bits that have to be exchanges between the workers in each communication round. Whilst the first schemes of this type have been presented for centralized topologies only (Alistarh et al., 2017; Wangni et al., 2018), many adaptations have been developed recently for optimization over arbitrary networks (Tang et al., 2018; Koloskova et al., 2019; Tang et al., 2019; Reisizadeh et al., 2019; Koloskova et al., 2020a).

All these, so far mentioned, decentralized schemes only converge *sublinearly* when using compressed messages, i.e. they need $\mathcal{O}\left(1/\epsilon^\tau\right)$ iterations to reach accuracy ϵ for a parameter $0 < \tau < \infty$ (most commonly $\tau \in \{1/2, 1, 2\}$). This is in sharp contrast to centralized approaches with parameter servers, where linear convergence rates of the form $\mathcal{O}\left(\log 1/\epsilon\right)$ can be attained even with communication compression, for instance when the objective function is strongly convex (Horváth et al., 2019). We believe that there is an intrinsic reason for this limitation: these early schemes for communication compression and optimization over arbitrary networks have been derived by *adapting the the decentralized gradient method* and compressing the gradient updates. However, decentralized gradient descent cannot achieve linear convergence on strongly convex problems, even without communication compression (Shi et al., 2015; Yuan et al., 2016; Koloskova et al., 2020b).

We develop new algorithms for quantized decentralized optimization based on the primal-dual gradient method (Chen and Rockafellar, 1997; Boyd et al., 2011) instead. This allows to overcome limitations of prior schemes. Most importantly, we are able to prove linear convergence on strongly convex functions for arbitrary unbiased randomized compressors.

Our results extend and improve the parallel work[1] in (Liu et al., 2020), that also applies to arbitrary compressors, and prior work (Magnússon et al., 2020) that is tied to a specific quantization scheme. We present a detailed comparison to these papers in Section 5.

Our main contributions can be summarized as follows:

(a) We design decentralized optimization algorithms for problem (1). For μ-strongly convex and L-smooth objective with condition number $\kappa := L/\mu$, our main algorithm converges linearly and achieves an ϵ accurate solution after at most

$$\mathcal{O}\left((\omega + \kappa(\rho + \omega\rho_\infty))\log\frac{1}{\epsilon}\right) \tag{2}$$

iterations, where $\rho \geq 1$ denotes the ratio between the largest and smallest non-zero eigenvalues of the Laplacian gossip matrix that encodes the communication topology, $\rho_\infty \leq \rho$ a new graph parameter we introduce later, and $\omega \geq 0$ quantifies the quality of an arbitrary unbiased quantization operator. For the special case $\omega = 0$ (no quantization) our rates recover the linear convergence rates of the primal-dual gradient method (Bertsekas, 1982; Alghunaim and Sayed, 2020). We provide further in-depth discussion of our convergence results in Section 5, see also Tables 1–2.

(b) Most notably, equation (2) reveals that for *any* compression parameter $\omega \leq \min\{\rho\rho_\infty^{-1}, \kappa\rho\}$ the complexity bound is $\mathcal{O}(\kappa\rho\log 1/\epsilon)$—the same as for the primal-dual method without compression. This means, that any communication saving achieved by quantization is *for free*, as they do not affect the total number of communication rounds but reduce the number of bits sent every round. We will show that the savings in communication can reach up to a factor of $\mathcal{O}(n)$ on certain problems.

(c) We give algorithms and convergence analysis for four important cases: (A) a primal-dual method for dual-friendly problems, (B) an incremental method only using primal gradient oracles, and especially for the machine learning context (C) a method for stochastic gradient oracles and (D) a variance-reduced method when the local functions have finite-sum structure.

(d) We illustrate in numerical experiments that the performance of our schemes matches with the theoretical rates and compare against prior baselines.

[1] Their proposed method is identical to **option B** (incremental primal update) our Algorithm 1.

2 Related Work

As decentralized optimization problems are special cases of linearly constrained (consensus constraint) optimization problems, algorithms based on augmented Lagrangian reformulations and primal dual algorithms, such as alternating method of multipliers (ADMM) (Glowinski and Marrocco, 1975; Gabay and Mercier, 1976), have been developed early on (Boyd et al., 2011). Linear convergence rates for primal-dual methods on strongly convex problems have been derived and refined over the past decades (Bertsekas, 1982; Tsitsiklis, 1984; Chen and Rockafellar, 1997; Shi et al., 2014; Alghunaim and Sayed, 2020). A variety of decentralized optimization schemes have been introduce and studied in the control and optimization communities (Duchi et al., 2012; Wei and Ozdaglar, 2012; Iutzeler et al., 2013; Rabbat, 2015; He et al., 2018; Lian et al., 2017; Wang and Joshi, 2018; Koloskova et al., 2020b), see also the review articles (Sayed, 2014; Xin et al., 2020; Nedić, 2020). Limitations of the distributed gradient method, such as for instance not attaining linear convergence rates, have been pointed out for instance in (Shi et al., 2015) and techniques such as EXTRA (Shi et al., 2015) and gradient tracking (Nedić et al., 2017) have been developed to achieve linear convergence on strongly convex problems with primal methods as well. Optimal decentralized algorithms based on accelerated gossip protocols have been presented in (Scaman et al., 2017) and (Uribe et al., 2018).

Quantization. Quantization techniques allow for (lossy) compression of the messages that are exchanged between the agents to reduce the number of bits that need to be exchanged in each round. Quantization has emerged in recent years as an important tool in parallel and distributed machine learning (Seide et al., 2014; Strom, 2015; Alistarh et al., 2017; Wen et al., 2017). Whilst these early schemes have suffered from increased variance due to the randomized compression schemes, schemes based on error-feedback can compensate these effects and attain faster convergence (Alistarh et al., 2018; Stich et al., 2018; Karimireddy

Table 1: Comparison to decentralized algorithms with communication compression and baseline results without compression. The rates show the most significant terms and indicate how many iterations are needed to reach $\|x - x^\star\|^2 \le \epsilon$ for all nodes. Here $\tilde{\rho} \approx \rho$, $\tilde{\omega} \ge \omega$ and $\tau \ge 1$ is an algorithm and function dependent constant, cf. the indicated references for definitions.

Algorithm[a] & Reference	linear rate	quantization	convergence to ϵ-accuracy
Decentralized Gradient Descent (Nedić+ 2009; Koloskova+ 2020b)			$\mathcal{O}\Big(\frac{\sqrt{\kappa}\tilde{\rho}^2}{\mu} \cdot \frac{1}{\sqrt{\epsilon}}\Big)$
QDGD (Reisizadeh+ 2019)		✓	$\mathcal{O}\Big(\frac{\kappa^2\tilde{\rho}^4 L^4 + \tilde{\omega}^2}{\mu^2} \cdot \frac{1}{\epsilon^2}\Big)$
Choco-SGD (Koloskova+ 2019)		✓	$\mathcal{O}\Big(\frac{\sqrt{\kappa}\tilde{\rho}^2(1+\omega)}{\mu} \cdot \frac{1}{\sqrt{\epsilon}}\Big)$
EXTRA (Shi+ 2015), Gradient Tracking (Qu+ 2016; Pu+ 2020)	✓		$\mathcal{O}\Big((\kappa^\tau \tilde{\rho}^2) \cdot \log\frac{1}{\epsilon}\Big)$
Primal Dual Gradient Method (Scaman+ 2017; Alghunaim+ 2020)	✓		$\mathcal{O}\Big((\kappa\rho) \cdot \log\frac{1}{\epsilon}\Big)$
LEAD (Liu+ 2020)	✓	✓	$\mathcal{O}\Big((\kappa\rho\omega) \cdot \log\frac{1}{\epsilon}\Big)$
this paper	✓	✓	$\mathcal{O}\Big((\kappa\rho + \kappa\rho_\infty\omega) \cdot \log\frac{1}{\epsilon}\Big)$

[a] Convergence rates for the non-accelerated versions of these schemes.

et al., 2019; Stich and Karimireddy, 2019) on centralized network topologies.

Quantization in the context of decentralized optimization has first been studied for the decentralized consensus problem where the agents aim to collaboratively compute the average of private data vectors. The effects of various quantization techniques have been studied in (Xiao et al., 2005; Nedić et al., 2008; Carli et al., 2010b) and many different techniques have been proposed to address quantization errors, such as decreasing stepsizes or adaptive coding schemes (Carli et al., 2010a; Yuan et al., 2012; Reisizadeh et al., 2019). Only recently, a first scheme with linear convergence to the exact solution was presented (Koloskova et al., 2019). However, this algorithm does not converge linearly on arbitrary strongly convex optimization problems that we consider here. Primal-dual methods with quantization have been introduced in (Magnússon et al., 2020; Liu et al., 2020). For more general, non-convex problems, further schemes with communication compression have been proposed by Tang et al. (2018, 2019); Koloskova et al. (2020a).

Variance Reduction. Variance reduction for finite-sum structured problems has been introduced in (Johnson and Zhang, 2013; Defazio et al., 2014) and previously been applied to the closely related saddle-point problems (Palaniappan and Bach, 2016) and specifically also for decentralized consensus optimization (Mokhtari and Ribeiro, 2016; Xin et al., 2019). Hendrikx et al. (2020) developed an optimal algorithm for finite-sum optimization. Variance reduction and in combination with communication compression has previously been studied in the context of distributed optimization with a parameter server only (Horváth et al., 2019). This method relies on efficient (and uncompressed) broadcast communication which we avoid here by supporting a fully decentralized topology.

3 Setup

We now specify the problem formulation, assumptions, and define several key concepts that will be used throughout the paper.

3.1 Regularity Assumptions

Assumption 1. *Each cost function $f_i\colon \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth, for parameters $0 < \mu \le L$ and condition number $\kappa := L/\mu$. That is $\forall\, x, y \in \mathbb{R}^d, i \in [n]$:*

$$f_i(y) \ge f_i(x) + \langle \nabla f_i(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2\,, \quad (3)$$

$$f_i(y) \le f_i(x) + \langle \nabla f_i(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2\,. \quad (4)$$

Sometimes we will also consider the stochastic setting:

$$f_i(x) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[f_i(x,\xi)\right]\,, \qquad \forall i \in [n], \quad (5)$$

where only stochastic gradients $\mathbb{E}_{\xi\sim\mathcal{D}}\left[\nabla f_i(x,\xi)\right] = \nabla f_i(x)$ are available. In this case we do need an additional assumption on the strength of the noise:

Assumption 2 (Bounded Variance and Smoothness)**.** *Function $f_i(x,\xi)$ is L-smooth in expectation and the stochastic variance at the optimum $x^\star := \arg\min f(x)$ is bounded. That is, for all $i \in [n]$ here exist $\sigma_i^2 \in \mathbb{R}_+$, such that*

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\|\nabla f_i(x^\star,\xi) - \nabla f_i(x^\star)\|_2^2\right] \le \sigma_i^2\,, \quad (6)$$

is bounded. We define $\sigma^2 := \frac{1}{n}\sum_{i=1}^n \sigma_i^2$. Further, smoothness implies the inequality

$$\mathbb{E}_{\xi\sim\mathcal{D}}\left[\|\nabla f_i(x,\xi) - \nabla f_i(y,\xi)\|_2^2\right] \leq 2LB_{f_i}(x,y)\,, \quad (7)$$

$\forall x, y \in \mathbb{R}^d, i \in [n]$, where $B_{f_i}(x,y)$ is a Bregman divergence $B_{f_i}(x,y) := f_i(x) - f_i(y) - \langle \nabla f_i(y), x-y\rangle$.

Remark 1. *For the special case of finite-sum structured problems on each worker, $f_i(x) = \frac{1}{m}\sum_{j=1}^m f_{ij}(x)$, equation (7) becomes*

$$\tfrac{1}{m}\textstyle\sum_{j=1}^m \|\nabla f_{ij}(x) - \nabla f_{ij}(y)\|_2^2 \leq 2LB_{f_i}(x,y)\,, \quad (8)$$

$\forall x, y \in \mathbb{R}^d$, $i \in [n]$.

3.2 Optimization over Networks

We model the network topology as an undirected graph $G = ([n], E)$ where $[n] := \{1, \ldots, n\}$ denotes the index set of the agents and $E \subset [n] \times [n]$ a set of pairs of communicating agents, $(i,j) \in E$ if and only if $(j,i) \in E$ (symmetric). If there exists an edge from agent i to agent j they may exchange information along this edge. Thus, agent i may send or receive messages from all its neighbors $\mathcal{N}_i = \{j \in [n] \mid (i,j) \in E\}$. We encode the communication links in a weighted Laplacian $\mathbf{W} \in \mathbb{S}_+^n$:

$$\mathbf{W}_{ij} = \begin{cases} -w_{ij}, & i \neq j, (i,j) \in E \\ 0, & i \neq j, (i,j) \notin E \\ \sum_{l \in \mathcal{N}_i} w_{il}, & i = j \end{cases}, \quad (9)$$

where $w_{ij} > 0$ for all $(i,j) \in E$. The mixing matrix is positive semidefinite $\mathbf{W} \in \mathbb{S}_+^n$, respects the graph structure, $\mathbf{W}_{ij} \neq 0$ only if $(i,j) \in E$, and $\ker \mathbf{W} = \operatorname{span}(\mathbf{1})$, where $\mathbf{1} = (1, \ldots, 1)^\top$. We denote by $\lambda_{\min}^+(\mathbf{W})$ the smallest non-zero eigenvalue of $\mathbf{W}$ and by $\lambda_{\max}(\mathbf{W})$ its largest eigenvalue. We define $\rho := \lambda_{\max}(\mathbf{W})/\lambda_{\min}^+(\mathbf{W})$ to be the ratio between the largest and the smallest non-zero eigenvalue of $\mathbf{W}$, and $\rho_\infty := \max_{(i,j)\in E} w_{ij}/\lambda_{\min}^+(\mathbf{W})$ the maximum normalized edge weight.

Remark 2. *It holds $\rho_\infty \leq \rho$ and the gap $\rho\rho_\infty^{-1} \geq 1$ can reach size $\Theta(n)$.*

Proof. For any Laplacian, we have[2] $\Delta \leq \lambda_{\max}(\mathbf{W})$ for maximal weighted degree $\Delta := \max_{i\in[n]} w_{ii}$. As $\max_{(i,j)\in E} w_{ij} \leq \max_{i\in[n]} w_{ii} = \Delta$, it follows $\rho_\infty \leq \rho$. For the second claim, consider a k-regular graph, for a parameter $1 \leq k \leq n-1$, and uniform weights, $w_{ij} = 1$ for $(i,j) \in E$. Then $\max_{(i,j)\in E} w_{ij} = \frac{\Delta}{k}$, and $\rho\rho_\infty^{-1} \geq k$. □

[2] Folklore; this bound can be shown by considering Rayleigh quotients $\Delta = \max_{i\in[n]} e_i^\top \mathbf{W} e_i \leq \lambda_{\max}(\mathbf{W})$.

Remark 3. *The consensus constraint, $x_i = x_j$ can compactly be written as $\mathbf{W}\big[x_1, \cdots, x_n\big] = \mathbf{0}$ in matrix form if the graph is connected. This observation can be utilized to derive the standard saddle point reformulations of problem (1), see for instance (Lan et al., 2018; Alghunaim et al., 2019).*

3.3 Unbiased Quantization

We consider unbiased randomized quantizers $\mathcal{Q}\colon \mathbb{R}^d \to \mathbb{R}^d$ as for instance in (Alistarh et al., 2017; Wangni et al., 2018; Horváth et al., 2019) with the following assumption on their variance.

Assumption 3 (ω-quanitzation)**.** *There exists a parameter $\omega \geq 0$ such that for all $x \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\mathcal{Q}(x)\right] = x\,, \quad \mathbb{E}\left[\|\mathcal{Q}(x) - x\|^2\right] \leq \omega\|x\|^2. \quad (10)$$

This general notion comprises many important examples of quantization operators currently used in applications. Below we just name a few (that we later use in the numerical experiments). However, it is important to note that our proposed method does *not* rely on a specific choice of quantization operator but can be used in combination with any arbitrary unbiased quantization scheme that satisfies Assumption 3.

Example 4 (rand-k and dit-k)**.** *Example compression operators and coding length, assuming that a single floating-point scalar is encoded with b bits with negligible loss in precision.*

- *no compression ($\omega = 0$). Each message has size db for this standard baseline.*
- rand-k*: random k-sparsification ($\omega = \frac{d}{k} - 1$) (Suresh et al., 2017; Wangni et al., 2018; Stich et al., 2018). $\mathcal{Q}(x) := \frac{d}{k}\mathcal{M}(x)$, where $\mathcal{M}(x)$ randomly selects k coordinates of x and masks the others to zero. The sparse vectors can be encoded with $kb + k\log d$ bits (non-zero coordinates and their indices).*
- dit-k*: random s-dithering ($\omega = \min\{\frac{d}{s^2}, \frac{\sqrt{d}}{s}\}$) (Goodall, 1951; Roberts, 1962; Alistarh et al., 2017). Each coordinate of the normalized vector $x/\|x\|$ is randomly rounded to one of s quanitzation levels, (often $s = 2^{k-1} - 1$ for integer k, so that the levels can be encoded with $k-1$ bits, plus one bit for the sign),*

  $$\mathcal{Q}(x) = \operatorname{sign}(x) \cdot \|x\|_2 \cdot \tfrac{1}{s} \cdot \left\lfloor s\tfrac{|x|}{\|x\|_2} + \xi \right\rfloor$$

  for random variable $\xi \sim_{\text{u.a.r.}} [0,1]^d$. As a special case for $s = 2$ one recovers Terngrad (Wen et al., 2017). A trivial upper bound for the encoding length is $dk+b$, but exploiting sparsity (encoding only non-zero quantized values and their indices) this bound can be improved to $\tilde{\mathcal{O}}(s(s+\sqrt{d})+b)$ (Alistarh et al., 2018).

Algorithm 1 Four Decentralized Quantized Optimization Algorithms

1: **Initialization**: $w_{ij} = w_{ji} > 0$ for $(i,j) \in E$, $z_1^0, \ldots, z_n^0 \in \mathbb{R}^d$ such that $\sum_{i=1}^n z_i^0 = 0$,
2: $x_1^0, \ldots, x_n^0 \in \mathbb{R}^d$, $h_1^0, \ldots, h_n^0 \in \mathbb{R}^d$, $\theta > 0$, $\alpha > 0$, $\eta > 0$
3: **for** $k = 0, 1, 2, \ldots$ **do**
4:     **for** $i = 1, \ldots, n$ **do** in parallel on each node — $\triangledown$ 4 options:
5:         • $x_i^{k+1} = \nabla f_i^\star(z_i^k)$ — $\triangleright$ **Option A** (dual update)
6:         • $x_i^{k+1} = x_i^k - \eta(\nabla f_i(x_i^k) - z_i^k)$ — $\triangleright$ **Option B** (incremental primal update)
7:         • Sample random $\xi_i^k \sim \mathcal{D}$
8:           $x_i^{k+1} = x_i^k - \eta(\nabla f_i(x_i^k, \xi_i^k) - z_i^k)$ — $\triangleright$ **Option C** (stochastic primal update)
9:         • Sample $j_i^k \in \{1, \ldots, m\}$ uniformly at random
10:           $g_i^k = \nabla f_{ij_i^k}(x_i^k) - \nabla f_{ij_i^k}(w_i^k) + \nabla f_i(w_i^k)$
11:           $w_i^{k+1} = \begin{cases} x_i^k, & \text{with probability } \frac{1}{m} \\ w_i^k, & \text{with probability } 1 - \frac{1}{m} \end{cases}$
12:           $x_i^{k+1} = x_i^k - \eta(g_i^k - z_i^k)$ — $\triangleright$ **Option D** (finite-sum structured problems)
13:         **for** $j \in \mathcal{N}_i$ **do**
14:             $\Delta_{ij}^k = \mathcal{Q}(x_i^{k+1} - h_i^k) + h_i^k$ — $\triangleright$ prepare quantized dual updates
15:         **end for**
16:         $h_i^{k+1} = h_i^k + \alpha \mathcal{Q}(x_i^{k+1} - h_i^k)$ — (communication with neighbors)
17:     **end for**
18:     **for** $i = 1, \ldots, n$ **do** in parallel on each node — $\triangledown$ update dual variables
19:         $z_i^{k+1} = z_i^k - \theta \sum_{j \in \mathcal{N}_i} w_{ij}(\Delta_{ij}^k - \Delta_{ji}^k)$ — (communication with neighbors)
20:     **end for**
21: **end for**

4 Algorithm

We give the pseudocode for our proposed schemes in Algorithm 1 above. We will give convergence rates for four different choices of updating the variables x_i^k (in this notation $i \in [n]$ range over the nodes, and $k \geq 0$ over the iterations).

Option A is applicable only if the dual functions $f_i^*\colon \mathbb{R}^d \to \mathbb{R}$ of each f_i are known and their gradients can be evaluated efficiently.[3]

Option B maintains dual variables z_i^k that are incrementally updated instead (accessing primal gradient $\nabla f(x_i^k)$ only). Similarly to the incremental version of the classic primal-dual gradient method, we will have $z_i^k \to z_i^+ := \nabla f_i(x^\star)$ for $k \to \infty$, which explains the intuition behind the z_i^k variables.

Option C is applicable when only stochastic gradient oracles are available.

Option D applies bias-corrected gradient updates for finite-sum structured f_i's (analogous to the bias corrected updates in SVRG (Johnson and Zhang, 2013)). Full batch gradients are re-computed after a random number of epochs (Hannah et al., 2018).

[3] The convex conjugate of $f_i^*\colon R^d \to \mathbb{R}$ of f_i is defined as $f_i^*(z) := \sup_{x \in \mathbb{R}^d}(\langle x, z\rangle - f_i(x))$.

We give the convergence rates for these variants in Section 5 below (see also Table 2).

The updates on lines 5–12 (depending on the chosen option) are performed in parallel on each agent. The auxiliary vectors h_i^k updated on line 16 are crucial component in our scheme that are required to achieve linear convergence: we will show in the appendix that $h_i^k \to x^\star$ for $(k \to \infty)$, so that for the quantization on line 14 we will be able to show (by virtue of (10)) that the quantization noise reduces linearly to zero as $x_i^k \to x^\star$ for $(k \to \infty)$. This would not be possible when quantizing the iterates x_i^k directly.

Implementation Details. It is easy to see that only quantized vectors need to be exchanged between the clients (every node needs to send two quantized vectors to each of its neighbors). To see this, assume that the vectors h_i^k are known to all neighbors of node i (maintaining h_i^k requires only quantized updates as per line 16: $h_i^{k+1} - h_i^k = \alpha q$, where q is a quantized vector). The update on line 19 can be rewritten as

$$z_i^k - z_i^{k+1} = \theta \sum_{j \in \mathcal{N}_i} \left(h_i^k - h_j^k + q_i - q_j\right),$$

where q_i, q_j are quantized vectors. Further note that the memory requirement is quite low per node: each agent needs to store its local copies of x_i^k, z_i^k, h_i^k and

Setting	Convergence Rate, $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors	Reference
A, dual $\nabla f_i^\star(z)$ available	$\mathcal{O}\left((\omega + \kappa(\rho + \omega\rho_\infty))\log\frac{1}{\epsilon}\right)$	Theorem 12
B, primal $\nabla f_i(x)$ available		Theorem 14
C, stochastic $\nabla f_i(x,\xi)$ available	$\tilde{\mathcal{O}}\left(\omega + (\rho + \omega\rho_\infty)\left(\kappa + \frac{\sigma\sqrt{1+\omega}}{\sqrt{\epsilon}\mu} + \frac{\sigma^2(\rho+\omega\rho_\infty)}{\epsilon\mu^2}\right)\right)$	Theorem 16
D, finite sum $f_i(x) = \frac{1}{m}\sum_{j=1}^m f_{ij}(x)$	$\mathcal{O}\left((m + \omega + \kappa(\rho + \omega\rho_\infty))\log\frac{1}{\epsilon}\right)$	Theorem 19

Table 2: Summary of the convergence results for Algorithm 1 to reach accuracy $\|x - x^\star\|^2 \leq \epsilon$ on all nodes. The depicted results are for stepsizes $\alpha = \frac{1}{\omega+1}$, $\eta = \frac{1}{L}$ and $\theta = \Theta\big(\frac{\mu}{\lambda_{\max}(\mathbf{W}) + \omega\max_{(i,j)\in E} w_{ij}}\big)$ for option A, B and D and chosen as in equation (50) for option C.

$\bar{h}_i^k := \frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i} h_j^k$ (but not each h_j^k individually). This memory efficient implementation is similar as the one explained in (Koloskova et al., 2019).

5 Convergence Analysis

We summarize the convergence results of Algorithm 1 in Theorem 5. All proofs are given in the supplementary materials, restated as Theorems 12, 14, 16, 19.

Theorem 5. *Under Assumptions 1–3, for any given $\epsilon > 0$ Algorithm 1 with stepsizes $\alpha = \frac{1}{\omega+1}$, $\eta = \frac{1}{L}$ and $\theta = \Theta\big(\frac{\mu}{\lambda_{\max}(\mathbf{W}) + \omega\max_{(i,j)\in E} w_{ij}}\big)$ for option A, B and D and chosen as in equation (50) for option C reaches accuracy $\|x - x^\star\|^2 \leq \epsilon$ on all nodes after the following number of iterations T:*

Options A/B: *dual $\nabla f_i^\star(z)$ / primal $\nabla f_i(x)$ available*

$$T = \mathcal{O}\left((\omega + \kappa(\rho + \omega\rho_\infty))\log\frac{1}{\epsilon}\right)$$

Option C: *stochastic $\nabla f_i(x,\xi)$ available*

$$T = \tilde{\mathcal{O}}\left(\omega + (\rho + \omega\rho_\infty)\left(\kappa + \frac{\sigma\sqrt{1+\omega}}{\sqrt{\epsilon}\mu} + \frac{\sigma^2(\rho+\omega\rho_\infty)}{\epsilon\mu^2}\right)\right)$$

Option D: *finite sum $f_i(x) = \frac{1}{m}\sum_{j=1}^m f_{ij}(x)$*

$$T = \mathcal{O}\left((m + \omega + \kappa(\rho + \omega\rho_\infty))\log\frac{1}{\epsilon}\right),$$

where $\tilde{\mathcal{O}}(\cdot)$ hides logarithmic factors.

For **Option A** and **B** we obtain the same linear convergence rate. For $\omega = 0$ the rate simplifies to $\tilde{\mathcal{O}}(\kappa\rho)$, which is the product of the condition number of f (difficulty of the optimization problem) and the spectral gap of $\mathbf{W}$ (how fast information diffuses in the graph).

The dependence on these parameters can be improved to their square roots with accelerated gradient methods cf. (Scaman et al., 2017; Uribe et al., 2018). In particular, our scheme fits the Catalyst framework (Lin et al., 2015) that can potentially be used to derive optimal accelerated rates with restarts. However, indirect acceleration via Catalyst might not give the best practical scheme, and direct acceleration would be preferred (though not derived in this work).

Our result recovers the best known rates for non-accelerated algorithms (Alghunaim and Sayed, 2020) and in the centralized setting ($\rho = 1$) we recover the rate of the standard gradient method. In contrast to the method proposed in (Reisizadeh et al., 2018, 2019) we are here able to show linear convergence for our scheme even with quantization, i.e., for $\omega > 0$. Liu et al. (2020) independently show convergence rate $\tilde{\mathcal{O}}(\kappa\rho\omega)$ for **option B**, whereas in our result, $\tilde{\mathcal{O}}(\kappa\rho + \kappa\rho_\infty\omega)$, the dependency on ω can be weaker. Magnússon et al. (2020) show linear convergence for **option A** but only for ω small enough, not an arbitrary parameter as considered here.[4]

The linear $\mathcal{O}(\omega)$ term that appears in all our results is not crucial, as sending ω-quantized vectors is typically $\mathcal{O}(\omega)$ times faster than sending uncompressed vectors (consider random-k quantization as a guiding example), thus $\mathcal{O}(\omega)$ is proportional to the time it takes to send one single unquantized vector between two nodes.

Compression for free. Note that for any choice of ω for which $\omega + \kappa(\rho + \omega\rho_\infty) = \mathcal{O}(\kappa\rho)$, or in other words,

$$\omega \leq \min\left\{\rho\rho_\infty^{-1}, \kappa\rho\right\}, \tag{11}$$

the total number of iteration does not increase but the number of bits send in each iteration can be decreased.

As explained in Remark 2, the ratio $\rho\rho_\infty^{-1}$ can reach size $\Theta(n)$, in particular for k-regular graphs with uniform weights, $\rho\rho_\infty^{-1} = \Theta(k)$. Hence ω can be chosen as large as $\Theta(n)$ for graphs with large maximal degree Δ. As a second example, consider a star graph with a

[4]Their quantization framework requires the compressed messages to have size (in bits b) at least $b \geq b_{\mathrm{c}}$, where b_{c} is a critical value, discussed below Theorem 1 in (Magnússon et al., 2020).

central node connected to all other nodes and uniform edge weights, $\lambda_{\max}(\mathbf{W})$ and the the spectral gap are both of order $\Omega(n)$, so that the choice $\omega = \mathcal{O}(n)$ is admissible. For well connected graphs (such as regular graphs), the second term in (11) becomes smaller, but for difficult optimization problems with $\kappa = \Omega(\Delta)$ we see that compression up to $\omega = \mathcal{O}(\Delta)$ is possible without affecting the convergence rate.

In **option D** we leverage the finite-sum structure of f_i. In each iteration only a single new gradient ∇f_{ij} has to be computed (unless a full pass over the local dataset is triggered). Our method combines SVRG-style variance reduction (reducing the variance of the stochastic gradients) with our new variance reduction technique for quantized communication to achieve linear convergence on decentralized networks. For $\omega = 0$ and $\rho = 1$ we recover the convergence rate of SVRG and for $\rho > 1$ our rate improves over the $\tilde{\mathcal{O}}(m + \kappa^2\rho^2)$ convergence rate of the recently proposed GT-SVRG (Xin et al., 2019) which does not support quantization.

For **option C**, with stochastic updates, we observe that our convergence rate recovers the linear rate of option **A** and **B**, when $\sigma^2 \to 0$. However, when σ^2 is large, the rate is dominated by the $\mathcal{O}\big(\frac{\sigma^2}{\epsilon}\big)$ term, and the algorithm only converges sublinearly.

6 Experiments

In this section we experimentally validate our theoretical findings.

Setup. We use rand-k and dit-k quantization functions (see Example 4). We choose two unweighted ($w_{ij} = 1$ for $(i,j) \in E$) graphs on n nodes for our experiments: The *ring*, where every node is connected to two neighbours. As it holds $\rho \approx \rho_\infty \approx n^2$ we see that this is a challenging topology, only allowing communication compression for $\omega = \mathcal{O}(1)$ (see also Remark 2). Further, the *star* graph, where $(n-1)$ nodes have no direct links between them, but are all connected to the central node. Here it holds $\rho = n$, $\rho_\infty = 1$ and compression for $\omega = \mathcal{O}(n)$ is suggested by our theory.

As baselines we use decentralized gradient descent algorithms with quantized communications designed for convex cases: QDGD (Reisizadeh et al., 2019), Choco-Gossip and Choco-SGD (Koloskova et al., 2019) for consensus and logistic regression correspondingly. Note that when the compression function is identity ($\omega = 0$), Choco-SGD recovers D-SGD (Nedić and Ozdaglar, 2009), and our Algorithm 1 recovers Primal Dual GD (Scaman et al., 2017; Alghunaim and Sayed, 2020). In all our experiments we tune the hyperparameters of these algorithms independently over a logarithmic grid.

Average consensus. First, we illustrate the performance of Algorithm 1 on the average consensus problem where every worker i has a vector $x_i \in \mathbb{R}^d$ and the goal is to find the average $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$. We generate vectors x_i from normal distribution $\mathcal{N}(0, \mathbf{I})$. This can be cast into decentralized optimization formulation (1) by considering functions of the form $f_i(x) = \frac{1}{2}\|x - x_i\|_2^2$. Note that these f_i's are strongly convex and smooth with $L = \mu = 1$ (Assumption 1), we set $\eta = \frac{1}{L} = 1$ and tune the stepsize θ for our algorithm. In this setup we can easily compute full gradients. Moreover, both **option A** and **option B** of Algorithm 1 lead to the same update.

In Figure 1 we see that for the challenging ring topology almost any quantization level ω leads to an increase in the total number of iterations. On the other hand, as predicted by theory, for the star graph there is a level up to which quantization does not affect the convergence, and we can achieve communication savings for free.

In Figure 2 we compare our algorithms to the baselines. Even after tuning the stepsizes, QDGD converges very slowly (in agreement with Table 1). On both graphs, iteration-wise our algorithm converges faster than Choco. However, in terms of number of bits, Algorithm 1 converges slightly slower than Choco on the ring graph. This is because our Algorithm 1 requires twice as large messages compared to Choco for the same quantization level. However, even with this slight disadvantage, our algorithm performs best on the star graph in term of bits.

Logistic regression. We further assess performance on logistic regression with the objective function

$$f(x) = \frac{1}{m}\sum_{j=1}^m \log(1 + \exp(-b_j \langle a_j, x\rangle)) + \frac{1}{2m}\|x\|_2^2,$$

where $a_j \in \mathbb{R}^d$, $b_j \in \{-1, 1\}$. We use the W8A dataset (Platt, 1998) and distribute the samples between machines equally in a non-iid way, sorted by label. We use ring topology with $n = 16$ nodes. We compare two cases: either the nodes compute gradients on their full local batch (Figure 4, top), or stochastic gradients with respect to one single (randomly selected) local data sample (bottom). We tune all algorithms to reach best performance after 200 epochs in the full batch case and for 300 epochs in stochastic case (left). To plot performance in terms of transmitted number of bits (right), we run the algorithms longer with found parameters.

With local gradients available, our algorithm converges faster than the baselines. This is supported by the theory, as we prove linear convergence for our **option B**[5],

[5]On Figure 3 we do not see perfect linear convergence

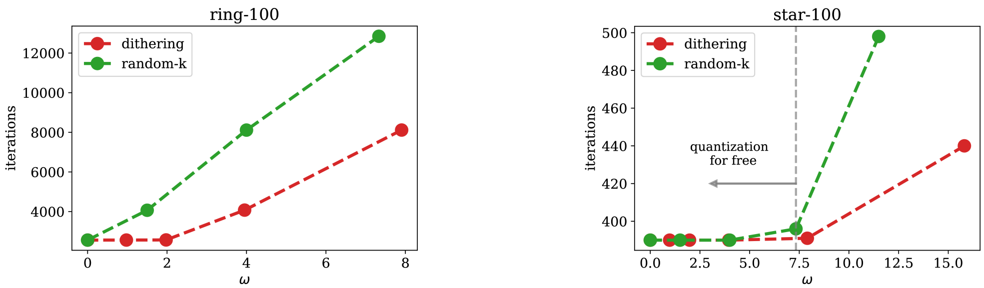

Figure 1: Illustrating quantization for free (right vs. left). Iterations to converge to 10^{-3} error for Algorithm 1 (option B) with different quantization functions. Average consensus problem on the *star* and *ring* topologies with $n = 100$ nodes, $d = 250$ and (rand-k) and (dit-k) compression.

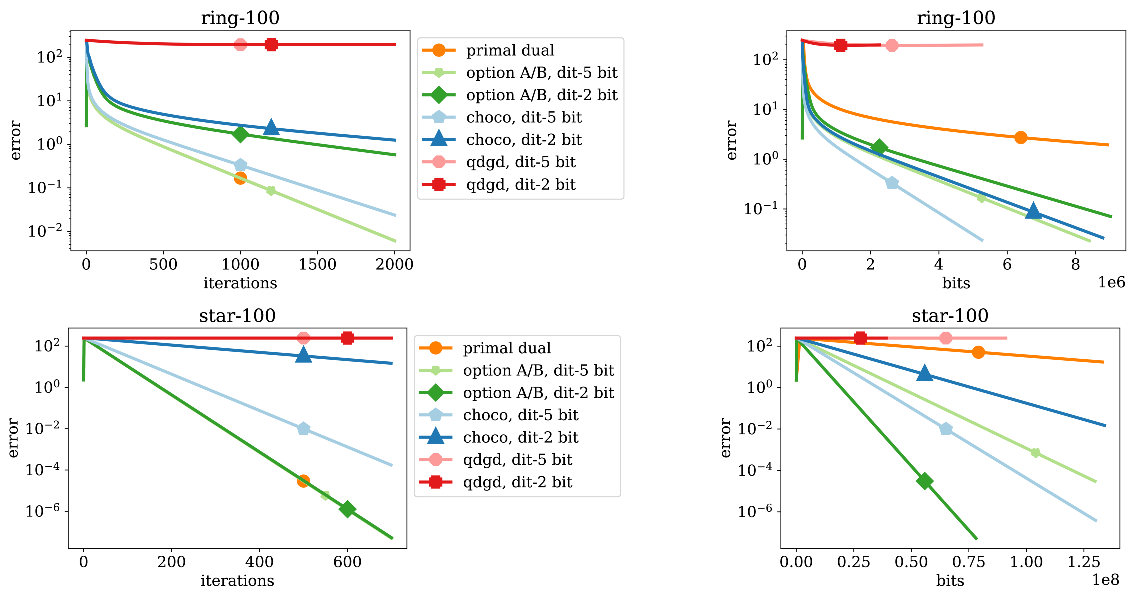

Figure 2: Comparison to the baselines. Average consensus problem on the star and ring topologies with $n = 100$ nodes, $d = 250$ and (rand-k) and (dit-k) compression.

while all other baselines converge only sublinearly (Table 1). With stochastic gradients, **option C** as good as the Choco baseline, while **option D** outperforms all schemes (we have proven linear rate).

Acknowledgements

We acknowledge funding from SNSF grant 200021_175796, as well as a Google Focused Research Award.

References

Sulaiman Alghunaim, Kun Yuan, and Ali H Sayed. A linearly convergent proximal gradient algorithm for decentralized optimization. In *NeurIPS - Advances in Neural Information Processing Systems 32*, pages 2848–2858. Curran Associates, Inc., 2019.

Sulaiman A. Alghunaim and Ali H. Sayed. Linear convergence of primal-dual gradient methods and their performance in distributed optimization. *arXiv preprint arXiv:1904.01196v2*, 2020. (improved rate in v2).

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *NIPS - Advances in Neural Information Processing Systems 30*, pages 1709–1720. Curran Associates, Inc., 2017.

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cedric Renggli. The convergence of sparsified gradient methods. In *NeuriPS - Advances in Neural Information Processing Systems 31*, pages 5977–5987. Curran Associates, Inc., 2018.

Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and sta-

for **option B** because algorithms were tuned for the fixed number of epochs, and the tuned learning rate could be bigger than suggested by the theory. In Appendix B we verify that by taking a smaller stepsize, **option B** attains linear convergence.

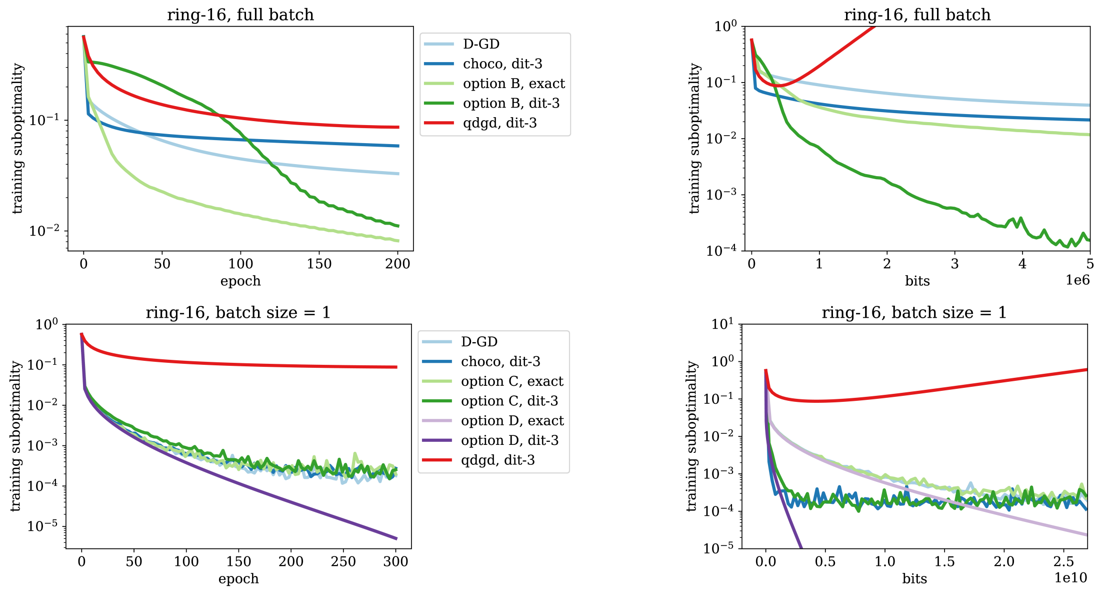

Figure 3: Logistic regression on w8a dataset. Comparison to the baselines for full batch GD (top) and stochastic GD (bottom).

tistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, 2011.

R. Carli, F. Bullo, and S. Zampieri. Quantized average consensus via dynamic coding/decoding schemes. *International Journal of Robust and Nonlinear Control*, 20: 156–175, 2010a.

R. Carli, P. Frasca, F. Fagnani, and S. Zampieri. Gossip consensus algorithms via quantized communication. *Automatica*, 46:70–80, 2010b.

George H-G. Chen and R. T. Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc D'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc V. Le, and Andrew Y. Ng. Large scale distributed deep networks. In *NIPS - Advances in Neural Information Processing Systems*, pages 1223–1231, 2012.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS - Advances in Neural Information Processing Systems 27*, pages 1646–1654. Curran Associates, Inc., 2014.

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

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17 – 40, 1976.

R. Glowinski and A. Marrocco. On the solution of a class of non linear dirichlet problems by a penalty-duality method and finite elements of order one. In Marchuk G.I., editor, *Optimization Techniques IFIP Technical Conference*, LNCS, 1975.

W. M. Goodall. Television by pulse code modulation. *The Bell System Technical Journal*, 30(1):33–49, 1951.

Robert Hannah, Yanli Liu, Daniel O'Connor, and Wotao Yin. Breaking the span assumption yields fast finite-sum minimization. *NeurIPS - Neural Information Processing Systems*, 2018.

Lie He, An Bian, and Martin Jaggi. COLA: Decentralized linear learning. In *NeurIPS - Advances in Neural Information Processing Systems 31*, pages 4541–4551, 2018.

Hadrien Hendrikx, Francis Bach, and Laurent Massoulie. An optimal algorithm for decentralized finite sum optimization. *arXiv preprint arXiv:2005.10675*, 2020.

Samuel Horváth, Dmitry Kovalev, Konstantin Mishchenko, Peter Richtárik, and Sebastian U. Stich. Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*, 2019.

Franck Iutzeler, Pascal Bianchi, Philippe Ciblat, and Walid Hachem. Asynchronous distributed optimization using a randomized alternating direction method of multipliers. In *Proceedings of the 52nd IEEE Conference on Decision and Control, CDC 2013, December 10-13, 2013, Firenze, Italy*, pages 3671–3676. IEEE, 2013.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS - Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian U. Stich, and Martin Jaggi. Error feedback fixes SignSGD and other gradient compression schemes. In *ICML - 36th International Conference on Machine Learning*, volume 97, pages 3252–3261. PMLR, 2019.

Anastasia Koloskova, Sebastian U. Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *ICML - 36th International Conference on Machine Learning*, volume 97, pages 3478–3487. PMLR, 2019.

Anastasia Koloskova, Tao Lin, Sebastian U. Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. *ICLR - International Conference on Learning Representations*, 2020a.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U. Stich. A unified theory of decentralized SGD with changing topology and local updates. *arXiv preprint arXiv:2003.10422*, 2020b.

Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 2018.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Advances in Neural Information Processing Systems 30*, pages 5330–5340. Curran Associates, Inc., 2017.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. In *NIPS - Advances in Neural Information Processing Systems 28*, 2015.

Xiaorui Liu, Yao Li, Rongrong Wang, Jiliang Tang, and Ming Yan. Linear convergent decentralized optimization with compression. *arXiv preprint arXiv:2007.00232*, 2020.

Sindri Magnússon, Hossein Shokri-Ghadikolaei, and Na Li. On maintaining linear convergence of distributed learning and optimization under limited communication. *IEEE Transactions on Signal Processing*, 68:6101–6116, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS 2017 - Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1273–1282, 2017.

H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016.

Aryan Mokhtari and Alejandro Ribeiro. DSA: Decentralized double stochastic averaging gradient algorithm. *Journal of Machine Learning Research*, 17(1): 2165–2199, 2016.

A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

A. Nedić, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4): 2597–2633, 2017.

Angelia Nedić. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020.

Angelia Nedić, Alex Olshevsky, Asuman Ozdaglar, and John N. Tsitsiklis. Distributed subgradient methods and quantization effects. In *Proceedings of the 47th IEEE Conference on Decision and Control, CDC 2008*, pages 4177–4184, 2008. ISBN 9781424431243.

Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems 29*, pages 1416–1424. Curran Associates, Inc., 2016.

John C. Platt. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods - Support Vector Learning, Cambridge, MA, 1998. MIT Press.*, 1998.

S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Math. Program.*, 2020.

G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 159–166, 2016.

M. Rabbat. Multi-agent mirror descent for decentralized stochastic optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 517–520, 2015.

A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani. An exact quantized decentralized gradient descent algorithm. *IEEE Transactions on Signal Processing*, 67(19): 4934–4947, 2019.

Amirhossein Reisizadeh, Aryan Mokhtari, S. Hamed Hassani, and Ramtin Pedarsani. Quantized decentralized consensus optimization. *arXiv preprint arXiv:1806.11536*, 2018.

L. Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8(2):145–154, 1962.

Ali Sayed. Adaptation, learning, and optimization over networks. *Found. Trends Mach. Learn.*, 7(4–5):311–801, 2014.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *ICML - 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3027–3036. PMLR, 2017.

Frank Seide and Amit Agarwal. CNTK: Microsoft's open-source deep-learning toolkit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 2135, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In Haizhou Li, Helen M. Meng, Bin Ma, Engsiong Chng, and Lei Xie, editors, *INTERSPEECH*, pages 1058–1062. ISCA, 2014.

W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2): 944–966, 2015.

Sebastian U. Stich and Sai P. Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*, 2019.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *NeurIPS - Advances in Neural Information Processing Systems 31*, pages 4452–4463. Curran Associates, Inc., 2018.

Nikko Strom. Scalable distributed dnn training using commodity gpu cloud computing. In *INTERSPEECH*, pages 1488–1492. ISCA, 2015.

Ananda Theertha Suresh, Felix X. Yu, Sanjiv Kumar, and H. Brendan McMahan. Distributed mean estimation with limited communication. In *ICML - 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3329–3337. PMLR, 2017.

Hanlin Tang, Shaoduo Gan, Ce Zhang, Tong Zhang, and Ji Liu. Communication compression for decentralized training. In *NeurIPS - Advances in Neural Information Processing Systems 31*, pages 7663–7673. Curran Associates, Inc., 2018.

Hanlin Tang, Xiangru Lian, Shuang Qiu, Lei Yuan, Ce Zhang, Tong Zhang, and Ji Liu. Deepsqueeze: Decentralization meets error-compensated compression. *arXiv preprint arXiv:1907.07346*, 2019.

John N. Tsitsiklis. *Problems in decentralized decision making and computation.* PhD thesis, Massachusetts Institute of Technology, 1984.

César A Uribe, Soomin Lee, and Alexander Gasnikov. A dual approach for optimal algorithms in distributed optimization over networks. *arXiv preprint arXiv:1809.00710*, 2018.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *NeurIPS - Advances in Neural Information Processing Systems 31*, pages 1306–1316. Curran Associates, Inc., 2018.

E. Wei and A. Ozdaglar. Distributed alternating direction method of multipliers. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5445–5450, 2012.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *NIPS - Advances in Neural Information Processing Systems 30*, pages 1509–1519. Curran Associates, Inc., 2017.

L. Xiao, S. Boyd, and S. Lall. A scheme for robust distributed sensor fusion based on average consensus. In *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.*, pages 63–70, 2005.

R. Xin, S. Kar, and U. A. Khan. Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence. *IEEE Signal Processing Magazine*, 37(3): 102–113, 2020.

Ran Xin, Usman A. Khan, and Soummya Kar. Variance-reduced decentralized stochastic optimization with accelerated convergence. *arXiv preprint arXiv:1912.04230*, 2019.

Deming Yuan, Shengyuan Xu, Huanyu Zhao, and Lina Rong. Distributed dual averaging method for multi-agent optimization with quantized communication. *Systems & Control Letters*, 61(11):1053 – 1061, 2012.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

A Parameters for the Numerical Experiments

In this section we give the hyperparameters we used for the for the experiments in the paper (found by grid search).

	ring-100	star-100
method	θ	θ
exact (no quantization)	1.26	1.58
dit-17	1.26	1.58
dit-9	1.26	1.58
dit-5	1.26	1.58
dit-4	1.26	1.58
dit-3	0.79	1.58
dit-2	0.4	1.58
dit-1	-	1.58
rand-100	0.8	1.58
rand-50	0.4	1.58
rand-30	0.25	1.58
rand-20	0.2	1.58
rand-10	0.1	1.0

Table 3: Hyperparameters found by tuning (lowest iteration number to reach target accuracy) in experiments for Fig. 1.

	ring-100	star-100
method	θ	θ
primal dual	1.26	1.58
option A/B, dit-5	1.26	1.58
option A/B, dit-2	0.4	1.58
method	γ	γ
choco, dit-5	1	1
choco, dit-2	0.25	0.2
method	ϵ, α	ϵ, α
qdgd, dit-5	(0.001, 1.0)	(0.0001, 10000.0)
qdgd, dit-2	(0.001, 1.0)	(0.0001, 10000.0)

Table 4: Hyperparameters found by tuning in experiments for Fig. 2.

batch size = 1

method	γ	η
D-GD	-	0.1
choco, dit-3	0.316	0.1
method	θ	η
option C, exact	0.1	0.1
option C, dit-3	3.16×10^{-3}	0.1
option D, exact	3.16	0.1
option D, dit-3	3.16×10^{-3}	0.1
method	ϵ	α
qdgd, dit-3	1e-06	100

full batch

method	γ	η
D-GD	-	21.5
choco, dit-3	0.316	4.64
method	θ	η
option B, exact	0.1	21.5
option B, dit-3	3.5×10^{-4}	40
method	ϵ	α
qdgd, dit-3	0.01	31.6

Table 5: Hyperparameters found by tuning (lowest error after 200 epochs for full batch, and 300 epochs for batch size 1) in experiments for Fig. 4.

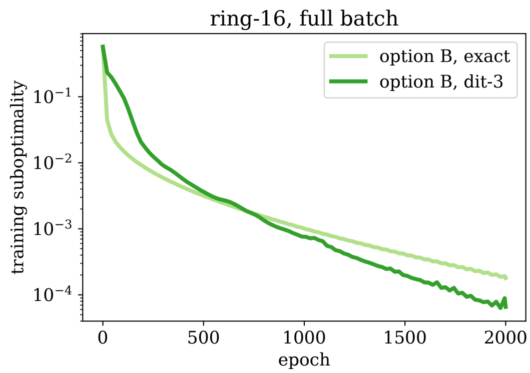

Figure 4: Logistic regression on w8a dataset. **option B, exact** with $\eta = 20$, $\gamma = 0.05$, **option B, dit-3** with $\eta = 30$, $\gamma = 3 \times 10^{-4}$.

B Additional Plots

In this section we verify that for the same setting as for the Figure 4 (Section 6) **option B** of Algorithm 1 converges linearly if we take the smaller stepsizes.

C Convergence Proof

To prove the convergence of Algorithm 1, we will use matrix notation for the iterates of the algorithm:

- primal iterates
$$\mathbf{X}^k = \left[x_1^k, \ldots, x_n^k\right] \in \mathbb{R}^{d\times n},$$
- and dual iterates
$$\mathbf{Z}^k = \left[z_1^k, \ldots, z_n^k\right] \in \mathbb{R}^{d\times n}$$
and
$$\mathbf{H}^k = \left[h_1^k, \ldots, h_n^k\right] \in \mathbb{R}^{d\times n}.$$

Let $\mathbf{X}^\star = \underbrace{[x^\star, \ldots, x^\star]}_{n \text{ times}}$, $\mathbf{Z}^\star = [z_1^\star, \ldots, z_n^\star]$, where $x^\star$ is a solution of (1), $z_i^\star = \nabla f_i(x^\star)$ for all $i = 1, \ldots, n$.

For arbitrary matrix $\mathbf{B} \in \mathbb{S}_+^n$, we define matrix semi-norm (norm in case $\mathbf{B}$ is positive-definite) $\|\cdot\|_{\mathbf{B}} : \mathbb{R}^{d\times n} \to \mathbb{R}_+$, which is defined as follows:

$$\|\mathbf{X}\|_{\mathbf{B}}^2 = \langle \mathbf{X}\mathbf{B}, \mathbf{X}\rangle = \mathbf{tr}\left(\mathbf{X}\mathbf{B}\mathbf{X}^\top\right). \tag{12}$$

Lemma 6 (Properties of $\mathbf{W}$)**.** $\ker \mathbf{W} = \operatorname{span}(\mathbf{1})$*, where* $\mathbf{1} = (1, \ldots, 1)^\top$.

Lemma 7. *For arbitrary* $\mathbf{X} \in \mathbb{R}^{d\times n}$*, the following inequalities hold:*

$$\|\mathbf{X}\|_{\mathbf{W}}^2 \le \lambda_{\max}(\mathbf{W})\|\mathbf{X}\|_{\mathbf{I}}^2, \tag{13}$$

$$\lambda_{\min}^+(\mathbf{W})\|\mathbf{X}\|_{\mathbf{W}^\dagger}^2 \le \|\mathbf{X}\|_{\mathbf{I}}^2. \tag{14}$$

Lemma 8 (h_i^k update)**.** *Let* $\alpha = \frac{1}{\omega+1}$*. The following inequality holds:*

$$\mathbb{E}\left[\|\mathbf{H}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2\right] \le (1-\alpha)\|\mathbf{H}^k - \mathbf{X}^\star\|_{\mathbf{I}}^2 + \alpha\mathbb{E}\left[\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2\right]. \tag{15}$$

Proof. Since $h_i^{k+1} = h_i^k + \alpha\mathcal{Q}(x_i^{k+1} - h_i^k)$, we can decompose

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{H}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2\right] &= \mathbb{E}\left[\|\mathbf{H}^k - \mathbf{X}^\star\|_{\mathbf{I}}^2\right] + 2\alpha\langle \mathbf{X}^{k+1} - \mathbf{H}^k, \mathbf{H}^k - \mathbf{X}^\star\rangle + \alpha^2\mathbb{E}\left[\|\mathbf{X}^{k+1} - \mathbf{H}^k\|_{\mathbf{I}}^2\right] \\
&\le \mathbb{E}\left[\|\mathbf{H}^k - \mathbf{X}^\star\|_{\mathbf{I}}^2\right] + 2\alpha\langle \mathbf{X}^{k+1} - \mathbf{H}^k, \mathbf{H}^k - \mathbf{X}^\star\rangle + \alpha^2(1+\omega)\|\mathbf{X}^{k+1} - \mathbf{H}^k\|_{\mathbf{I}}^2 \\
&\le \mathbb{E}\left[\|\mathbf{H}^k - \mathbf{X}^\star\|_{\mathbf{I}}^2\right] + 2\alpha\langle \mathbf{X}^{k+1} - \mathbf{H}^k, \mathbf{H}^k - \mathbf{X}^\star\rangle + \alpha\|\mathbf{X}^{k+1} - \mathbf{H}^k\|_{\mathbf{I}}^2 \\
&= \mathbb{E}\left[\|\mathbf{H}^k - \mathbf{X}^\star\|_{\mathbf{I}}^2\right] + \alpha\langle \mathbf{X}^{k+1} - \mathbf{H}^k, \mathbf{X}^{k+1} + \mathbf{H}^k - 2\mathbf{X}^\star\rangle \\
&= \mathbb{E}\left[\|\mathbf{H}^k - \mathbf{X}^\star\|_{\mathbf{I}}^2\right] + \alpha\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2 - \alpha\|\mathbf{H}^k - \mathbf{X}^\star\|_{\mathbf{I}}^2 \\
&\le (1-\alpha)\|\mathbf{H}^k - \mathbf{X}^\star\|_{\mathbf{I}}^2 + \alpha\mathbb{E}\left[\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2\right].
\end{aligned}$$

□

Lemma 9 (Dual step)**.**

$$\mathbb{E}\left[\|\mathbf{Z}^{k+1} - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2\right] \le \|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \mathbb{E}\left[-2\theta\langle \mathbf{X}^{k+1} - \mathbf{X}^\star, \mathbf{Z}^k - \mathbf{Z}^\star\rangle + \theta^2\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{W}}^2 + \Sigma^k\right], \tag{16}$$

where Σ^k is a variance:

$$\Sigma^k = \mathbb{E}_Q\left[\|\mathbf{Z}^{k+1} - \mathbb{E}_Q\left[\mathbf{Z}^{k+1}\right]\|_{\mathbf{W}^\dagger}^2\right]. \tag{17}$$

Proof. Firstly, we prove that

$$\sum_{i=1}^n z_i^k = 0 \tag{18}$$

for all $k = 0, 1, 2, \ldots$ by induction. For $k = 0$, (18) follows trivially from initialization step of Algorithm 1. Now, assuming that (18) holds, we have

$$\begin{aligned}
\sum_{i=1}^n z_i^{k+1} &= \sum_{i=1}^n z_i^k - \theta\sum_{j\in\mathcal{N}_i} w_{ij}(\Delta_{ij}^k - \Delta_{ji}^k) = -\theta\sum_{(i,j)\in E} w_{ij}\Delta_{ij}^k + \theta\sum_{(i,j)\in E} w_{ij}\Delta_{ji}^k \\
&= -\theta\sum_{(i,j)\in E} w_{ij}\Delta_{ij}^k + \theta\sum_{(i,j)\in E} w_{ji}\Delta_{ij}^k = \theta\sum_{(i,j)\in E}\Delta_{ij}^k(w_{ji} - w_{ij}) = 0,
\end{aligned}$$

which proves (18) for all $k = 0, 1, 2, \ldots$.

Next, we show that

$$(\mathbf{Z}^k - \mathbf{Z}^\star)\mathbf{W}^\dagger\mathbf{W} = \mathbf{Z}^k - \mathbf{Z}^\star, \tag{19}$$

which holds, since $(\mathbf{Z}^k - \mathbf{Z}^\star)\mathbf{1} = \mathbf{0}$, where $\mathbf{1} = (\mathbf{1}, \ldots, \mathbf{1})^\top \in \mathbb{R}^{\mathbf{n}}$, and hence rows of matrix $\mathbf{Z}^k - \mathbf{Z}^\star$ lie in range of $\mathbf{W}$.

Now, we rewrite $\mathbb{E}_Q\left[\|\mathbf{Z}^{k+1} - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2\right]$:

$$\mathbb{E}_Q\left[\|\mathbf{Z}^{k+1} - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2\right] = \|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \mathbb{E}_Q\left[2\langle \mathbf{Z}^{k+1} - \mathbf{Z}^k, (\mathbf{Z}^k - \mathbf{Z}^\star)\mathbf{W}^\dagger\rangle\right] + \mathbb{E}_Q\left[\|\mathbf{Z}^{k+1} - \mathbf{Z}^k\|_{\mathbf{W}^\dagger}^2\right]. \tag{20}$$

To simplify the second term in (20), we first rewrite $\mathbb{E}_Q\left[z_i^{k+1} - z_i^k\right]$ as follows:

$$\begin{aligned}
\mathbb{E}_Q\left[z_i^{k+1} - z_i^k\right] &= -\theta\sum_{j\in\mathcal{N}_i} w_{ij}\mathbb{E}\left[\Delta_{ij}^k - \Delta_{ji}^k\right] = -\theta\sum_{j\in\mathcal{N}_i} w_{ij}(x_i^{k+1} - x^\star) + \theta\sum_{j\in\mathcal{N}_i} w_{ij}(x_j^{k+1} - x^\star) \\
&= -\theta\sum_{j=1}^n \mathbf{W}_{ij}(x_j^{k+1} - x^\star),
\end{aligned}$$

which gives

$$\mathbb{E}_Q\left[\mathbf{Z}^{k+1} - \mathbf{Z}^k\right] = -\theta(\mathbf{X}^{k+1} - \mathbf{X}^\star)\mathbf{W}. \tag{21}$$

Hence,

$$\mathbb{E}\left[2\langle \mathbf{Z}^{k+1} - \mathbf{Z}^k, (\mathbf{Z}^k - \mathbf{Z}^\star)\mathbf{W}^\dagger\rangle\right] \overset{(21)}{=} -2\theta\langle \mathbf{X}^{k+1} - \mathbf{X}^\star, (\mathbf{Z}^k - \mathbf{Z}^\star)\mathbf{W}^\dagger\mathbf{W}\rangle \overset{(19)}{=} -2\theta\langle \mathbf{X}^{k+1} - \mathbf{X}^\star, \mathbf{Z}^k - \mathbf{Z}^\star\rangle. \tag{22}$$

Now, we simplify the last term in (20):

$$\mathbb{E}_Q\left[\|\mathbf{Z}^{k+1}-\mathbf{Z}^k\|^2_{\mathbf{W}^\dagger}\right] = \|\mathbb{E}_Q\left[\mathbf{Z}^{k+1}-\mathbf{Z}^k\right]\|^2_{\mathbf{W}^\dagger} + \mathbb{E}_Q\left[\|\mathbf{Z}^{k+1}-\mathbb{E}_Q\left[\mathbf{Z}^{k+1}\right]\|^2_{\mathbf{W}^\dagger}\right]$$
$$\overset{(21)}{=} \theta^2\|(\mathbf{X}^{k+1}-\mathbf{X}^\star)\mathbf{W}\|^2_{\mathbf{W}^\dagger} + \Sigma^k = \theta^2\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{W}} + \Sigma^k.$$

Plugging this and (22) into (20) gives

$$\mathbb{E}_Q\left[\|\mathbf{Z}^{k+1}-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger}\right] \le \|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} - 2\theta\langle \mathbf{X}^{k+1}-\mathbf{X}^\star, \mathbf{Z}^k-\mathbf{Z}^\star\rangle + \theta^2\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{W}} + \Sigma^k.$$

Taking full expectation concludes the proof. □

Lemma 10 (Variance bound)**.** *For each $k \ge 0$ we have*

$$\Sigma^k \le 4\theta^2\omega \max_{(i,j)\in E} w_{ij}\left[\|\mathbf{X}^{k+1}-\mathbf{X}_\star\|^2_{\mathbf{I}} + \|\mathbf{H}^k-\mathbf{X}_\star\|^2_{\mathbf{I}}\right]. \tag{23}$$

Proof.

$$\Sigma^k = \mathbb{E}_Q\left[\|\mathbf{Z}^{k+1}-\mathbb{E}_Q\left[\mathbf{Z}^{k+1}\right]\|^2_{\mathbf{W}^\dagger}\right] = \mathbb{E}_Q\sum_{i=1}^n\sum_{j=1}^n \mathbf{W}^\dagger_{ij}\langle z_i^{k+1}-\mathbb{E}_Q z_i^{k+1}, z_j^{k+1}-\mathbb{E}_Q z_j^{k+1}\rangle \tag{24}$$

$$= \sum_{i=1}^n\sum_{j=1}^n \mathbf{W}^\dagger_{ij}\mathbb{E}_Q\left[\langle z_i^{k+1}-\mathbb{E}_Q z_i^{k+1}, z_j^{k+1}-\mathbb{E}_Q z_j^{k+1}\rangle\right], \tag{25}$$

where $\mathbf{W}^\dagger_{ij}$ is number in the i-th row of j-th column of $\mathbf{W}^\dagger$.

Next we observe, that

$$z_i^{k+1}-\mathbb{E}_Q z_i^{k+1} = -\theta\sum_{j\in\mathcal{N}_i} w_{ij}(\bar{\Delta}^k_{ij}-\bar{\Delta}^k_{ji}), \tag{26}$$

where $\bar{\Delta}^k_{ij} := \Delta^k_{ij} - \mathbb{E}_Q\Delta_{ij} = \Delta^k_{ij} - x_i^{k+1}$.

From the construction of Δ^k_{ij} it follows that

$$\mathbb{E}_Q\left[\langle\bar{\Delta}^k_{ab},\bar{\Delta}^k_{cd}\rangle\right] = \begin{cases}\sigma^k_a, & a=c \text{ and } b=d\\ 0, & \text{otherwise}\end{cases} = \delta_{ac}\delta_{bd}\sigma^k_a, \tag{27}$$

where $\sigma^k_i = \mathbb{E}_Q\left[\|Q(x_i^{k+1}-h_i^k)-(x_i^{k+1}-h_i^k)\|_2^2\right]$ and δ_{ij} is a Kronecker delta.

Now, we rewrite $\mathbb{E}_Q\left[\langle z_i^{k+1}-\mathbb{E}_Q z_i^{k+1}, z_j^{k+1}-\mathbb{E}_Q z_j^{k+1}\rangle\right]$:

$$\begin{aligned}
\mathbb{E}_Q\left[\langle z_i^{k+1}-\mathbb{E}_Q z_i^{k+1}, z_j^{k+1}-\mathbb{E}_Q z_j^{k+1}\rangle\right] &\overset{(26)}{=} \theta^2\mathbb{E}_Q\left[\left\langle\sum_{p\in\mathcal{N}_i} w_{ip}(\bar{\Delta}^k_{ip}-\bar{\Delta}^k_{pi}), \sum_{q\in\mathcal{N}_i} w_{jq}(\bar{\Delta}^k_{jq}-\bar{\Delta}^k_{qj})\right\rangle\right]\\
&= \theta^2\mathbb{E}_Q\left[\left\langle\sum_{p\ne i}\mathbf{W}_{ip}(\bar{\Delta}^k_{ip}-\bar{\Delta}^k_{pi}), \sum_{q\ne j}\mathbf{W}_{jq}(\bar{\Delta}^k_{jq}-\bar{\Delta}^k_{qj})\right\rangle\right]\\
&= \theta^2\sum_{p\ne i}\sum_{q\ne j}\mathbf{W}_{ip}\mathbf{W}_{jq}\mathbb{E}_Q\left[\langle\bar{\Delta}^k_{ip}-\bar{\Delta}^k_{pi}, \bar{\Delta}^k_{jq}-\bar{\Delta}^k_{qj}\rangle\right]\\
&= \theta^2\sum_{p\ne i}\sum_{q\ne j}\mathbf{W}_{ip}\mathbf{W}_{jq}\left[\delta_{ij}\delta_{pq}(\sigma^k_i+\sigma^k_p) - \delta_{iq}\delta_{pj}(\sigma^k_i+\sigma^k_p)\right]\\
&= \theta^2\sum_{p\ne i}\mathbf{W}_{ip}(\sigma^k_i+\sigma^k_p)\sum_{q\ne j}\mathbf{W}_{jq}\left[\delta_{ij}\delta_{pq}-\delta_{iq}\delta_{pj}\right]\\
&= \theta^2\delta_{ij}\sum_{p\ne i}\mathbf{W}^2_{ip}(\sigma^k_i+\sigma^k_p)\sum_{q\ne i}\delta_{pq} - \theta^2\mathbf{W}^2_{ij}(\sigma^k_i+\sigma^k_j)\sum_{p\ne i}\delta_{pj}\sum_{q\ne j}\delta_{iq}\\
&= \theta^2\delta_{ij}\sum_{p\ne i}\mathbf{W}^2_{ip}(\sigma^k_i+\sigma^k_p) - \theta^2(1-\delta_{ij})\mathbf{W}^2_{ij}(\sigma^k_i+\sigma^k_j).
\end{aligned}$$

Plugging this into (25) gives

$$
\begin{aligned}
\Sigma^k &= \theta^2 \sum_{i=1}^n \sum_{j=1}^n \mathbf{W}^\dagger_{ij} \delta_{ij} \sum_{p \neq i} \mathbf{W}^2_{ip} (\sigma_i^k + \sigma_p^k) - \theta^2 \sum_{i=1}^n \sum_{j=1}^n \mathbf{W}^\dagger_{ij} (1 - \delta_{ij}) \mathbf{W}^2_{ij} (\sigma_i^k + \sigma_j^k) \\
&= \theta^2 \sum_{i=1}^n \mathbf{W}^\dagger_{ii} \sum_{p \neq i} \mathbf{W}^2_{ip} (\sigma_i^k + \sigma_p^k) - \theta^2 \sum_{i=1}^n \sum_{j \neq i} \mathbf{W}^\dagger_{ij} \mathbf{W}^2_{ij} (\sigma_i^k + \sigma_j^k) \\
&= \theta^2 \sum_{i=1}^n \mathbf{W}^\dagger_{ii} \sum_{j \neq i} \mathbf{W}^2_{ij} (\sigma_i^k + \sigma_j^k) - \theta^2 \sum_{i=1}^n \sum_{j \neq i} \mathbf{W}^\dagger_{ij} \mathbf{W}^2_{ij} (\sigma_i^k + \sigma_j^k) \\
&= \theta^2 \sum_{i=1}^n \sum_{j \neq i} (\mathbf{W}^\dagger_{ii} - \mathbf{W}^\dagger_{ij}) \mathbf{W}^2_{ij} (\sigma_i^k + \sigma_j^k) = \theta^2 \sum_{i=1}^n \sum_{j \neq i} (\mathbf{W}^\dagger_{ii} - \mathbf{W}^\dagger_{ij}) \mathbf{W}^2_{ij} (\sigma_i^k + \sigma_j^k) \\
&= 2\theta^2 \sum_{i=1}^n \sum_{j \neq i} (\mathbf{W}^\dagger_{ii} - \mathbf{W}^\dagger_{ij}) \mathbf{W}^2_{ij} \sigma_i^k = 2\theta^2 \sum_{i=1}^n \sigma_i^k \sum_{j \neq i} (\mathbf{W}^\dagger_{ii} - \mathbf{W}^\dagger_{ij}) \mathbf{W}^2_{ij} \\
&= 2\theta^2 \sum_{i=1}^n \sigma_i^k \left(\mathbf{W}^\dagger_{ii} \sum_{j \neq i} \mathbf{W}^2_{ij} - \sum_{j \neq i} \mathbf{W}^\dagger_{ij} \mathbf{W}^2_{ij} \right) = 2\theta^2 \sum_{i=1}^n \sigma_i^k \sum_{j=1}^n \mathbf{W}^\dagger_{ij} \hat{\mathbf{W}}_{ij},
\end{aligned}
$$

where $\hat{\mathbf{W}}$ is another Laplacian matrix:

$$
\hat{\mathbf{W}}_{ij} = \begin{cases} -\mathbf{W}^2_{ij}, & i \neq j \\ \sum_{l \neq j} \mathbf{W}^2_{il}, & i = j \end{cases}. \tag{28}
$$

This gives us

$$
\begin{aligned}
\Sigma^k &= 2\theta^2 \sum_{i=1}^n \sigma_i^k \sum_{j=1}^n \mathbf{W}^\dagger_{ij} \hat{\mathbf{W}}_{ij} = 2\theta^2 \sum_{i=1}^n \sigma_i^k [\mathbf{W}^\dagger \hat{\mathbf{W}}]_{ii} = 2\theta^2 \sum_{i=1}^n \sigma_i^k [\mathbf{W}^{\dagger/2} \hat{\mathbf{W}} \mathbf{W}^{\dagger/2}]_{ii} \\
&\leq 2\theta^2 \lambda_{\max}(\mathbf{W}^{\dagger/2} \hat{\mathbf{W}} \mathbf{W}^{\dagger/2}) \sum_{i=1}^n \sigma_i^k.
\end{aligned}
$$

Now, we use the definition of σ_i^k and get

$$
\begin{aligned}
\Sigma^k &\leq 2\theta^2 \lambda_{\max}(\mathbf{W}^{\dagger/2} \hat{\mathbf{W}} \mathbf{W}^{\dagger/2}) \sum_{i=1}^n \mathbb{E}_Q \left[\| Q(x_i^{k+1} - h_i^k) - (x_i^{k+1} - h_i^k) \|_2^2 \right] && (29) \\
&\leq 2\theta^2 \omega \lambda_{\max}(\mathbf{W}^{\dagger/2} \hat{\mathbf{W}} \mathbf{W}^{\dagger/2}) \sum_{i=1}^n \omega \| x_i^{k+1} - h_i^k \|_2^2 = 2\theta^2 \omega \lambda_{\max}(\mathbf{W}^{\dagger/2} \hat{\mathbf{W}} \mathbf{W}^{\dagger/2}) \| \mathbf{X}^{k+1} - \mathbf{H}^k \|_{\mathbf{I}}^2 && (30) \\
&\leq 4\theta^2 \omega \lambda_{\max}(\mathbf{W}^{\dagger/2} \hat{\mathbf{W}} \mathbf{W}^{\dagger/2}) \left[\| \mathbf{X}^{k+1} - \mathbf{X}_\star \|_{\mathbf{I}}^2 + \| \mathbf{H}^k - \mathbf{X}_\star \|_{\mathbf{I}}^2 \right]. && (31)
\end{aligned}
$$

It remains to upper-bound $\lambda_{\max}(\mathbf{W}^{\dagger/2} \hat{\mathbf{W}} \mathbf{W}^{\dagger/2})$. We note that for all $u \in \mathbb{R}^n$

$$
\begin{aligned}
u^\top \hat{\mathbf{W}} u &= \sum_{i=1}^n \sum_{j=1}^n \hat{\mathbf{W}}_{ij} u_i u_j = \sum_{i=1}^n u_i^2 \sum_{j \neq i} \mathbf{W}^2_{ij} - \sum_{i=1}^n \sum_{j \neq i} u_i u_j \mathbf{W}^2_{ij} \\
&= \frac{1}{2} \sum_{i=1}^n \sum_{j \neq i} (u_i^2 + u_j^2) \mathbf{W}^2_{ij} - \sum_{i=1}^n \sum_{j \neq i} u_i u_j \mathbf{W}^2_{ij} = \frac{1}{2} \sum_{i=1}^n \sum_{j \neq i} (u_i - u_j)^2 \mathbf{W}^2_{ij} \\
&\leq \frac{1}{2} \max_{(i,j) \in E} w_{ij} \sum_{i=1}^n \sum_{j \neq i} (u_i - u_j)^2 \mathbf{W}_{ij} = \max_{(i,j) \in E} w_{ij} \cdot u^\top \mathbf{W} u.
\end{aligned}
$$

Hence, for all $u \in \mathbb{R}^n$

$$
u^\top \mathbf{W}^{\dagger/2} \hat{\mathbf{W}} \mathbf{W}^{\dagger/2} u \leq \max_{(i,j) \in E} w_{ij} \cdot u^\top \mathbf{W}^{\dagger/2} \mathbf{W} \mathbf{W}^{\dagger/2} u = \max_{(i,j) \in E} w_{ij} \cdot u^\top \mathbf{W}^\dagger \mathbf{W} u \leq \max_{(i,j) \in E} w_{ij} \cdot \|u\|_2^2
$$

and thus $\lambda_{\max}(\mathbf{W}^{\dagger/2}\hat{\mathbf{W}}\mathbf{W}^{\dagger/2})$ is bounded by

$$\lambda_{\max}(\mathbf{W}^{\dagger/2}\hat{\mathbf{W}}\mathbf{W}^{\dagger/2}) \leq \max_{(i,j)\in E} w_{ij}. \tag{32}$$

plugging this into (31) gives (23) and concludes the proof.

□

C.1 Option A

Lemma 11 (Primal step, Option A)**.** *The following inequality holds:*

$$-2\theta\langle \mathbf{X}^{k+1} - \mathbf{X}^\star, \mathbf{Z}^k - \mathbf{Z}^\star\rangle \leq -\frac{\theta}{L}\|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{I}}^2 - \theta\mu\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2. \tag{33}$$

Proof. From Line 5 of Algorithm 1 it follows that

$$\begin{aligned}
-2\theta\langle \mathbf{X}^{k+1} - \mathbf{X}^\star, \mathbf{Z}^k - \mathbf{Z}^\star\rangle &= -2\theta\sum_{i=1}^n \langle \nabla f_i^\star(z_i^k) - \nabla f_i^\star(z_i^\star), z_i^k - z_i^\star\rangle\\
&\leq -\frac{\theta}{L}\sum_{i=1}^n \|z_i^k - z_i^\star\|_2^2 - \theta\mu\sum_{i=1}^n \|\nabla f_i^*(z_i^k) - \nabla f_i^*(z_i^\star)\|_2^2\\
&= -\frac{\theta}{L}\|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{I}}^2 - \theta\mu\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2,
\end{aligned}$$

where we used strong monotonicity and co-coercivity of ∇f_i^* in the inequality. □

Theorem 12 (Convergence of Algorithm 1, Option A)**.** *Let Ψ_A^k be a Lyapunov function which is defined as follows:*

$$\Psi_A^k = \|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k - \mathbf{X}^\star\|_{\mathbf{I}}^2. \tag{34}$$

Let ρ_A be defined as follows:

$$\rho_A = \max\left\{2(\omega+1), \frac{L(\lambda_{\max}(\mathbf{W}) + 12\omega\max_{(i,j)\in E} w_{ij})}{\mu\lambda_{\min}^+(\mathbf{W})}\right\}^{-1}. \tag{35}$$

Choosing the stepsize θ as

$$\theta = \frac{\mu}{\lambda_{\max}(\mathbf{W}) + 12\omega\max_{(i,j)\in E} w_{ij}} \tag{36}$$

and stepsize $\alpha = \frac{1}{\omega+1}$ gives the following inequality:

$$\mathbb{E}\left[\Psi_A^{k+1}\right] \leq (1-\rho_A)\Psi_A^k. \tag{37}$$

Proof. We start with rewriting (16):

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{Z}^{k+1} - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2\right] &\overset{(16)}{\leq} \|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \mathbb{E}\left[-2\theta\langle \mathbf{X}^{k+1} - \mathbf{X}^\star, \mathbf{Z}^k - \mathbf{Z}^\star\rangle + \theta^2\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{W}}^2 + \Sigma^k\right]\\
&\overset{(33)}{\leq} \|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 - \frac{\theta}{L}\|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{I}}^2 - \theta\mu\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2 + \theta^2\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{W}}^2 + \mathbb{E}\left[\Sigma^k\right]\\
&\overset{(13),(14)}{\leq} \left(1 - \frac{\theta\lambda_{\min}^+(\mathbf{W})}{L}\right)\|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 - \theta(\mu - \theta\lambda_{\max}(\mathbf{W}))\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2 + \mathbb{E}\left[\Sigma^k\right]\\
&\overset{(23)}{\leq} \left(1 - \frac{\theta\lambda_{\min}^+(\mathbf{W})}{L}\right)\|\mathbf{Z}^k - \mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + 4\theta^2\omega\max_{(i,j)\in E} w_{ij}\|\mathbf{H}^k - \mathbf{X}_\star\|_{\mathbf{I}}^2\\
&- \quad \theta\left(\mu - \theta\left[\lambda_{\max}(\mathbf{W}) + 4\omega\max_{(i,j)\in E} w_{ij}\right]\right)\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|_{\mathbf{I}}^2.
\end{aligned}$$

Now, we combine this with (15):

$$\begin{aligned}
\mathbb{E}\left[\Psi_A^{k+1}\right] &\leq \left(1-\frac{\theta\lambda^+_{\min}(\mathbf{W})}{L}\right)\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + 4\theta^2\omega \max_{(i,j)\in E} w_{ij}\|\mathbf{H}^k-\mathbf{X}_\star\|^2_{\mathbf{I}} \\
&\quad - \theta\left(\mu - \theta\left[\lambda_{\max}(\mathbf{W}) + 4\omega \max_{(i,j)\in E} w_{ij}\right]\right)\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}} \\
&\quad + (1-\alpha)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} + 8\theta^2\omega\max_{(i,j)\in E} w_{ij}\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}} \\
&= \left(1-\frac{\theta\lambda^+_{\min}(\mathbf{W})}{L}\right)\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + \left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} \\
&\quad + \theta\left(\mu - \theta\left[\lambda_{\max}(\mathbf{W}) + 12\omega \max_{(i,j)\in E} w_{ij}\right]\right)\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}.
\end{aligned}$$

Using (36) we get

$$\begin{aligned}
\mathbb{E}\left[\Psi_A^{k+1}\right] &\leq \left(1-\frac{\theta\lambda^+_{\min}(\mathbf{W})}{L}\right)\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + \left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} \\
&\leq (1-\rho_A)\Psi_A^k,
\end{aligned}$$

which concludes the proof. □

C.2 Option B

Lemma 13 (Primal step, Option B)**.** *Let $\eta \leq \frac{1}{L}$. Then the following inequality holds:*

$$\begin{aligned}
-2\theta\langle \mathbf{X}^{k+1}-\mathbf{X}^\star, \mathbf{Z}^k-\mathbf{Z}^\star\rangle &\leq -\eta\theta\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{I}} - \frac{\theta\mu}{2}\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}} \\
&\quad + (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}.
\end{aligned} \tag{38}$$

Proof. From Line 6 of Algorithm 1 it follows that

$$\begin{aligned}
\|x_i^{k+1} - x^\star - \eta(z_i^k - z_i^\star)\|_2^2 &= \|x_i^k - x^\star - \eta(\nabla f_i(x_i^k) - \nabla f_i(x^\star))\|_2^2 \\
&\leq (1-\eta\mu)\|x_i^k - x_i^\star\|_2^2
\end{aligned}$$

for any stepsize $\eta \leq \frac{1}{L}$, which leads to

$$\begin{aligned}
(1-\eta\mu)\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} &= \sum_{i=1}^n (1-\eta\mu)\|x_i^k - x_i^\star\|_2^2 \geq \sum_{i=1}^n \|x_i^{k+1} - x^\star - \eta(z_i^k - z_i^\star)\|_2^2 \\
&= \|\mathbf{X}^{k+1}-\mathbf{X}^\star - \eta(\mathbf{Z}^k-\mathbf{Z}^\star)\|^2_{\mathbf{I}} \\
&= \|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}} + \eta^2\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{I}} - 2\eta\langle \mathbf{X}^{k+1}-\mathbf{X}^\star, \mathbf{Z}^k-\mathbf{Z}^\star\rangle.
\end{aligned}$$

After rearranging, we get

$$\begin{aligned}
-2\theta\langle \mathbf{X}^{k+1}-\mathbf{X}^\star, \mathbf{Z}^k-\mathbf{Z}^\star\rangle &\leq -\eta\theta\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{I}} - \frac{\theta\mu}{2}\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}} \\
&\quad + (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}},
\end{aligned}$$

which concludes the proof.

□

Theorem 14 (Convergence of Algorithm 1, Option B)**.** *Let Ψ_B^k be a Lyapunov function defined as follows:*

$$\Psi_B^k = \|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + \frac{(1-\eta\mu/2)\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} + \frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|^2_{\mathbf{I}}. \tag{39}$$

Let ρ_B be defined as follows:

$$\rho_B = \max\left\{2(\omega+1), \frac{2L(\lambda_{\max}(\mathbf{W}) + 12\omega \max_{(i,j)\in E} w_{ij})}{\mu\lambda_{\min}^{+}(\mathbf{W})}\right\}^{-1}. \tag{40}$$

Choosing stepsize θ

$$\theta = \frac{\mu}{2\lambda_{\max}(\mathbf{W}) + 24\omega \max_{(i,j)\in E} w_{ij}}, \tag{41}$$

stepsize $\eta = \frac{1}{L}$ and stepsize $\alpha = \frac{1}{\omega+1}$ gives the following inequality:

$$\mathbb{E}\left[\Psi_B^{k+1}\right] \le (1-\rho_B)\Psi_B^k. \tag{42}$$

Proof. We start with rewriting (16):

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{Z}^{k+1} - \mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger}\right] &\overset{(16)}{\le} \|\mathbf{Z}^k - \mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + \mathbb{E}\left[-2\theta\langle \mathbf{X}^{k+1} - \mathbf{X}^\star, \mathbf{Z}^k - \mathbf{Z}^\star\rangle + \theta^2\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{W}} + \Sigma^k\right] \\
&\overset{(38)}{\le} \|\mathbf{Z}^k - \mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} - \eta\theta\|\mathbf{Z}^k - \mathbf{Z}^\star\|^2_{\mathbf{I}} - \frac{\theta\mu}{2}\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{I}} + \theta^2\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{W}} + \mathbb{E}\left[\Sigma^k\right] \\
&+ (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k - \mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1 - \frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{I}} \\
&\overset{(13),(14)}{\le} (1-\eta\theta\lambda_{\min}^{+}(\mathbf{W}))\|\mathbf{Z}^k - \mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} - \theta\left(\frac{\mu}{2} - \theta\lambda_{\max}(\mathbf{W})\right)\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{I}} + \mathbb{E}\left[\Sigma^k\right] \\
&+ (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k - \mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1 - \frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{I}} \\
&\overset{(23)}{\le} (1-\eta\theta\lambda_{\min}^{+}(\mathbf{W}))\|\mathbf{Z}^k - \mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + 4\theta^2\omega \max_{(i,j)\in E} w_{ij}\|\mathbf{H}^k - \mathbf{X}_\star\|^2_{\mathbf{I}} \\
&- \theta\left(\frac{\mu}{2} - \theta\left[\lambda_{\max}(\mathbf{W}) + 4\omega\max_{(i,j)\in E} w_{ij}\right]\right)\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{I}} \\
&+ (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k - \mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1 - \frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{I}}.
\end{aligned}$$

Now, we combine this with (15):

$$\begin{aligned}
\mathbb{E}\left[\Psi_B^{k+1}\right] &\le (1-\eta\theta\lambda_{\min}^{+}(\mathbf{W}))\|\mathbf{Z}^k - \mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k - \mathbf{X}^\star\|^2_{\mathbf{I}} + 4\theta^2\omega\max_{(i,j)\in E} w_{ij}\|\mathbf{H}^k - \mathbf{X}_\star\|^2_{\mathbf{I}} \\
&- \theta\left(\frac{\mu}{2} - \theta\left[\lambda_{\max}(\mathbf{W}) + 4\omega\max_{(i,j)\in E} w_{ij}\right]\right)\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{I}} \\
&+ (1-\alpha)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k - \mathbf{X}^\star\|^2_{\mathbf{I}} + 8\theta^2\omega\max_{(i,j)\in E} w_{ij}\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{I}} \\
&= (1-\eta\theta\lambda_{\min}^{+}(\mathbf{W}))\|\mathbf{Z}^k - \mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k - \mathbf{X}^\star\|^2_{\mathbf{I}} \\
&+ \left(1 - \frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k - \mathbf{X}^\star\|^2_{\mathbf{I}} \\
&- \theta\left(\frac{\mu}{2} - \theta\left[\lambda_{\max}(\mathbf{W}) + 12\omega\max_{(i,j)\in E} w_{ij}\right]\right)\|\mathbf{X}^{k+1} - \mathbf{X}^\star\|^2_{\mathbf{I}}.
\end{aligned}$$

Using (41) we get

$$
\begin{aligned}
\mathbb{E}\left[\Psi_B^{k+1}\right] &\le (1-\eta\theta\lambda_{\min}^{+}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \left(1-\frac{\eta\mu}{2-\eta\mu}\right)\frac{(1-\eta\mu/2)\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 \\
&+ \left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 \\
&\le (1-\eta\theta\lambda_{\min}^{+}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \left(1-\frac{\eta\mu}{2}\right)\frac{(1-\eta\mu/2)\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 \\
&+ \left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 \\
&\le (1-\rho_B)\Psi_B^k,
\end{aligned}
$$

which concludes the proof. □

C.3 Option C

Lemma 15 (Primal step, Option C)**.** *Let $\eta \le \frac{1}{2L}$. Then the following inequality holds:*

$$
\begin{aligned}
-2\theta\mathbb{E}\left[\langle\mathbf{X}^{k+1}-\mathbf{X}^\star,\mathbf{Z}^k-\mathbf{Z}^\star\rangle\right] &\le -\eta\theta\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{I}}^2 - \frac{\theta\mu}{2}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2\right] + 2n\eta\theta\sigma^2 && (43)\\
&+ (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2\right]. && (44)
\end{aligned}
$$

Proof. From Line 8 of Algorithm 1 it follows that

$$
\begin{aligned}
\mathbb{E}\left[\|x_i^{k+1}-x^\star-\eta(z_i^k-z_i^\star)\|_2^2\right] &= \mathbb{E}\left[\|x_i^k-x^\star-\eta(\nabla f_i(x_i^k,\xi_i^k)-\nabla f_i(x^\star))\|_2^2\right] \\
&= \|x_i^k-x^\star\|_2^2 - 2\eta\mathbb{E}\left[\langle\nabla f_i(x_i^k,\xi_i^k)-\nabla f_i(x^\star), x_i^k-x^\star\rangle\right] \\
&+ \eta^2\mathbb{E}\left[\|\nabla f_i(x_i^k,\xi_i^k)-\nabla f_i(x^\star)\|_2^2\right] \\
&\overset{(5)}{\le} \|x_i^k-x^\star\|_2^2 - 2\eta\langle\nabla f_i(x_i^k)-\nabla f_i(x^\star), x_i^k-x^\star\rangle \\
&+ 2\eta^2\mathbb{E}\left[\|\nabla f_i(x_i^k,\xi_i^k)-\nabla f_i(x^\star,\xi_i^k)\|_2^2 + \|\nabla f_i(x^\star,\xi_i^k)-\nabla f_i(x^\star)\|_2^2\right] \\
&\overset{(3),(6),(7)}{\le} (1-\eta\mu)\|x_i^k-x^\star\|_2^2 - 2\eta B_{f_i}(x_i^k,x_i^\star) + 4L\eta^2 B_{f_i}(x_i^k,x^\star) + 2\eta^2\sigma_i^2 \\
&\le (1-\eta\mu)\|x_i^k-x_i^\star\|_2^2 + 2\eta^2\sigma_i^2,
\end{aligned}
$$

where we used $\eta \le \frac{1}{2L}$ in the last inequality. This leads to

$$
\begin{aligned}
(1-\eta\mu)\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 &= \sum_{i=1}^n(1-\eta\mu)\|x_i^k-x_i^\star\|_2^2 \ge \mathbb{E}\left[\sum_{i=1}^n\|x_i^{k+1}-x^\star-\eta(z_i^k-z_i^\star)\|_2^2\right] - 2n\eta^2\sigma^2 \\
&= \mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star-\eta(\mathbf{Z}^k-\mathbf{Z}^\star)\|_{\mathbf{I}}^2\right] - 2n\eta^2\sigma^2 \\
&= \eta^2\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{I}}^2 + \mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2 - 2\eta\langle\mathbf{X}^{k+1}-\mathbf{X}^\star,\mathbf{Z}^k-\mathbf{Z}^\star\rangle\right] - 2n\eta^2\sigma^2.
\end{aligned}
$$

After rearranging, we get

$$
\begin{aligned}
-2\theta\mathbb{E}\left[\langle\mathbf{X}^{k+1}-\mathbf{X}^\star,\mathbf{Z}^k-\mathbf{Z}^\star\rangle\right] &\le -\eta\theta\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{I}}^2 - \frac{\theta\mu}{2}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2\right] + 2n\eta\theta\sigma^2 \\
&+ (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2\right],
\end{aligned}
$$

which concludes the proof.

□

Theorem 16 (Convergence of Algorithm 1, Option C)**.** *Let Ψ_C^k be a Lyapunov function defined as follows:*

$$
\Psi_C^k = \|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \frac{(1-\eta\mu/2)\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 + \frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2. \tag{45}
$$

Let ρ_C be defined as follows:

$$\rho_C = \max\left\{2(\omega+1), \frac{2(\lambda_{\max}(\mathbf{W}) + 12\omega\max_{(i,j)\in E} w_{ij})}{\eta\mu\lambda^+_{\min}(\mathbf{W})}\right\}^{-1}. \tag{46}$$

Choosing stepsize θ

$$\theta = \frac{\mu}{2\lambda_{\max}(\mathbf{W}) + 24\omega\max_{(i,j)\in E} w_{ij}}, \tag{47}$$

stepsize $\eta \le \frac{1}{2L}$ and stepsize $\alpha = \frac{1}{\omega+1}$ gives the following inequality:

$$\mathbb{E}\left[\Psi_C^{k+1}\right] \le (1-\rho_C)\Psi_C^k + 2n\eta\theta\sigma^2. \tag{48}$$

Proof. We start with rewriting (16):

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{Z}^{k+1}-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger}\right] &\overset{(16)}{\le} \|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + \mathbb{E}\left[-2\theta\langle\mathbf{X}^{k+1}-\mathbf{X}^\star, \mathbf{Z}^k-\mathbf{Z}^\star\rangle + \theta^2\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{W}} + \Sigma^k\right]\\
&\overset{(44)}{\le} \|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} - \eta\theta\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{I}} - \frac{\theta\mu}{2}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right] + \theta^2\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{W}}\right]\\
&+ \mathbb{E}\left[\Sigma^k\right] + (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right] + 2n\eta\theta\sigma^2\\
&\overset{(13),(14)}{\le} (1-\eta\theta\lambda^+_{\min}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} - \theta\left(\frac{\mu}{2}-\theta\lambda_{\max}(\mathbf{W})\right)\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right] + \mathbb{E}\left[\Sigma^k\right]\\
&+ (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right] + 2n\eta\theta\sigma^2\\
&\overset{(23)}{\le} (1-\eta\theta\lambda^+_{\min}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + 4\theta^2\omega\max_{(i,j)\in E} w_{ij}\|\mathbf{H}^k-\mathbf{X}_\star\|^2_{\mathbf{I}}\\
&- \theta\left(\frac{\mu}{2}-\theta\left[\lambda_{\max}(\mathbf{W}) + 4\omega\max_{(i,j)\in E} w_{ij}\right]\right)\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right]\\
&+ (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right] + 2n\eta\theta\sigma^2
\end{aligned}$$

Now, we combine this with (15):

$$\begin{aligned}
\mathbb{E}\left[\Psi_C^{k+1}\right] &\le (1-\eta\theta\lambda^+_{\min}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} + 4\theta^2\omega\max_{(i,j)\in E} w_{ij}\|\mathbf{H}^k-\mathbf{X}_\star\|^2_{\mathbf{I}}\\
&- \theta\left(\frac{\mu}{2}-\theta\left[\lambda_{\max}(\mathbf{W}) + 4\omega\max_{(i,j)\in E} w_{ij}\right]\right)\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right]\\
&+ (1-\alpha)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} + 8\theta^2\omega\max_{(i,j)\in E} w_{ij}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right] + 2n\eta\theta\sigma^2\\
&= (1-\eta\theta\lambda^+_{\min}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} + 2n\eta\theta\sigma^2\\
&+ \left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|^2_{\mathbf{I}}\\
&- \theta\left(\frac{\mu}{2}-\theta\left[\lambda_{\max}(\mathbf{W}) + 12\omega\max_{(i,j)\in E} w_{ij}\right]\right)\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right].
\end{aligned}$$

Using (47) we get

$$\begin{aligned}\mathbb{E}\left[\Psi_C^{k+1}\right] &\leq (1-\eta\theta\lambda_{\min}^+(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \left(1-\frac{\eta\mu}{2-\eta\mu}\right)\frac{(1-\eta\mu/2)\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2\\
&+\left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 + 2n\eta\theta\sigma^2\\
&\leq (1-\eta\theta\lambda_{\min}^+(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \left(1-\frac{\eta\mu}{2}\right)\frac{(1-\eta\mu/2)\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2\\
&+\left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 + 2n\eta\theta\sigma^2\\
&\leq (1-\rho_C)\Psi_C^k + 2n\eta\theta\sigma^2,\end{aligned}$$

which concludes the proof. ◻

Corollary 17. *Let*

$$\hat{x}^k = \frac{1}{n}\sum_{i=1}^n x_i^k. \tag{49}$$

Then for any $\epsilon > 0$, choosing stepsize

$$\eta = \min\left\{\frac{1}{2L}, \frac{\sqrt{\epsilon}}{4\sigma\sqrt{\omega+1}}, \frac{\epsilon\mu\lambda_{\min}^+(\mathbf{W})}{16\sigma^2(\lambda_{\max}(\mathbf{W})+12\omega\max_{(i,j)\in E} w_{ij})}\right\} \tag{50}$$

and number of iterations

$$\begin{aligned}k \geq \max\Bigg\{&2(\omega+1), \frac{4L(\lambda_{\max}(\mathbf{W})+12\omega\max_{(i,j)\in E} w_{ij})}{\mu\lambda_{\min}^+(\mathbf{W})},\\
&\frac{8\sigma\sqrt{\omega+1}(\lambda_{\max}(\mathbf{W})+12\omega\max_{(i,j)\in E} w_{ij})}{\mu\lambda_{\min}^+(\mathbf{W})\sqrt{\epsilon}},\\
&\frac{32\sigma^2(\lambda_{\max}(\mathbf{W})+12\omega\max_{(i,j)\in E} w_{ij})^2}{\epsilon\mu^2(\lambda_{\min}^+(\mathbf{W}))^2}\Bigg\}\log\frac{4\eta\Psi_C^0}{n\theta\epsilon}\end{aligned} \tag{51}$$

gives

$$\mathbb{E}\left[\|\hat{x}^k - x^\star\|_2^2\right] \leq \epsilon. \tag{52}$$

Proof. Using definition of $\hat{x}^k$ and (48) we get

$$\begin{aligned}\mathbb{E}\left[\|\hat{x}^k-x^\star\|_2^2\right] &\leq \frac{1}{n}\mathbb{E}\left[\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2\right] \leq \mathbb{E}\left[\frac{\eta}{n\theta(1-\eta\mu/2)}\Psi_C^k\right] \leq \frac{2\eta}{n\theta}\mathbb{E}\left[\Psi_C^k\right]\\
&\leq \frac{2\eta}{n\theta}\left((1-\rho_C)^k\Psi_C^0 + \frac{2n\eta\theta\sigma^2}{\rho_C}\right) = (1-\rho_C)^k\frac{2\eta}{n\theta}\Psi_C^0 + \frac{4\eta^2\sigma^2}{\rho_C}.\end{aligned}$$

From (50) and (46) it follows that $\frac{4\eta^2\sigma^2}{\rho_C} \leq \frac{\epsilon}{2}$ and hence

$$\mathbb{E}\left[\|\hat{x}^k-x^\star\|_2^2\right] \leq (1-\rho_C)^k\frac{2\eta}{n\theta}\Psi_C^0 + \frac{\epsilon}{2}.$$

From (51), (50) and (46) it follows that $(1-\rho_C)^k \leq \frac{n\theta\epsilon}{4\eta\Psi_C^0}$ and hence

$$\mathbb{E}\left[\|\hat{x}^k-x^\star\|_2^2\right] \leq \frac{\epsilon}{2}+\frac{\epsilon}{2} = \epsilon.$$

◻

C.4 Option D

Lemma 18 (Primal step, Option D)**.** *Let $\eta \leq \frac{1}{6L}$. Then the following inequality holds:*

$$
\begin{aligned}
-2\theta\mathbb{E}\left[\langle \mathbf{X}^{k+1}-\mathbf{X}^\star, \mathbf{Z}^k-\mathbf{Z}^\star\rangle\right] \leq & -\eta\theta\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{I}}^2 - \frac{\theta\mu}{2}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2\right] \\
&+ (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2\right] \\
&+ 8mL\eta\theta\sum_{i=1}^n\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star) - \mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right].
\end{aligned} \tag{53}
$$

Proof. From Line 12 of Algorithm 1 it follows that

$$
\begin{aligned}
\mathbb{E}\left[\|x_i^{k+1}-x^\star-\eta(z_i^k-z_i^\star)\|_2^2\right] &= \mathbb{E}\left[\|x_i^k-x^\star-\eta(\nabla f_{ij_i^k}(x_i^k)-\nabla f_{ij_i^k}(w_i^k)+\nabla f_i(w_i^k)-\nabla f_i(x^\star))\|_2^2\right] \\
&= \|x_i^k-x^\star\|_2^2 - 2\eta\langle\nabla f_i(x_i^k)-\nabla f_i(x^\star), x_i^k-x^\star\rangle \\
&+ \eta^2\mathbb{E}\left[\|\nabla f_{ij_i^k}(x_i^k)-\nabla f_{ij_i^k}(w_i^k)+\nabla f_i(w_i^k)-\nabla f_i(x^\star)\|_2^2\right] \\
&\overset{(3)}{\leq} (1-\eta\mu)\|x_i^k-x^\star\|_2^2 - 2\eta B_{f_i}(x_i^k,x^\star) + 2\eta^2\mathbb{E}\left[\|\nabla f_{ij_i^k}(x_i^k)-\nabla f_{ij_i^k}(x^\star)\|_2^2\right] \\
&+ 2\eta^2\mathbb{E}\left[\|\nabla f_{ij_i^k}(x^\star)-\nabla f_i(x^\star)-\nabla f_{ij_i^k}(w_i^k)+\nabla f_i(w_i^k)\|_2^2\right] \\
&\leq (1-\eta\mu)\|x_i^k-x^\star\|_2^2 - 2\eta B_{f_i}(x_i^k,x^\star) + \frac{2\eta^2}{m}\sum_{j=1}^m\|\nabla f_{ij}(x_i^k)-\nabla f_{ij}(x^\star)\|_2^2 \\
&+ \frac{2\eta^2}{m}\sum_{j=1}^m\|\nabla f_{ij}(w_i^k)-\nabla f_{ij}(x^\star)\|_2^2 \\
&\overset{(8)}{\leq} (1-\eta\mu)\|x_i^k-x^\star\|_2^2 - 2\eta B_{f_i}(x_i^k,x^\star) + 4L\eta^2 B_{f_i}(x_i^k,x^\star) + 4L\eta^2 B_{f_i}(w_i^k,x^\star).
\end{aligned}
$$

From Line 11 of Algorithm 1 it follows that

$$
\mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right] = \left(1-\frac{1}{m}\right)B_{f_i}(w_i^k,x^\star) + \frac{1}{m}B_{f_i}(x_i^k,x_i^\star),
$$

which gives

$$
B_{f_i}(w_i^k,x_i^\star) \leq 2m\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star) - \mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right] + 2B_{f_i}(x_i^k,x_i^\star),
$$

and hence using stepsize $\eta \leq \frac{1}{6L}$ we get

$$
\begin{aligned}
\mathbb{E}\left[\|x_i^{k+1}-x^\star-\eta(z_i^k-z_i^\star)\|_2^2\right] &\leq (1-\eta\mu)\|x_i^k-x^\star\|_2^2 - 2\eta B_{f_i}(x_i^k,x^\star) + 12L\eta^2 B_{f_i}(x_i^k,x^\star) \\
&+ 8mL\eta^2\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star) - \mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right] \\
&\leq (1-\eta\mu)\|x_i^k-x^\star\|_2^2 + 8mL\eta^2\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star) - \mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right].
\end{aligned}
$$

This leads to

$$
\begin{aligned}
(1-\eta\mu)\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 = \sum_{i=1}^n(1-\eta\mu)\|x_i^k-x_i^\star\|_2^2 &\geq \sum_{i=1}^n \mathbb{E}\left[\|x_i^{k+1}-x^\star-\eta(z_i^k-z_i^\star)\|_2^2\right] \\
&- 8mL\eta^2\sum_{i=1}^n\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star)-\mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right] \\
&= \mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star-\eta(\mathbf{Z}^k-\mathbf{Z}^\star)\|_{\mathbf{I}}^2\right] \\
&- 8mL\eta^2\sum_{i=1}^n\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star)-\mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right] \\
&= \mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2\right]+\eta^2\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{I}}^2-2\eta\mathbb{E}\left[\langle\mathbf{X}^{k+1}-\mathbf{X}^\star,\mathbf{Z}^k-\mathbf{Z}^\star\rangle\right] \\
&- 8mL\eta^2\sum_{i=1}^n\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star)-\mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right].
\end{aligned}
$$

After rearranging, we get

$$
\begin{aligned}
-2\theta\mathbb{E}\left[\langle\mathbf{X}^{k+1}-\mathbf{X}^\star,\mathbf{Z}^k-\mathbf{Z}^\star\rangle\right] &\leq -\eta\theta\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{I}}^2-\frac{\theta\mu}{2}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2\right] \\
&+ (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2-\left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|_{\mathbf{I}}^2\right] \\
&+ 8mL\eta\theta\sum_{i=1}^n\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star)-\mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right],
\end{aligned}
$$

which concludes the proof.

□

Theorem 19 (Convergence of Algorithm 1, Option D)**.** *Let Ψ_D^k be a Lyapunov function which is defined as follows:*

$$
\Psi_D^k = \|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2+\frac{(1-\eta\mu/2)\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2+\frac{8\theta^2\omega\max_{(i,j)\in E}w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2+8mL\eta\theta\sum_{i=1}^n B_{f_i}(w_i^k,x^\star). \tag{54}
$$

Let ρ_D be defined as follows:

$$
\rho_D = \max\left\{2m, 2(\omega+1), \frac{12L(\lambda_{\max}(\mathbf{W})+12\omega\max_{(i,j)\in E}w_{ij})}{\mu\lambda_{\min}^+(\mathbf{W})}\right\}^{-1}. \tag{55}
$$

Choosing stepsize θ

$$
\theta = \frac{\mu}{2\lambda_{\max}(\mathbf{W})+24\omega\max_{(i,j)\in E}w_{ij}}, \tag{56}
$$

stepsize $\eta = \frac{1}{6L}$ and stepsize $\alpha = \frac{1}{\omega+1}$ gives the following inequality:

$$
\mathbb{E}\left[\Psi_D^{k+1}\right] \leq (1-\rho_D)\Psi_D^k. \tag{57}
$$

Proof. We start with rewriting (16):

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{Z}^{k+1}-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger}\right] &\overset{(16)}{\leq} && \|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + \mathbb{E}\left[-2\theta\langle\mathbf{X}^{k+1}-\mathbf{X}^\star,\mathbf{Z}^k-\mathbf{Z}^\star\rangle + \theta^2\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{W}} + \Sigma^k\right]\\
&\overset{(53)}{\leq} && \|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} - \eta\theta\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{I}} - \mathbb{E}\left[\frac{\theta\mu}{2}\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}} + \theta^2\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{W}} + \Sigma^k\right]\\
&+ && (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right]\\
&+ && 8mL\eta\theta\sum_{i=1}^n\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star) - \mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right]\\
&\overset{(13),(14)}{\leq} && (1-\eta\theta\lambda^+_{\min}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} - \theta\left(\frac{\mu}{2}-\theta\lambda_{\max}(\mathbf{W})\right)\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right] + \mathbb{E}\left[\Sigma^k\right]\\
&+ && (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right]\\
&+ && 8mL\eta\theta\sum_{i=1}^n\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star) - \mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right]\\
&\overset{(23)}{\leq} && (1-\eta\theta\lambda^+_{\min}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + 4\theta^2\omega\max_{(i,j)\in E} w_{ij}\|\mathbf{H}^k-\mathbf{X}_\star\|^2_{\mathbf{I}}\\
&- && \theta\left(\frac{\mu}{2}-\theta\left[\lambda_{\max}(\mathbf{W})+4\omega\max_{(i,j)\in E} w_{ij}\right]\right)\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right]\\
&+ && (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} - \left(1-\frac{\eta\mu}{2}\right)\frac{\theta}{\eta}\mathbb{E}\left[\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\right]\\
&+ && 8mL\eta\theta\sum_{i=1}^n\left[\left(1-\frac{1}{2m}\right)B_{f_i}(w_i^k,x^\star) - \mathbb{E}\left[B_{f_i}(w_i^{k+1},x^\star)\right]\right].
\end{aligned}
$$

Now, we combine this with (15):

$$
\begin{aligned}
\mathbb{E}\left[\Psi_D^{k+1}\right] \leq{}& (1-\eta\theta\lambda^+_{\min}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} + \left(1-\frac{1}{2m}\right)8mL\eta\theta\sum_{i=1}^n B_{f_i}(w_i^k,x^\star)\\
&+ 4\theta^2\omega\max_{(i,j)\in E} w_{ij}\|\mathbf{H}^k-\mathbf{X}_\star\|^2_{\mathbf{I}} - \theta\left(\frac{\mu}{2}-\theta\left[\lambda_{\max}(\mathbf{W})+4\omega\max_{(i,j)\in E} w_{ij}\right]\right)\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\\
&+ (1-\alpha)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|^2_{\mathbf{I}} + 8\theta^2\omega\max_{(i,j)\in E} w_{ij}\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}\\
&= (1-\eta\theta\lambda^+_{\min}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|^2_{\mathbf{W}^\dagger} + (1-\eta\mu)\frac{\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|^2_{\mathbf{I}}\\
&+ \left(1-\frac{1}{2m}\right)8mL\eta\theta\sum_{i=1}^n B_{f_i}(w_i^k,x^\star) + \left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|^2_{\mathbf{I}}\\
&- \theta\left(\frac{\mu}{2}-\theta\left[\lambda_{\max}(\mathbf{W})+12\omega\max_{(i,j)\in E} w_{ij}\right]\right)\|\mathbf{X}^{k+1}-\mathbf{X}^\star\|^2_{\mathbf{I}}.
\end{aligned}
$$

Using (56) we get

$$
\begin{aligned}
\mathbb{E}\left[\Psi_D^{k+1}\right] &\leq (1-\eta\theta\lambda_{\min}^{+}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \left(1-\frac{\eta\mu}{2-\eta\mu}\right)\frac{(1-\eta\mu/2)\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 \\
&+ \left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 + \left(1-\frac{1}{2m}\right)8mL\eta\theta\sum_{i=1}^n B_{f_i}(w_i^k,x^\star) \\
&\leq (1-\eta\theta\lambda_{\min}^{+}(\mathbf{W}))\|\mathbf{Z}^k-\mathbf{Z}^\star\|_{\mathbf{W}^\dagger}^2 + \left(1-\frac{\eta\mu}{2}\right)\frac{(1-\eta\mu/2)\theta}{\eta}\|\mathbf{X}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 \\
&+ \left(1-\frac{\alpha}{2}\right)\frac{8\theta^2\omega\max_{(i,j)\in E} w_{ij}}{\alpha}\|\mathbf{H}^k-\mathbf{X}^\star\|_{\mathbf{I}}^2 + \left(1-\frac{1}{2m}\right)8mL\eta\theta\sum_{i=1}^n B_{f_i}(w_i^k,x^\star) \\
&\leq (1-\rho_D)\Psi_D^k,
\end{aligned}
$$

which concludes the proof. □