Towards a Better Theoretical Understanding of Independent Subnetwork Training

Egor Shulgin [1] **Peter Richtárik** [1]

Abstract

Modern advancements in large-scale machine learning would be impossible without the paradigm of data-parallel distributed computing. Since distributed computing with large-scale models imparts excessive pressure on communication channels, significant recent research has been directed toward co-designing communication compression strategies and training algorithms with the goal of reducing communication costs. While pure data parallelism allows better data scaling, it suffers from poor model scaling properties. Indeed, compute nodes are severely limited by memory constraints, preventing further increases in model size. For this reason, the latest achievements in training giant neural network models also rely on some form of model parallelism. In this work, we take a closer theoretical look at Independent Subnetwork Training (IST), which is a recently proposed and highly effective technique for solving the aforementioned problems. We identify fundamental differences between IST and alternative approaches, such as distributed methods with compressed communication, and provide a precise analysis of its optimization performance on a quadratic model.

1. Introduction

A huge part of today's machine learning success is driven by the possibility of building more and more complex models and training them on increasingly larger datasets. This rapid progress has become feasible due to advancements in distributed optimization, which is necessary for proper scaling when the size of the training data grows (Zinkevich et al., 2010). In a typical scenario, data parallelism is used for efficiency and implies sharding the dataset across computing devices. This allowed very efficient scaling and acceleration of training moderately sized models by using additional hardware (Goyal et al., 2018). However, this data parallel approach can suffer from communication bottleneck, which has sparked extensive research on distributed optimization with compressed communication of the parameters between nodes (Alistarh et al., 2017; Konečný et al., 2016; Seide et al., 2014).

1.1. The Need for Model Parallelism

Despite its efficiency, data parallelism has some fundamental limitations when it comes to scaling up the model size. As the dimensions of a model increase, the amount of memory required to store and update the parameters also increases, which becomes problematic due to resource constraints on individual devices. This has led to the development of model parallelism (Dean et al., 2012; Richtárik & Takáč, 2016), which splits a large model across multiple nodes, with each node responsible for computations of parts of the model (Farber & Asanovic, 1997; Zhang et al., 1989). However, naive model parallelism also poses challenges because each node can only update its portion of the model based on the data it has access to. This creates a need for very careful management of communication between devices. Thus, a combination of both data and model parallelism is often necessary to achieve efficient and scalable training of huge models.

IST. Independent Subnetwork Training (IST) is a technique that suggests dividing a neural network into smaller subparts, training them in a distributed parallel fashion, and then aggregating the results to update the weights of the whole model. In IST, every subnetwork can operate independently and has fewer parameters than the full model, which not only reduces the load on computing nodes but also results in faster synchronization. A generalized analog of the described method is formalized as an iterative procedure in Algorithm 1 and schematically depicted in Figure 3. IST paradigm was pioneered by Yuan et al. (2022) for networks with fully connected layers and was later extended to ResNets (Dun et al., 2022) and Graph architectures (Wolfe et al., 2023). Previous experimental studies have shown

[1]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. Correspondence to: Egor Shulgin <egor.shulgin@kaust.edu.sa>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Algorithm 1 Distributed Submodel (Stochastic) Gradient Descent

Parameters: learning rate $\gamma > 0$; sketches $\mathbf{C}_1, \dots, \mathbf{C}_n$; initial model $x^0 \in \mathbb{R}^d$
for $k = 0, 1, 2 \dots$ **do**
  Select submodels $w_i^k = \mathbf{C}_i^k x^k$ for $i \in [n]$ and broadcast to all computing nodes
  for $i = 1, \dots, n$ in parallel **do**
    Compute local (stochastic) gradient w.r.t. submodel: $\mathbf{C}_i^k \nabla f_i(w_i^k)$
    Take (maybe multiple) gradient descent step $z_i^+ = w_i^k - \gamma \mathbf{C}_i^k \nabla f_i(w_i^k)$
    Send z_i^+ to the server
  end for
  Aggregate/merge received submodels: $x^{k+1} = \frac{1}{n} \sum_{i=1}^n z_i^+$
end for

that IST is a very promising approach for various applications as it allows to effectively combine data and model parallelism and train larger models with limited compute. In addition, Liao & Kyrillidis (2022) performed theoretical analysis of IST for overparameterized single hidden layer neural networks with ReLU activations. The idea of IST was also recently extended to the federated setting via an asynchronous distributed dropout technique (Dun et al., 2023).

Federated Learning. Another important setting when the data is distributed (due to privacy reasons) is Federated Learning (Kairouz et al., 2021; Konečný et al., 2016; McMahan et al., 2017). In this scenario, computing devices are often heterogeneous and more resource-constrained (Caldas et al., 2018) (e.g. mobile phones) in comparison to datacenter settings. Such challenges have prompted extensive research efforts into selecting smaller and more efficient submodels for local on-device training (Alam et al., 2022; Charles et al., 2022; Chen et al., 2023; Diao et al., 2021; Horvath et al., 2021; Jiang et al., 2022; Lin et al., 2022; Qiu et al., 2022; Wen et al., 2022; Yang et al., 2022). Many of these works propose approaches to adapt submodels, often tailored to specific neural network architectures, based on the capabilities of individual clients for various machine learning tasks. However, there is a lack of comprehension regarding the theoretical properties of these methods.

1.2. Summary of Contributions

After reviewing the literature, we identified a glaring gap in the rigorous understanding of IST convergence, directly motivating our research. The main contributions of this paper include:

- A novel approach to analyzing distributed methods that combine data and model parallelism by operating with sparse submodels for a quadratic model.
- The first analysis of independent subnetwork training in homogeneous and heterogeneous scenarios without restrictive assumptions on gradient estimators.
- Identification of the settings when IST can optimize very efficiently or not converge to the optimal solution but only to an irreducible neighborhood that is also tightly characterized.
- Empirical validation of the proposed theory through experiments for several practical settings.

2. Formalism and Setup

We consider the standard optimization formulation of a distributed/federated learning problem (Wang et al., 2021)

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x) \right], \tag{1}$$

where n is the number of clients/workers, and each $f_i : \mathbb{R}^d \to \mathbb{R}$ represents the loss of the model parameterized by vector $x \in \mathbb{R}^d$ on the data of client i.

A typical Stochastic Gradient Descent (SGD)-type method for solving this problem has the form

$$x^{k+1} = x^k - \gamma g^k, \qquad g^k = \frac{1}{n} \sum_{i=1}^n g_i^k, \tag{2}$$

where $\gamma > 0$ is the stepsize and g_i^k is a suitably constructed estimator of $\nabla f_i(x^k)$. In the distributed setting, computation of gradient estimators g_i^k is typically performed by clients, and the results are sent to the server, which subsequently performs aggregation via averaging $g^k = \frac{1}{n} \sum_{i=1}^n g_i^k$. The average is then used to update the model x^{k+1} via a gradient-type method (2), and at the next iteration, the model is broadcasted back to the clients. The process is repeated iteratively until a suitable model is found.

One of the main techniques used to accelerate distributed training is lossy *communication compression* (Alistarh et al., 2017; Konečný et al., 2016; Seide et al., 2014), which suggests applying a (possibly randomized) lossy compression mapping $\mathcal{C}$ to a vector/matrix/tensor x before broadcasting. This reduces the bits sent per communication round

at the cost of transmitting a less accurate estimate $\mathcal{C}(x)$ of x. Described technique can be formalized in the following definition.

Definition 2.1 (Unbiased compressor)**.** A randomized mapping $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is an **unbiased compression operator** ($\mathcal{C} \in \mathbb{U}(\omega)$ for brevity) if for some $\omega \geq 0$ and $\forall x \in \mathbb{R}^d$

$$\mathbb{E}\left[\mathcal{C}(x)\right] = x, \qquad \mathbb{E}\left[\|\mathcal{C}(x) - x\|^2\right] \leq \omega \|x\|^2. \tag{3}$$

A notable example of a mapping from this class is the *random sparsification* (`Rand-q` for $q \in \{1, \ldots, d\}$) operator defined by

$$\mathcal{C}_{\texttt{Rand-q}}(x) \coloneqq \mathbf{C}_q x = \frac{d}{q} \sum_{i \in S} e_i e_i^\top x, \tag{4}$$

where $e_1, \ldots, e_d \in \mathbb{R}^d$ are standard unit basis vectors, and S is a random subset of $[d] \coloneqq \{1, \ldots, d\}$ sampled from the uniform distribution on the all subsets of $[d]$ with cardinality q. `Rand-q` belongs to $\mathbb{U}\left(d/q - 1\right)$, which means that the more elements are "dropped" (lower q), the higher the variance ω of the compressor.

In this work, we are mainly interested in a somewhat more general class of operators than mere sparsifiers. In particular, we are interested in compressing via the application of random matrices, i.e., via *sketching*. A sketch $\mathbf{C}_i^k \in \mathbb{R}^{d \times d}$ can be used to represent submodel computations in the following way:

$$g_i^k \coloneqq \mathbf{C}_i^k \nabla f_i(\mathbf{C}_i^k x^k),$$

where we require $\mathbf{C}_i^k$ to be a symmetric positive semi-definite matrix. Such gradient estimates correspond to computing the local gradient with respect to a sparse submodel $\mathbf{C}_i^k x^k$, and additionally sketching the resulting gradient with the same matrix $\mathbf{C}_i^k$ to guarantee that the resulting update lies in the lower-dimensional subspace.

Using this notion, Algorithm 1 (with one local gradient step) can be represented in the following form:

$$x^{k+1} = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbf{C}_i^k x^k - \gamma \mathbf{C}_i^k \nabla f_i(\mathbf{C}_i^k x^k)\right], \tag{5}$$

which is equivalent to the SGD-type update (2) when the *perfect reconstruction* property holds (with probability one)

$$\mathbf{C}^k \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbf{C}_i^k = \mathbf{I},$$

where $\mathbf{I}$ is the identity matrix. This property is inherent for a specific class of compressors that are particularly useful for capturing the concept of an *independent* subnetwork partition.

Definition 2.2 (Permutation sketch)**.** Assume that model size is greater than the number of clients $d \geq n$ and $d = qn$, where $q \geq 1$ is an integer[1]. Let $\pi = (\pi_1, \ldots, \pi_d)$ be a random permutation of $[d]$. Then for all $i \in [n]$, we define `Perm-q` operator

$$\mathbf{C}_i \coloneqq n \cdot \sum_{j=q(i-1)+1}^{qi} e_{\pi_j} e_{\pi_j}^\top. \tag{6}$$

`Perm-q` is unbiased and can be conveniently used for representing a structured decomposition of the model, such that every client i is responsible for computations over a submodel $\mathbf{C}_i x^k$.

Our convergence analysis relies on the assumption that was previously used for coordinate descent-type methods.

Assumption 2.3 (Matrix smoothness)**.** A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is $\mathbf{L}$-smooth, if there exists a positive semi-definite matrix $\mathbf{L} \in \mathbb{R}^{d \times d}$ such that $\forall x, h \in \mathbb{R}^d$

$$f(x + h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \langle \mathbf{L} h, h \rangle. \tag{7}$$

A standard L-smoothness condition is obtained as a special case of (7) for $\mathbf{L} = L \cdot \mathbf{I}$. Matrix smoothness was previously used for designing data-dependent gradient sparsification to accelerate optimization in communication-constrained settings (Safaryan et al., 2021; Wang et al., 2022).

2.1. Issues with Existing Approaches

Consider the simplest gradient descent method with a compressed model in the single-node setting:

$$x^{k+1} = x^k - \gamma \nabla f(\mathcal{C}(x^k)). \tag{8}$$

Algorithms belonging to this family require a different analysis in comparison to SGD (Gorbunov et al., 2020; Gower et al., 2019), Distributed Compressed Gradient Descent (Alistarh et al., 2017; Khirirat et al., 2018), and Randomized Coordinate Descent (Nesterov, 2012; Richtárik & Takáč, 2014)-type methods because the gradient estimator is no longer unbiased

$$\mathbb{E}\left[\nabla f(\mathcal{C}(x))\right] \neq \nabla f(x) = \mathbb{E}\left[\mathcal{C}(\nabla f(x))\right].$$

This is why such kind of algorithms (8) are harder to analyze. So, prior results for *unbiased* SGD (Khaled & Richtárik, 2023) cannot be directly reused. Furthermore, the nature of

[1]While this condition may look restrictive, it naturally holds for distributed learning in a data-center setting. Permutation sparsifiers were introduced by (Szlendak et al., 2022) and generalized to other scenarios (like $n \geq d$).

the bias in this type of gradient estimator does not exhibit additive noise, thereby preventing the application of previous analyses for biased SGD (Ajalloeian & Stich, 2020).

An assumption like the bounded stochastic gradient norm extensively used in previous works (Lin et al., 2019; Zhou et al., 2022) hinders an accurate understanding of such methods. This assumption hides the fundamental difficulty of analyzing a biased gradient estimator:

$$\mathbb{E}\left[\|\nabla f(\mathcal{C}(x))\|^2\right] \leq G \tag{9}$$

and may not hold, even for quadratic functions $f(x) = x^\top \mathbf{A} x$. In addition, in the distributed setting, such a condition can result in vacuous bounds (Khaled et al., 2020) as it does not capture heterogeneity accurately.

2.2. Simplifications Taken

To conduct a thorough theoretical analysis of methods that combine data with model parallelism, we simplify the algorithm and problem setting to isolate the unique effects of this approach. The following considerations are made:

- **(a)** We assume that every node i computes the true gradient at the submodel $\mathbf{C}_i \nabla f_i(\mathbf{C}_i x^k)$.
- **(b)** A notable difference compared to the original IST Algorithm 1 is that workers perform a single gradient descent step (or just gradient computation).
- **(c)** Finally, we consider a special case of a quadratic model (10) as a loss function (1).

Condition **(a)** is mainly for the sake of simplicity and clarity of exposition and can be generalized to stochastic gradient estimators with bounded variance. Condition **(b)** is imposed because local steps did not bring any theoretical efficiency improvements for heterogeneous settings until very recently (Mishchenko et al., 2022), and even then, only with the introduction of additional control variables, which goes against the requirements of resource-constrained device settings. The reason behind **(c)** is that despite its apparent simplicity, the quadratic problem has been used extensively to study properties of neural networks (Zhang et al., 2019; Zhu et al., 2023). Moreover, it is a non-trivial model, which makes it possible to understand complex optimization algorithms (Arjevani et al., 2020; Cunha et al., 2022; Goujaud et al., 2022). The quadratic problem is suitable for observing complex phenomena and providing theoretical insights, which can also be observed in practical scenarios. Finally, Appendix C presents a generalization for smooth functions.

Having said that, we consider a special case of problem (1) for symmetric matrices $\mathbf{L}_i$

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \quad f_i(x) \equiv \frac{1}{2} x^\top \mathbf{L}_i x - x^\top \mathrm{b}_i\,. \tag{10}$$

In this case, $f(x)$ is $\overline{\mathbf{L}}$-smooth, and $\nabla f(x) = \overline{\mathbf{L}} x - \overline{\mathrm{b}}$, where $\overline{\mathbf{L}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{L}_i$ and $\overline{\mathrm{b}} := \frac{1}{n}\sum_{i=1}^{n} \mathrm{b}_i$.

3. Results in the Interpolation Case

First, let us examine the case of $\mathrm{b}_i \equiv 0$, which we call interpolation for quadratics, and perform the analysis for general sketches $\mathbf{C}_i^k$. In this case, the gradient estimator (2) takes the form

$$g^k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{C}_i^k \nabla f_i(\mathbf{C}_i^k x^k) = \overline{\mathbf{B}}^k x^k, \tag{11}$$

where $\overline{\mathbf{B}}^k := \frac{1}{n}\sum_{i=1}^{n} \mathbf{C}_i^k \mathbf{L}_i \mathbf{C}_i^k$. We prove the following result for a method with such an estimator.

Theorem 3.1. *Consider the method (2) with estimator (11) for a quadratic problem (10) with $\overline{\mathbf{L}} \succ 0$ and $\mathrm{b}_i \equiv 0$. Then if $\overline{\mathbf{W}} := \frac{1}{2}\mathbb{E}\left[\overline{\mathbf{L}}\,\overline{\mathbf{B}}^k + \overline{\mathbf{B}}^k\,\overline{\mathbf{L}}\right] \succeq 0$ and there exists a constant $\theta > 0$:*

$$\mathbb{E}\left[\overline{\mathbf{B}}^k\,\overline{\mathbf{L}}\,\overline{\mathbf{B}}^k\right] \preceq \theta\,\overline{\mathbf{W}}, \tag{12}$$

and the step size is chosen as $0 < \gamma \leq \frac{1}{\theta}$, the iterates satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla f(x^k)\|^2_{\overline{\mathbf{L}}^{-1}\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-1}}\right] \leq \frac{2\delta^K}{\gamma K}, \tag{13}$$

for $\delta^K := f(x^0) - \mathbb{E}\left[f(x^K)\right]$ and

$$\mathbb{E}\left[\|x^k - x^\star\|^2_{\overline{\mathbf{L}}}\right] \leq (1-\gamma\Lambda)^k \|x^0 - x^\star\|^2_{\overline{\mathbf{L}}}, \tag{14}$$

where $\Lambda := \lambda_{\min}\left(\overline{\mathbf{L}}^{-\frac{1}{2}}\,\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-\frac{1}{2}}\right)$.

This theorem establishes an $\mathcal{O}(1/K)$ convergence rate with a constant step size up to a stationary point (13) and linear convergence (14) for the expected distance to the optimum $x^\star := \arg\min f(x)$. Note that we employ weighted norms in our analysis, as the considered class of loss functions satisfies the matrix $\overline{\mathbf{L}}$-smoothness Assumption 2.3. The use of standard Euclidean distance may result in loose bounds that do not recover correct rates for special cases like gradient descent.

It is important to highlight that the inequality (12) may not hold (for any $\theta > 0$) in the general case as the matrix $\overline{\mathbf{W}}$ is not guaranteed to be positive (semi-)definite in the case of general sampling. The intuition behind this issue is that arbitrary sketches $\mathbf{C}_i^k$ can result in the gradient estimator g^k,

which is misaligned with the true gradient $\nabla f(x^k)$. Specifically, the inner product $\langle \nabla f(x^k), g^k \rangle$ can be negative, and there is no expected descent after one step.

Next, we give examples of samplings for which the inequality (12) can be satisfied.

1. Identity. Consider $\mathbf{C}_i \equiv \mathbf{I}$. Then $\overline{\mathbf{B}}^k = \overline{\mathbf{L}}$, $\overline{\mathbf{B}}^k \overline{\mathbf{L}} \overline{\mathbf{B}}^k = \overline{\mathbf{L}}^3$, $\overline{\mathbf{W}} = \overline{\mathbf{L}}^2 \succ 0$ and hence (12) is satisfied for $\theta = \lambda_{\max}(\overline{\mathbf{L}})$. So, (13) says that if we choose $\gamma = 1/\theta$, then

$$\frac{1}{K}\sum_{k=0}^{K-1} \left\|\nabla f(x^k)\right\|_{\mathbf{I}}^2 \leq \frac{2\lambda_{\max}(\overline{\mathbf{L}})\left(f(x^0) - f(x^K)\right)}{K},$$

which exactly matches the rate of gradient descent in the non-convex setting. As for convergence of the iterates, the rate in (14) is $\lambda_{\max}(\overline{\mathbf{L}})/\lambda_{\min}(\overline{\mathbf{L}})$ which corresponds to the precise gradient descent result for strongly convex functions.

2. Permutation. Assume[2] $n = d$ and the use of `Perm-1` (special case of Definition 2.2) sketch $\mathbf{C}_i^k = n e_{\pi_i^k} e_{\pi_i^k}^\top$, where $\pi^k = (\pi_1^k, \ldots, \pi_n^k)$ is a random permutation of $[n]$. Then

$$\mathbb{E}\left[\overline{\mathbf{B}}^k\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\mathbf{C}_i^k \mathbf{L}_i \mathbf{C}_i^k\right] = \frac{1}{n}\sum_{i=1}^n n\mathrm{Diag}(\mathbf{L}_i) = n\,\overline{\mathbf{D}}$$

where $\overline{\mathbf{D}} := \frac{1}{n}\sum_{i=1}^n \mathbf{D}_i, \mathbf{D}_i := \mathrm{Diag}(\mathbf{L}_i)$. Then inequality (12) leads to

$$n\,\overline{\mathbf{D}}\,\overline{\mathbf{L}}\,\overline{\mathbf{D}} \preceq \frac{\theta}{2}\left(\overline{\mathbf{L}}\,\overline{\mathbf{D}} + \overline{\mathbf{D}}\,\overline{\mathbf{L}}\right),$$

which may not always hold as $\overline{\mathbf{L}}\,\overline{\mathbf{D}} + \overline{\mathbf{D}}\,\overline{\mathbf{L}}$ is not guaranteed to be positive-definite—even in the case of $\overline{\mathbf{L}} \succ 0$. However, such a condition can be enforced via a slight modification of the permutation sketches, which is done in Section 3.2. The limitation of such an approach is that the resulting compressors are no longer unbiased.

Remark 3.2. Matrix $\overline{\mathbf{W}}$ in the case of permutation sketches may not be positive-definite. Consider the following example of a homogeneous ($\mathbf{L}_i \equiv \mathbf{L}$) two-dimensional problem:

$$\mathbf{L} = \begin{bmatrix} a & c \\ c & b \end{bmatrix}.$$

Then $\overline{\mathbf{W}}$ equals to

$$\frac{1}{2}\left[\overline{\mathbf{L}}\,\overline{\mathbf{D}} + \overline{\mathbf{D}}\,\overline{\mathbf{L}}\right] = \begin{bmatrix} a^2 & c(a+b)/2 \\ c(a+b)/2 & b^2 \end{bmatrix},$$

which for $c > \frac{2ab}{a+b}$ has $\det(\overline{\mathbf{W}}) < 0$, and thus $\overline{\mathbf{W}} \not\succ 0$ according to Sylvester's criterion.

[2]This is mainly done to simplify the presentation. Results can be generalized to the case of $n \neq d$ in a similar manner as in (Szlendak et al., 2022), which can be found in the Appendix.

Next, we focus on the particular case of **permutation** sketches, which are the most suitable for model partitioning according to Independent Subnetwork Training (IST). In the rest of this section, we discuss how the condition (12) can be enforced via a specially designed preconditioning of the problem (10) or modification of the sketch mechanism (6).

3.1. Homogeneous Problem Preconditioning

To start, consider a homogeneous setting $f_i(x) = \frac{1}{2}x^\top \mathbf{L} x$, so $\mathbf{L}_i \equiv \mathbf{L}$. Now define $\mathbf{D} = \mathrm{Diag}(\mathbf{L})$ – a diagonal matrix with elements equal to the diagonal of $\mathbf{L}$. Then, the problem can be converted to

$$f_i(\mathbf{D}^{-\frac{1}{2}}x) = \frac{1}{2}\left(\mathbf{D}^{-\frac{1}{2}}x\right)^\top \mathbf{L}\left(\mathbf{D}^{-\frac{1}{2}}x\right) = \frac{1}{2}x^\top \tilde{\mathbf{L}}\, x,$$

where $\tilde{\mathbf{L}} := \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$. It is equivalent to the original problem after changing the variables $\tilde{x} := \mathbf{D}^{-\frac{1}{2}}x$. Note that $\mathbf{D} = \mathrm{Diag}(\mathbf{L})$ is positive-definite as $\mathbf{L} \succ 0$, and therefore $\tilde{\mathbf{L}} \succ 0$. Moreover, the preconditioned matrix $\tilde{\mathbf{L}}$ has all ones on the diagonal: $\mathrm{Diag}(\tilde{\mathbf{L}}) = \mathbf{I}$. If we now combine reformulated problem with `Perm-1` sketches

$$\mathbb{E}\left[\overline{\mathbf{B}}^k\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{C}_i\,\tilde{\mathbf{L}}\,\mathbf{C}_i\right] = n\mathrm{Diag}(\tilde{\mathbf{L}}) = n\mathbf{I}.$$

Therefore, inequality (12) takes the form $\tilde{\mathbf{W}} = n\tilde{\mathbf{L}} \succeq \frac{1}{\theta}n^2\tilde{\mathbf{L}}$, which holds for $\theta \geq n$, and the left-hand side of (13) can be transformed (for an accurate comparison to standard methods) in the following way:

$$\begin{aligned}\left\|\nabla f(x^k)\right\|_{\tilde{\mathbf{L}}^{-1}\tilde{\mathbf{W}}\tilde{\mathbf{L}}^{-1}}^2 &\geq n\lambda_{\min}\left(\tilde{\mathbf{L}}^{-1}\right)\left\|\nabla f(x^k)\right\|_{\mathbf{I}}^2 \\ &= n\lambda_{\max}(\tilde{\mathbf{L}})\left\|\nabla f(x^k)\right\|_{\mathbf{I}}^2\end{aligned}$$

The resulting convergence guarantee is

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|_{\mathbf{I}}^2\right] \leq \frac{2\lambda_{\max}(\tilde{\mathbf{L}})\delta^K}{K},$$

which matches classical gradient descent.

3.2. Heterogeneous Sketch Preconditioning

In contrast to the homogeneous case, the heterogeneous problem $f_i(x) = \frac{1}{2}x^\top \mathbf{L}_i x$ cannot be so easily preconditioned by a simple change of variables $\tilde{x} := \mathbf{D}^{-\frac{1}{2}}x$, as every client i has its own matrix $\mathbf{L}_i$. However, this problem can be fixed via the following modification of `Perm-1`, which scales the output according to the diagonal elements of the local smoothness matrix $\mathbf{L}_i$:

$$\tilde{\mathbf{C}}_i := \sqrt{n/\left[\mathbf{L}_i\right]_{\pi_i,\pi_i}}\, e_{\pi_i} e_{\pi_i}^\top. \tag{15}$$

In this case, $\mathbb{E}\left[\tilde{\mathbf{C}}_i \mathbf{L}_i \tilde{\mathbf{C}}_i\right] = \mathbf{I}$, $\mathbb{E}\left[\overline{\mathbf{B}}^k\right] = \mathbf{I}$, and $\overline{\mathbf{W}} = \overline{\mathbf{L}}$. Then inequality (12) is satisfied for $\theta \geq 1$.

If one inputs these results into (13), such convergence guarantee can be obtained

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|_{\mathbf{I}}^2\right] \leq \frac{2\lambda_{\max}(\overline{\mathbf{L}})\delta^K}{K},$$

which matches the gradient descent result as well. Thus, we can conclude that heterogeneity does not bring such a fundamental challenge in this scenario. In addition, the method with `Perm-1` is significantly better in terms of computational and communication complexity, as it requires calculation of the local gradients with respect to much smaller submodels and transmits only sparse updates.

This construction also shows that for $\gamma = 1/\theta = 1$

$$\gamma\lambda_{\min}\left(\overline{\mathbf{L}}^{-\frac{1}{2}}\,\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-\frac{1}{2}}\right) = 1,$$

which, after plugging into the bound for the iterates (14), shows that the method basically converges in one iteration. This observation indicates that sketch preconditioning can be extremely efficient, although it uses only the diagonal elements of matrices $\mathbf{L}_i$.

Now that we understand that the method can perform very well in the special case of $\tilde{\mathrm{b}}_i \equiv 0$, we can move on to a more complicated situation.

4. Irreducible Bias in the General Case

Now we look at the most general heterogeneous case with different matrices and linear terms $f_i(x) \equiv \frac{1}{2}x^\top \mathbf{L}_i x - x^\top \mathrm{b}_i$. In this instance, the gradient estimator (2) takes the form

$$g^k = \frac{1}{n}\sum_{i=1}^{n}\mathbf{C}_i^k\left(\mathbf{L}_i\mathbf{C}_i^k x^k - \mathrm{b}_i\right) = \overline{\mathbf{B}}^k x^k - \overline{\mathbf{C}\mathrm{b}}, \quad (16)$$

where $\overline{\mathbf{C}\mathrm{b}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{C}_i^k\,\mathrm{b}_i$. Herewith let us use a heterogeneous permutation sketch preconditioner (15), as in Section 3.2. Then $\mathbb{E}\left[\overline{\mathbf{B}}^k\right] = \mathbf{I}$ and $\mathbb{E}\left[\overline{\mathbf{C}\mathrm{b}}\right] = \frac{1}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}}$, where $\widetilde{\mathbf{D}\,\mathrm{b}} := \frac{1}{n}\sum_{i=1}^{n}\mathbf{D}_i^{-\frac{1}{2}}\,\mathrm{b}_i$. Furthermore, the expected gradient estimator (16) results in $\mathbb{E}\left[g^k\right] = x^k - \frac{1}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}}$ and can be transformed in the following manner:

$$\begin{aligned}\mathbb{E}\left[g^k\right] &= \overline{\mathbf{L}}^{-1}\,\overline{\mathbf{L}}\,x^k \pm \overline{\mathbf{L}}^{-1}\,\overline{\mathrm{b}} - \frac{1}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}} \\ &= \overline{\mathbf{L}}^{-1}\,\nabla f(x^k) + \underbrace{\overline{\mathbf{L}}^{-1}\,\overline{\mathrm{b}} - \frac{1}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}}}_{h},\end{aligned} \quad (17)$$

which reflects the decomposition of the estimator into the optimally preconditioned true gradient and a bias, depending on the linear terms b_i.

4.1. Bias of the Method

Estimator (17) can be directly plugged (with proper conditioning) into the general SGD update (2)

$$\begin{aligned}\mathbb{E}\left[x^{k+1}\right] &= (1-\gamma)x^k + \frac{\gamma}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}} \\ &= (1-\gamma)^{k+1}\,x^0 + \frac{\gamma}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}}\sum_{j=0}^{k}(1-\gamma)^j.\end{aligned} \quad (18)$$

The resulting recursion (18) is exact, and its asymptotic limit can be analyzed. Thus, for constant $\gamma < 1$, by using the formula for the sum of the first k terms of a geometric series, one gets

$$\mathbb{E}\left[x^k\right] = (1-\gamma)^k\,x^0 + \frac{1-(1-\gamma)^k}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}} \underset{k\to\infty}{\longrightarrow} \frac{1}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}},$$

which shows that in the limit, the first initialization term (with x^0) vanishes while the second converges to $\frac{1}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}}$. This reasoning shows that the method does not converge to the exact solution

$$\mathbb{E}\left[x^k\right] \to x^\infty \neq x^\star \in \operatorname*{arg\,min}_{x\in\mathbb{R}^d}\left\{\frac{1}{2}x^\top\,\overline{\mathbf{L}}\,x - x^\top\,\overline{\mathrm{b}}\right\},$$

which for the positive-definite $\overline{\mathbf{L}}$ can be defined as $x^\star = \overline{\mathbf{L}}^{-1}\,\overline{\mathrm{b}}$, while $x^\infty = \frac{1}{n\sqrt{n}}\sum_{i=1}^{n}\mathbf{D}_i^{-\frac{1}{2}}\,\mathrm{b}_i$. So, in general, there is an unavoidable bias. However, in the limit case: $n = d \to \infty$, the bias diminishes.

4.2. Generic Convergence Analysis

While the analysis in Section 4.1 is precise, it does not allow us to compare the convergence of IST to standard optimization methods. Therefore, we also analyze the non-asymptotic behavior of the method to understand the convergence speed. Our result is formalized in the following theorem:

Theorem 4.1. *Consider the method (2) with the estimator (16) for the quadratic problem (10) with the positive-definite matrix $\overline{\mathbf{L}} \succ 0$. Assume that for every $\mathbf{D}_i := \mathrm{Diag}(\mathbf{L}_i)$ matrices $\mathbf{D}_i^{-\frac{1}{2}}$ exist, scaled permutation sketches (15) are used, and heterogeneity is bounded as $\mathbb{E}\left[\left\|g^k - \mathbb{E}\left[g^k\right]\right\|_{\overline{\mathbf{L}}}^2\right] \leq \sigma^2$. Then, for the step size chosen as follows:*

$$0 < \gamma \leq \gamma_\beta := \frac{1/2-\beta}{\beta+1/2},$$

where $\gamma_{c,\beta} \in (0,1]$ for $\beta \in (0,1/2)$, the iterates satisfy

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|_{\overline{\mathbf{L}}^{-1}}^2\right] &\leq \frac{f(x^0)-\mathbb{E}\left[f(x^K)\right]}{0.5\gamma K} \\ &+ \left(\frac{1-\gamma}{0.5\beta}+\gamma\right)\|h\|_{\overline{\mathbf{L}}}^2 + \gamma\sigma^2,\end{aligned} \quad (19)$$

where $\overline{\mathbf{L}} = \frac{1}{n}\sum_{i=1}^n \mathbf{L}_i, h = \overline{\mathbf{L}}^{-1}\,\overline{\mathrm{b}} - \frac{1}{n^{3/2}}\sum_{i=1}^n \mathbf{D}_i^{-\frac{1}{2}}\,\mathrm{b}_i$ *and* $\overline{\mathrm{b}} = \frac{1}{n}\sum_{i=1}^n \mathrm{b}_i$.

Note that the derived convergence upper bound has a neighborhood proportional to the bias of the gradient estimator h and level of heterogeneity σ^2. Some of these terms with factor γ can be eliminated by decreasing the learning rate (e.g., $\sim 1/\sqrt{K}$). However, such a strategy does not diminish the term with a multiplier $2\beta^{-1}(1-\gamma)$, making the neighborhood irreducible. Moreover, this term can be eliminated for $\gamma = 1$, which also minimizes the first term that decreases as $1/K$. However, this step size choice maximizes the terms with factor γ. Thus, there exists an inherent trade-off between convergence speed and the size of the neighborhood.

In addition, convergence to the stationary point is measured by the weighted $\overline{\mathbf{L}}^{-1}$ squared norm of the gradient. At the same time, the neighborhood term depends on the weighted by $\overline{\mathbf{L}}$ norm of h. This fine-grained decoupling is achieved by carefully applying the Fenchel-Young inequality and provides a tighter characterization of the convergence compared to using standard Euclidean distances.

Homogeneous Case. In this scenario, every worker has access to all data $f_i(x) \equiv \frac{1}{2}x^\top \mathbf{L}x - x^\top \mathrm{b}$. Then diagonal preconditioning of the problem can be used, as in the previous Section 3.1. This results in a gradient $\nabla f(x) = \tilde{\mathbf{L}}\,x - \tilde{\mathrm{b}}$ for $\tilde{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$ and $\tilde{\mathrm{b}} = \mathbf{D}^{-\frac{1}{2}}\,\mathrm{b}$. If this expression is further combined with a permutation sketch scaled by $1/\sqrt{n}\,\mathbf{C}_i' := \sqrt{n}e_{\pi_i}e_{\pi_i}^\top$, the resulting gradient estimator is:

$$g^k = x^k - \frac{1}{\sqrt{n}}\,\tilde{\mathrm{b}} = \tilde{\mathbf{L}}^{-1}\,\nabla f(x^k) + \tilde{h}, \qquad (20)$$

for $\tilde{h} = \tilde{\mathbf{L}}^{-1}\,\tilde{\mathrm{b}} - \frac{1}{\sqrt{n}}\,\tilde{\mathrm{b}}$. In this case, the heterogeneity term σ^2 from the upper bound (19) disappears as $\mathbb{E}\left[\left\|g^k - \mathbb{E}\left[g^k\right]\right\|_{\tilde{\mathbf{L}}}^2\right] = 0$, which significantly decreases the neighborhood size. However, the bias term depending on $\tilde{h}$ still remains, as the method does not converge to the exact solution $x^k \to x^\infty \neq x^\star = \tilde{\mathbf{L}}^{-1}\,\tilde{\mathrm{b}}$ for positive-definite $\tilde{\mathbf{L}}$. Nevertheless the method's fixed point $x^\infty = \tilde{\mathrm{b}}/\sqrt{n}$ and solution $x^\star$ can coincide when $\tilde{\mathbf{L}}^{-1}\,\tilde{\mathrm{b}} = \frac{1}{\sqrt{n}}\,\tilde{\mathrm{b}}$, which means that $\tilde{\mathrm{b}}$ is the right eigenvector of matrix $\tilde{\mathbf{L}}^{-1}$ with eigenvalue $\frac{1}{\sqrt{n}}$.

Let us contrast the obtained result (19) with the non-convex rate of SGD (Khaled & Richtárik, 2023) with constant step size γ for L-smooth and lower-bounded f

$$\min_{k<K-1} \mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \frac{f(x^0) - \inf f}{0.1\gamma K} + \gamma LC, \quad (21)$$

where constant C depends, for example, on the variance of the stochastic gradient estimator. Observe that the first term in the compared upper bounds (21) and (19) is almost identical and decreases with speed $1/K$. However, unlike (19), the neighborhood for SGD can be completely eliminated by reducing the step size γ. This highlights a fundamental difference between our results and unbiased methods. The intuition behind this issue is that for SGD-type methods like compressed gradient descent

$$x^{k+1} = x^k - \mathcal{C}(\nabla f(x^k))$$

the gradient estimate is unbiased and enjoys the property that variance

$$\mathbb{E}\left[\|\mathcal{C}(\nabla f(x^k)) - \nabla f(x^k)\|^2\right] \leq \omega\|\nabla f(x^k)\|^2$$

goes down to zero as the method progresses because $\nabla f(x^k) \to \nabla f(x^\star) = 0$ in the unconstrained case. In addition, any stationary point $x^\star$ ceases to be a fixed point of the iterative procedure as

$$x^\star \neq x^\star - \nabla f(\mathcal{C}(x^\star)),$$

in the general case, unlike for compressed gradient descent with both biased and unbiased compressors $\mathcal{C}$. Thus, even if the method—computing the gradient with a sparse model—is initialized from the *solution* after one gradient step, the method may get away from the optimum.

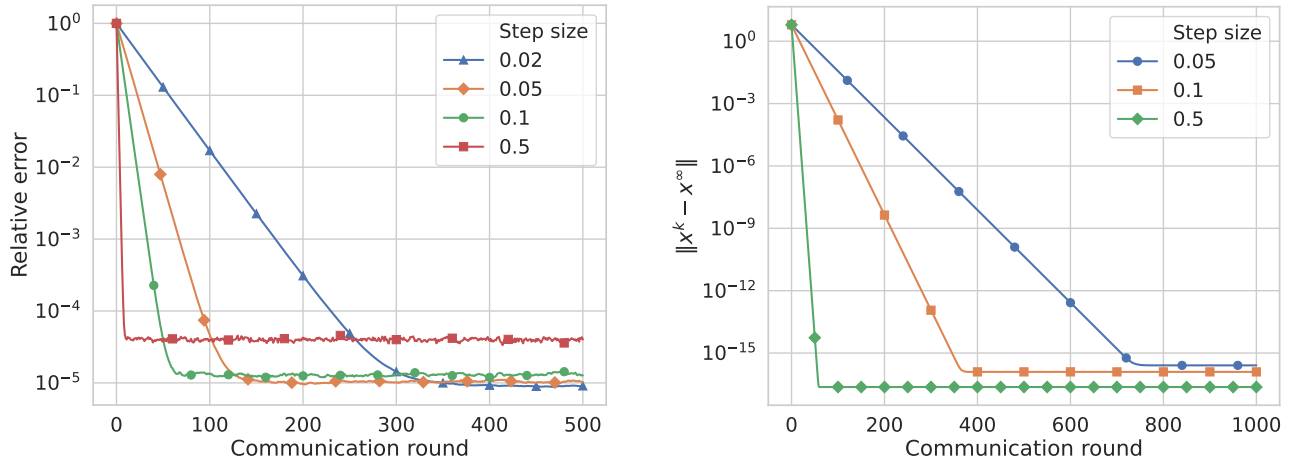
4.3. Comparison to Previous Works

Independent Subnetwork Training (Yuan et al., 2022). There are several improvements over the previous works that tried to theoretically analyze the convergence of distributed IST.

The first difference is that our results allow for an almost arbitrary level of model sparsification, i.e., will work for any $\omega \geq 0$ as permutation sketches can be viewed as a special case of compression operators (2.1). This represents a significant improvement over the work of Yuan et al. (2022), which demands[3] $\omega \lesssim \mu^2/L^2$. Such a requirement is very restrictive as the condition number L/μ of the loss function f is typically very large for any non-trivial optimization problem. Thus, the sparsifier's (4) variance $\omega = d/q - 1$ has to be very close to 0 and $q \approx d$. Thus, the previous theory allows almost no compression (sparsification) because it is based on the analysis of gradient descent with compressed iterates (Khaled & Richtárik, 2019).

The second distinction is that the original IST work (Yuan et al., 2022) considered a single node setting, and thus their convergence bounds did not capture the effect of heterogeneity, which we believe is of crucial importance for distributed settings (Chraibi et al., 2019; Shulgin & Richtárik, 2022).

[3] μ refers to a constant from the Polyak-Łojasiewicz (or strong convexity) condition. In case of a quadratic problem with positive-definite matrix $\mathbf{A}$ constant μ equals to $\lambda_{\min}(\mathbf{A})$



(a) Function convergence for heterogeneous case.

(b) Iterates convergence for homogeneous case.

Figure 1. Performance of simplified IST on a quadratic problem for varying step size values.

Moreover, the original work considers the Lipschitz continuity of the loss function f, which is not satisfied for a simple quadratic model. A more detailed comparison, including additional assumptions on the gradient estimator made by Yuan et al. (2022), is presented in the Appendix D.

FL with Model Pruning. In a recent work, Zhou et al. (2022) made an attempt to analyze a variant of the FedAvg algorithm with sparse local initialization and compressed gradient training (pruned local models). They considered a case of L-smooth loss and a sparsification operator satisfying a similar condition to (2.1). However, they also assumed that the squared norm of the stochastic gradient is uniformly bounded (9), which is "pathological" (Khaled et al., 2020)—especially in the case of local methods—as it does not allow the analysis to capture the very important effect of heterogeneity and can result in vacuous bounds.

In the Appendix D, we present some limitations of other relevant previous approaches to training with compressed models: too restrictive assumptions on the algorithm (Mohtashami et al., 2022) or non-applicability in our problem setting (Chayti & Karimireddy, 2024). In addition, we discuss the differences between IST and 3D parallelism (Shoeybi et al., 2019).

5. Experiments

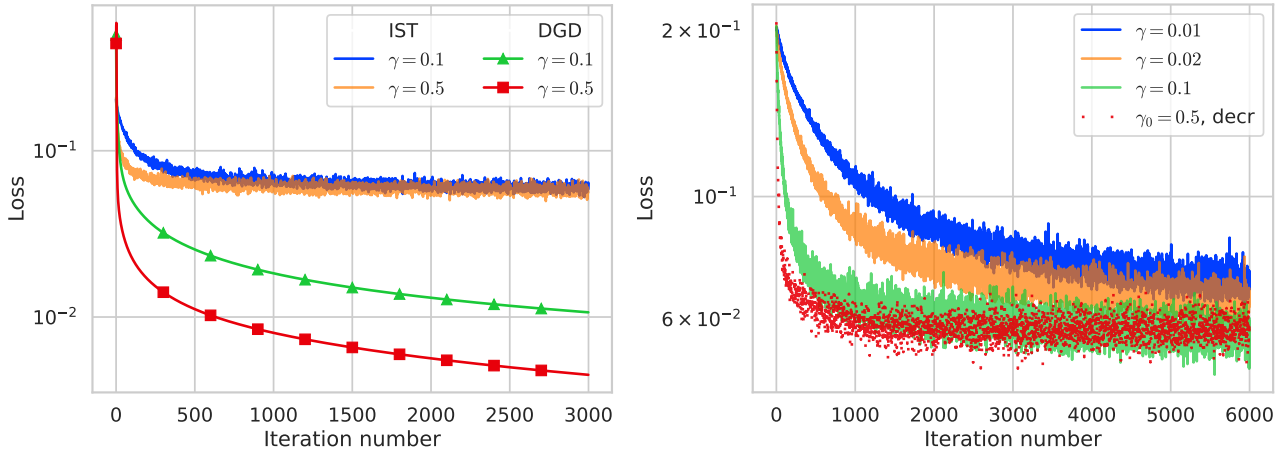
To empirically validate our theoretical framework and its implications, we focus on carefully controlled setting that satisfy the assumptions of our work. Specifically, we consider a quadratic problem defined in (10), where $\mathbf{L}_i = \mathbf{B}_i^\top \mathbf{B}_i$. Entries of the matrices $\mathbf{B}_i \in \mathbb{R}^{d\times d}$, vectors $\mathrm{b}_i \in \mathbb{R}^d$, and initialization $x^0 \in \mathbb{R}^d$ are generated from a standard Gaussian distribution $\mathcal{N}(0,1)$.

Heterogeneous setting. In Figure 1(a), we present the performance of the simplified Independent Subnetwork Training (IST) algorithm (update (2) with estimator (16)) for a heterogeneous problem. We fix the dimension d to 1000 and the number of computing nodes n to 10. We evaluate the logarithm of a relative functional error $\log(f(x^k) - f(x^\star))/(f(x^0) - f(x^\star))$, while the horizontal axis denotes the number of communication rounds required to achieve a certain error tolerance. According to our theory (19) and (54), the method converges to a neighborhood of the solution, which depends on the chosen step size. Specifically, a larger step size allows for faster convergence but results in a larger neighborhood.

Homogeneous setting. In Figure 1(b), we demonstrate the convergence of the iterates x^k for a homogeneous problem with $d = n = 50$. The results are in close agreement with our theoretical predictions for the estimator (20). We observe that the distance to the method's expected fixed point $x^\infty = \tilde{\mathrm{b}}/\sqrt{n}$ decreases linearly for different step size values. This confirms that IST may not converge to the optimal solution $x^\star = \tilde{\mathbf{L}}^{-1}\tilde{\mathrm{b}}$ of the original problem (10) in general (no interpolation) cases. In addition, there are no visible oscillations in comparison to the heterogeneous case.

5.1. Neural Network Results

We closely follow the experimental setup of Liao & Kyrillidis (2022) and use a one-hidden-layer neural network with ReLU activations. For completeness, we repeat some of the details here. ResNet-50 model (He et al., 2016) pre-trained on ImageNet is used as a feature extractor and concatenated



(a) Comparison of IST and Distributed Gradient Descent (DGD).

(b) Performance of IST with different step sizes.

Figure 2. Experimental study of IST on a neural network problem.

with two fully connected layers. The resulting model is then trained on the CIFAR-10 (Krizhevsky et al., 2009) dataset. We take the outputs of the re-trained ResNet-50 as the input features resulting in $d = 2048$, and the logit outputs of the combined model are used as the labels.

The goal of the experiment is to study the optimization performance of the IST method described in Section 2. Namely, we consider Algorithm 1 with 1) $\mathbf{C}_i$ chosen as `Perm-q` (6) for IST and 2) $\mathbf{C}_i = \mathbf{I}$ for Distributed Gradient Descent (DGD). Both methods are implemented across $n = 10$ nodes, employing constant step sizes γ, and one local step per communication round.

Figure 2(a) shows the logarithm of the average loss of submodels after every iteration. The main observation is that IST's performance stagnates (starts oscillating) at some point, unlike DGD's. Namely, the studied method converges to the neighborhood whose size is basically the same for different step size values. This phenomenon distinguishes the method from standard SGD. Moreover, larger γ speeds up DGD and allows reaching the error floor (oscillation level) faster for IST. These observations agree well with our Theorems 4.1 and B.4 as the convergence upper bound (19) has an irreducible term proportional to the bias norm $\|h\|^2_{\mathbf{L}}$ of the gradient estimator. This term can not be eliminated completely by decreasing the step size γ, unlike SGD (21).

In Figure 2(b), we take a closer look at the training loss during IST optimization. For this problem, the situation differs from Figure 1(a) as for even smaller step size values ($\gamma \in \{0.01, 0.02\}$), the method converges to a higher error floor. Interestingly, if γ is decreased by 10 every 1000 iterations, the method's performance (red dotted curve) almost does not change. This can be explained by the second term $(1-\gamma)\,\beta^{-1}\,\|h\|^2_{\mathbf{L}}$ from Theorem 4.1 convergence bound (19), which increases for smaller γ. The observed effect distinguishes IST from the typical training situation in deep learning.

6. Conclusions and Future Work

In this study, we introduced a novel approach to understanding training with combined model and data parallelism for a quadratic model. Our framework sheds light on distributed submodel optimization, which reveals the advantages and limitations of Independent Subnetwork Training (IST). Moreover, we accurately characterized the behavior of the considered method in both homogeneous and heterogeneous scenarios without imposing restrictive assumptions on the gradient estimators.

In future research, it would be valuable to explore extensions of our findings to settings that are closer to scenarios, such as cross-device federated learning. This could involve investigating partial participation support, leveraging local training benefits, and ensuring robustness against stragglers.

It would be interesting to generalize our analysis to non-quadratic scenarios without relying on pathological assumptions. This work shows a somewhat negative result regarding worst-case IST performance for standard Empirical Risk Minimization problem (1). However, IST was recently shown to be effective at solving an alternative optimization problem formulation by Demidovich et al. (2023). Another potential promising research direction is algorithmic modifications of the original IST to solve the fundamental problems highlighted in this work and acceleration of training.

Acknowledgements

We would like to thank anonymous reviewers, Avetik Karagulyan (KAUST) and Andrea Devlin (KAUST) for their helpful comments and suggestions to improve the manuscript.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ajalloeian, A. and Stich, S. U. On the convergence of SGD with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020. (Cited on page 4)

Alam, S., Liu, L., Yan, M., and Zhang, M. FedRolex: Model-heterogeneous federated learning with rolling sub-model extraction. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=OtxyysUdBE. (Cited on page 2)

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30, 2017. (Cited on pages 1, 2, and 3)

Arjevani, Y., Shamir, O., and Srebro, N. A tight convergence analysis for stochastic gradient descent with delayed updates. In *Algorithmic Learning Theory*, pp. 111–132. PMLR, 2020. (Cited on page 4)

Bian, S., Li, D., Wang, H., Xing, E. P., and Venkataraman, S. Does compressing activations help model parallel training? *arXiv preprint arXiv:2301.02654*, 2023. (Cited on page 28)

Caldas, S., Konečny, J., McMahan, H. B., and Talwalkar, A. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018. (Cited on page 2)

Charles, Z., Bonawitz, K., Chiknavaryan, S., McMahan, B., et al. Federated select: A primitive for communication- and memory-efficient federated learning. *arXiv preprint arXiv:2208.09432*, 2022. (Cited on page 2)

Chayti, E. M. and Karimireddy, S. P. Optimization with access to auxiliary information. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=kxYqgSkH8I. (Cited on pages 8 and 28)

Chen, Y., Chen, Z., Wu, P., and Yu, H. FEDOBD: Opportunistic block dropout for efficiently training large-scale neural networks through federated learning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 3541–3549, 2023. (Cited on page 2)

Chraibi, S., Khaled, A., Kovalev, D., Richtárik, P., Salim, A., and Takáč, M. Distributed fixed point methods with compressed iterates. *arXiv preprint arXiv:2102.07245*, 2019. (Cited on page 7)

Cunha, L., Gidel, G., Pedregosa, F., Scieur, D., and Paquette, C. Only tails matter: Average-case universality and robustness in the convex regime. In *International Conference on Machine Learning*, pp. 4474–4491. PMLR, 2022. (Cited on page 4)

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, M., Senior, A., Tucker, P., Yang, K., et al. Large scale distributed deep networks. *Advances in Neural Information Processing Systems*, 25, 2012. (Cited on page 1)

Demidovich, Y., Malinovsky, G., Shulgin, E., and Richtárik, P. MAST: Model-agnostic sparsified training. *arXiv preprint arXiv:2311.16086*, 2023. (Cited on page 9)

Diao, E., Ding, J., and Tarokh, V. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=TNkPBBYFkXg. (Cited on page 2)

Dun, C., Wolfe, C. R., Jermaine, C. M., and Kyrillidis, A. ResIST: Layer-wise decomposition of resnets for distributed training. In *Uncertainty in Artificial Intelligence*, pp. 610–620. PMLR, 2022. (Cited on page 1)

Dun, C., Hipolito, M., Jermaine, C., Dimitriadis, D., and Kyrillidis, A. Efficient and light-weight federated learning via asynchronous distributed dropout. In *International Conference on Artificial Intelligence and Statistics*, pp. 6630–6660. PMLR, 2023. (Cited on pages 2 and 28)

Farber, P. and Asanovic, K. Parallel neural network training on multi-spert. In *Proceedings of 3rd International Conference on Algorithms and Architectures for Parallel Processing*, pp. 659–666. IEEE, 1997. (Cited on page 1)

Gorbunov, E., Hanzely, F., and Richtárik, P. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 680–690. PMLR, 2020. (Cited on page 3)

Goujaud, B., Scieur, D., Dieuleveut, A., Taylor, A. B., and Pedregosa, F. Super-acceleration with cyclical step-sizes. In *International Conference on Artificial Intelligence and Statistics*, pp. 3028–3065. PMLR, 2022. (Cited on page 4)

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. SGD: General analysis and improved rates. *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*, 2019. (Cited on page 3)

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2018. (Cited on page 1)

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. (Cited on page 8)

Horvath, S., Laskaridis, S., Almeida, M., Leontiadis, I., Venieris, S., and Lane, N. FjORD: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, 34:12876–12889, 2021. (Cited on page 2)

Jiang, Y., Wang, S., Valls, V., Ko, B. J., Lee, W.-H., Leung, K. K., and Tassiulas, L. Model pruning enables efficient federated learning on edge devices. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. (Cited on page 2)

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K. A., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021. doi: 10.1561/2200000083. URL https://doi.org/10.1561/2200000083. (Cited on page 2)

Khaled, A. and Richtárik, P. Gradient descent with compressed iterates. *arXiv preprint arXiv:1909.04716*, 2019. (Cited on pages 7 and 24)

Khaled, A. and Richtárik, P. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=AU4qHN2VkS. Survey Certification. (Cited on pages 3, 7, and 23)

Khaled, A., Mishchenko, K., and Richtárik, P. Tighter theory for local SGD on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529. PMLR, 2020. (Cited on pages 4 and 8)

Khirirat, S., Feyzmahdavian, H. R., and Johansson, M. Distributed learning with compressed gradients. *arXiv preprint arXiv:1806.06573*, 2018. (Cited on page 3)

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *NIPS Private Multi-Party Machine Learning Workshop*, 2016. (Cited on pages 1 and 2)

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6(1):1, 2009. (Cited on page 9)

Li, Z., Bao, H., Zhang, X., and Richtárik, P. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pp. 6286–6295. PMLR, 2021. (Cited on page 22)

Liao, F. and Kyrillidis, A. On the convergence of shallow neural network training with randomly masked neurons. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=e7mYYMSyZH. (Cited on pages 2, 8, and 28)

Lin, R., Xiao, Y., Yang, T.-J., Zhao, D., Xiong, L., Motta, G., and Beaufays, F. Federated pruning: Improving neural network efficiency with federated learning. *arXiv preprint arXiv:2209.06359*, 2022. (Cited on page 2)

Lin, T., Stich, S. U., Barba, L., Dmitriev, D., and Jaggi, M. Dynamic model pruning with feedback. In *International Conference on Learning Representations*, 2019. (Cited on page 4)

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017. (Cited on page 2)

Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In *International Conference on Machine Learning*, pp. 15750–15769. PMLR, 2022. (Cited on page 4)

Mohtashami, A., Jaggi, M., and Stich, S. Masked training of neural networks with partial gradients. In *International Conference on Artificial Intelligence and Statistics*, pp. 5876–5890. PMLR, 2022. (Cited on pages 8 and 27)

Nesterov, Y. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. (Cited on page 3)

Qiu, X., Fernandez-Marques, J., Gusmao, P. P., Gao, Y., Parcollet, T., and Lane, N. D. ZeroFL: Efficient on-device training for federated learning with local sparsity. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=2sDQwC_hmnM. (Cited on page 2)

Richtárik, P. and Takáč, M. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144 (1-2):1–38, 2014. (Cited on page 3)

Richtárik, P. and Takáč, M. Distributed coordinate descent method for learning with big data. *Journal of Machine Learning Research*, 17(75):1–25, 2016. (Cited on page 1)

Safaryan, M., Hanzely, F., and Richtárik, P. Smoothness matrices beat smoothness constants: Better communication compression techniques for distributed optimization. *Advances in Neural Information Processing Systems*, 34: 25688–25702, 2021. (Cited on page 3)

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. (Cited on pages 1 and 2)

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019. (Cited on pages 8 and 28)

Shulgin, E. and Richtárik, P. Shifted compression framework: Generalizations and improvements. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. (Cited on page 7)

Stich, S. U. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019. (Cited on page 25)

Szlendak, R., Tyurin, A., and Richtárik, P. Permutation compressors for provably faster distributed nonconvex optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=GugZ5DzzAu. (Cited on pages 3, 5, 21, 22, and 23)

Wang, B., Safaryan, M., and Richtárik, P. Theoretically better and numerically faster distributed optimization with smoothness-aware quantization techniques. *Advances in Neural Information Processing Systems*, 35:9841–9852, 2022. (Cited on page 3)

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021. (Cited on page 2)

Wen, D., Jeon, K.-J., and Huang, K. Federated dropout—a simple approach for enabling federated learning on resource constrained devices. *IEEE Wireless Communications Letters*, 11(5):923–927, 2022. (Cited on page 2)

Wolfe, C. R., Yang, J., Liao, F., Chowdhury, A., Dun, C., Bayer, A., Segarra, S., and Kyrillidis, A. GIST: Distributed training for large-scale graph convolutional networks. *Journal of Applied and Computational Topology*, pp. 1–53, 2023. (Cited on page 1)

Yang, T.-J., Guliani, D., Beaufays, F., and Motta, G. Partial variable training for efficient on-device federated learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4348–4352. IEEE, 2022. (Cited on page 2)

Yuan, B., Wolfe, C. R., Dun, C., Tang, Y., Kyrillidis, A., and Jermaine, C. Distributed learning of fully connected neural networks using independent subnet training. *Proceedings of the VLDB Endowment*, 15(8):1581–1590, 2022. (Cited on pages 1, 7, 8, 14, 24, and 26)

Zhang, G., Li, L., Nado, Z., Martens, J., Sachdeva, S., Dahl, G., Shallue, C., and Grosse, R. B. Which algorithmic choices matter at which batch sizes? insights from a noisy quadratic model. *Advances in neural information processing systems*, 32, 2019. (Cited on page 4)

Zhang, X., Mckenna, M., Mesirov, J., and Waltz, D. An efficient implementation of the back-propagation algorithm on the connection machine CM-2. *Advances in neural information processing systems*, 2, 1989. (Cited on page 1)

Zhou, H., Lan, T., Venkataramani, G., and Ding, W. On the convergence of heterogeneous federated learning with arbitrary adaptive online model pruning. *arXiv preprint arXiv:2201.11803*, 2022. (Cited on pages 4 and 8)

Zhu, L., Liu, C., Radhakrishnan, A., and Belkin, M. Quadratic models for understanding neural network dynamics. In *The Twelfth International Conference on Learning Representations*, 2023. (Cited on page 4)

Zinkevich, M., Weimer, M., Li, L., and Smola, A. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010. (Cited on page 1)

Contents

1 Introduction **1**

1.1 The Need for Model Parallelism . . . 1

1.2 Summary of Contributions . . . 2

2 Formalism and Setup **2**

2.1 Issues with Existing Approaches . . . 3

2.2 Simplifications Taken . . . 4

3 Results in the Interpolation Case **4**

3.1 Homogeneous Problem Preconditioning . . . 5

3.2 Heterogeneous Sketch Preconditioning . . . 5

4 Irreducible Bias in the General Case **6**

4.1 Bias of the Method . . . 6

4.2 Generic Convergence Analysis . . . 6

4.3 Comparison to Previous Works . . . 7

5 Experiments **8**

5.1 Neural Network Results . . . 8

6 Conclusions and Future Work **9**

A Basic and Auxiliary Facts **14**

B Proofs **14**

B.1 Permutation Sketch Computations . . . 14

B.1.1 Heterogeneous sketch preconditioning. . . . 15

B.2 Interpolation Case: Proof of Theorem 3.1 . . . 16

B.3 Non-zero Solution . . . 17

B.3.1 Convergence analysis for heterogeneous case: proof of Theorem 4.1. . . . 18

B.3.2 Homogeneous case . . . 19

B.4 Generalization to $n \neq d$ Case. . . . 21

C Generalization Beyond Quadratics **22**

C.1 Preliminary Facts . . . 22

C.2 Convergence Analysis . . . 23

D Comparison to Related Works **26**

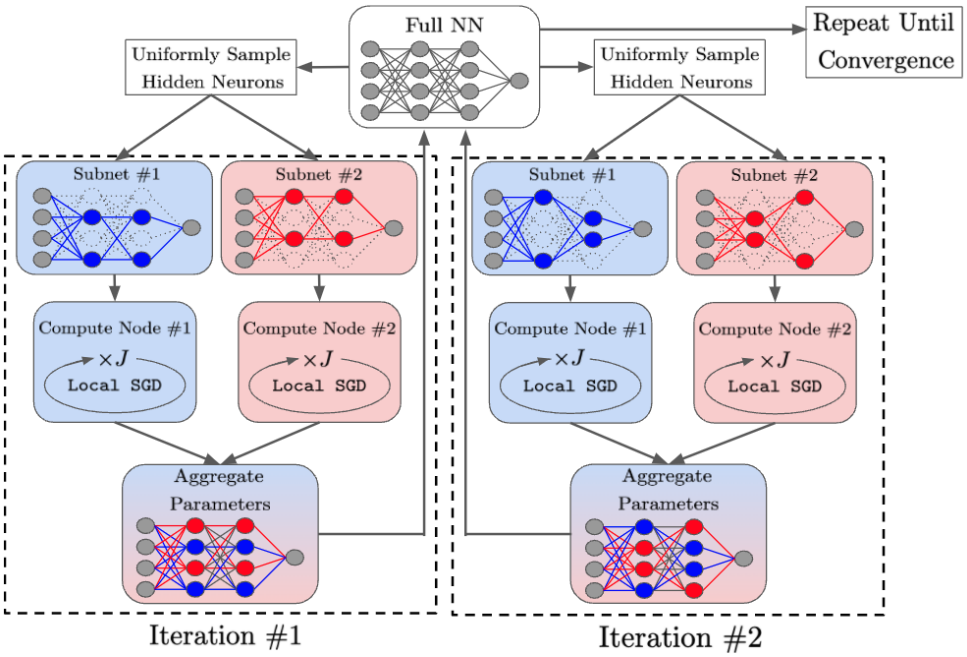


Figure 3. Schematic depiction of a Neural Network trained with IST across two nodes (source: Yuan et al. (2022)).

A. Basic and Auxiliary Facts

L-matrix smoothness:

$$f(x+h) \leq f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \langle \mathbf{L}h, h \rangle, \qquad \forall x, h \in \mathbb{R}^d. \tag{22}$$

Basic Inequalities. For all vectors $a, b \in \mathbb{R}^d$ and random vector $X \in \mathbb{R}^d$:

$$2\langle a, b \rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2, \tag{23}$$

$$\|a+b\|^2 \leq (1+\beta)\|a\|^2 + (1+\beta^{-1})\|b\|^2, \text{ for } \beta > 0 \tag{24}$$

$$\mathbf{E}\,\|X-a\|^2 = \mathbf{E}\,\|X - \mathbf{E}\,X\|^2 + \|\,\mathbf{E}\,X - a\|^2. \tag{25}$$

For a set of $n \geq 1$ vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ it holds

$$\left\| \frac{1}{n} \sum_{i=1}^n a_i \right\|^2 \leq \frac{1}{n} \sum_{i=1}^n \|a_i\|^2. \tag{26}$$

Lemma A.1 (Fenchel–Young inequality)**.** *For any function f and its convex conjugate f^*, Fenchel's inequality (also known as the Fenchel–Young inequality) holds for every $x, y \in \mathbb{R}^d$*

$$\langle x, y \rangle \leq f(x) + f^*(y).$$

The proof follows from the definition of conjugate: $f^*(y) := \sup_{x'} \{\langle y, x' \rangle - f(x')\} \geq \langle y, x \rangle - f(x)$.

In the case of a quadratic function $f(x) = \beta\|x\|^2_{\mathbf{L}}$, we can compute $f^*(y) = \frac{1}{4}\beta^{-1}\|y\|^2_{\mathbf{L}^{-1}}$. Thus

$$\langle x, y \rangle \leq \beta\|x\|^2_{\mathbf{L}} + \frac{1}{4}\beta^{-1}\|y\|^2_{\mathbf{L}^{-1}}. \tag{27}$$

B. Proofs

B.1. Permutation Sketch Computations

All derivations in this section are performed for the $n = d$ case.

Classical Permutation Sketching. `Perm-1`: $\mathbf{C}_i = ne_{\pi_i}e_{\pi_i}^\top$, where $\pi = (\pi_1, \ldots, \pi_n)$ is a random permutation of $[n]$. For the homogeneous problem $\mathbf{L}_i \equiv \mathbf{L}$:

$$\mathbb{E}\left[\overline{\mathbf{B}}^k\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{C}_i\,\mathbf{L}\,\mathbf{C}_i\right] = n\mathrm{Diag}(\mathbf{L}) \tag{28}$$

Then

$$2\,\overline{\mathbf{W}} = \mathbb{E}\left[\mathbf{L}\,\overline{\mathbf{B}}^k + \overline{\mathbf{B}}^k\,\mathbf{L}\right] = n\left(\mathbf{L}\mathrm{Diag}(\mathbf{L}) + \mathrm{Diag}(\mathbf{L})\mathbf{L}\right) \tag{29}$$

and

$$\mathbb{E}\left[\overline{\mathbf{B}}^k\,\mathbf{L}\,\overline{\mathbf{B}}^k\right] = n^2\mathrm{Diag}(\mathbf{L})\mathbf{L}\mathrm{Diag}(\mathbf{L}). \tag{30}$$

By repeating basically the same calculations for $\mathbf{C}_i' = \sqrt{n}e_{\pi_i}e_{\pi_i}^\top$ we have that

$$\mathbb{E}\left[\overline{\mathbf{B}}^k\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{C}_i'\mathbf{L}\mathbf{C}_i'\right] = \mathrm{Diag}(\mathbf{L}), \tag{31}$$

and $\mathbb{E}\left[\overline{\mathbf{B}}^k\,\mathbf{L}\,\overline{\mathbf{B}}^k\right] = \mathrm{Diag}(\mathbf{L})\mathbf{L}\mathrm{Diag}(\mathbf{L})$, $2\,\overline{\mathbf{W}} = \mathbb{E}\left[\mathbf{L}\,\overline{\mathbf{B}}^k + \overline{\mathbf{B}}^k\,\mathbf{L}\right] = \mathbf{L}\mathrm{Diag}(\mathbf{L}) + \mathrm{Diag}(\mathbf{L})\mathbf{L}$.

B.1.1. Heterogeneous sketch preconditioning.

We recall the following modification of `Perm-1`:

$$\tilde{\mathbf{C}}_i \coloneqq \sqrt{n/\left[\mathbf{L}_i\right]_{\pi_i,\pi_i}}e_{\pi_i}e_{\pi_i}^\top. \tag{32}$$

Then

$$\mathbb{E}\left[\tilde{\mathbf{C}}_i\mathbf{L}_i\tilde{\mathbf{C}}_i\right] = \mathbb{E}\left[n[\mathbf{L}_i]_{\pi_i,\pi_i}^{-1}e_{\pi_i}e_{\pi_i}^\top\mathbf{L}_ie_{\pi_i}e_{\pi_i}^\top\right] = \frac{1}{n}\sum_{j=1}^n ne_j\mathbf{I}_{j,j}e_j^\top = \mathbf{I}. \tag{33}$$

and

$$\begin{aligned}
\mathbb{E}\left[\overline{\mathbf{B}}^k\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{C}}_i\mathbf{L}_i\tilde{\mathbf{C}}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[n[\mathbf{L}_i]_{\pi_i,\pi_i}^{-1}e_{\pi_i}e_{\pi_i}^\top\mathbf{L}_ie_{\pi_i}e_{\pi_i}^\top\right] \\
&= \frac{1}{n}\sum_{i=1}^n\frac{1}{n}\sum_{j=1}^n n[\mathbf{L}_i]_{j,j}^{-1}e_j[\mathbf{L}_i]_{j,j}e_j^\top \\
&= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n e_je_j^\top \\
&= \mathbf{I}.
\end{aligned}$$

Thus $\overline{\mathbf{W}} = \frac{1}{2}\mathbb{E}\left[\overline{\mathbf{L}}\,\overline{\mathbf{B}}^k + \overline{\mathbf{B}}^k\,\overline{\mathbf{L}}\right] = \overline{\mathbf{L}}$. On the left hand side of inequality (12), we have

$$\begin{aligned}
\mathbb{E}\left[\overline{\mathbf{B}}^k\,\overline{\mathbf{L}}\,\overline{\mathbf{B}}^k\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \tilde{\mathbf{C}}_i\mathbf{L}_i\tilde{\mathbf{C}}_i\,\overline{\mathbf{L}}\,\frac{1}{n}\sum_{i=j}^n \tilde{\mathbf{C}}_j\mathbf{L}_j\tilde{\mathbf{C}}_j\right] \\
&= \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}\left[\tilde{\mathbf{C}}_i\mathbf{L}_i\tilde{\mathbf{C}}_i\,\overline{\mathbf{L}}\,\tilde{\mathbf{C}}_j\mathbf{L}_j\tilde{\mathbf{C}}_j\right] \\
&= \sum_{i,j=1}^n e_ie_i^\top\,\overline{\mathbf{L}}\,e_je_j^\top \\
&= \mathbf{I}\,\overline{\mathbf{L}}\,\mathbf{I} \\
&= \overline{\mathbf{L}}\,.
\end{aligned}$$

B.2. Interpolation Case: Proof of Theorem 3.1

In the quadratic interpolation regime, the linear term is zero $f_i(x) = \frac{1}{2}x^\top \mathbf{L}_i x$, and the gradient estimator has the form

$$g^k = \frac{1}{n}\sum_{i=1}^n \mathbf{C}_i^k \nabla f_i(\mathbf{C}_i^k x^k) = \frac{1}{n}\sum_{i=1}^n \mathbf{C}_i^k \mathbf{L}_i \mathbf{C}_i^k x^k = \overline{\mathbf{B}}^k x^k. \tag{34}$$

Proof. First, we prove the **stationary point** convergence result (13).

Using the $\overline{\mathbf{L}}$-smoothness of function f, we get

$$\begin{aligned} f(x^{k+1}) \overset{(2)}{=} f(x^k - \gamma g^k) &\overset{(7)}{\leq} f(x^k) - \left\langle \nabla f(x^k), \gamma g^k \right\rangle + \frac{\gamma^2}{2}\left\|g^k\right\|^2_{\overline{\mathbf{L}}} \\ &\overset{(11)}{=} f(x^k) - \gamma \left\langle \overline{\mathbf{L}}\, x^k, \overline{\mathbf{B}}^k x^k \right\rangle + \frac{\gamma^2}{2}\left\|\overline{\mathbf{B}}^k x^k\right\|^2_{\overline{\mathbf{L}}} \\ &= f(x^k) - \gamma (x^k)^\top \overline{\mathbf{L}}\, \overline{\mathbf{B}}^k x^k + \frac{\gamma^2}{2}(x^k)^\top \overline{\mathbf{B}}^k \overline{\mathbf{L}}\, \overline{\mathbf{B}}^k x^k. \end{aligned}$$

After applying conditional expectation, using its linearity, and the fact that

$$x^\top \mathbf{A} x = \frac{1}{2} x^\top \left(\mathbf{A} + \mathbf{A}^\top\right) x$$

we get

$$\begin{aligned} \mathbb{E}\left[f(x^{k+1}) \mid x^k\right] &\leq f(x^k) - \gamma (x^k)^\top \mathbb{E}\left[\overline{\mathbf{L}}\, \overline{\mathbf{B}}^k\right] x^k + \frac{\gamma^2}{2}(x^k)^\top \mathbb{E}\left[\overline{\mathbf{B}}^k \overline{\mathbf{L}}\, \overline{\mathbf{B}}^k\right] x^k \\ &= f(x^k) - \gamma (x^k)^\top \overline{\mathbf{W}}\, x^k + \frac{\gamma^2}{2}(x^k)^\top \mathbb{E}\left[\overline{\mathbf{B}}^k \overline{\mathbf{L}}\, \overline{\mathbf{B}}^k\right] x^k \\ &= f(x^k) - \gamma (\nabla f(x^k))^\top \overline{\mathbf{L}}^{-1} \overline{\mathbf{W}}\, \overline{\mathbf{L}}^{-1} \nabla f(x^k) \\ &\quad + \frac{\gamma^2}{2}(\nabla f(x^k))^\top \overline{\mathbf{L}}^{-1} \mathbb{E}\left[\overline{\mathbf{B}}^k \overline{\mathbf{L}}\, \overline{\mathbf{B}}^k\right] \overline{\mathbf{L}}^{-1} \nabla f(x^k) \\ &\overset{(12)}{\leq} f(x^k) - \gamma \|\nabla f(x^k)\|^2_{\overline{\mathbf{L}}^{-1}\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-1}} + \frac{\theta\gamma^2}{2}\|\nabla f(x^k)\|^2_{\overline{\mathbf{L}}^{-1}\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-1}} \\ &= f(x^k) - \gamma\left(1 - \theta\gamma/2\right)\|\nabla f(x^k)\|^2_{\overline{\mathbf{L}}^{-1}\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-1}} \\ &\leq f(x^k) - \frac{\gamma}{2}\|\nabla f(x^k)\|^2_{\overline{\mathbf{L}}^{-1}\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-1}}, \end{aligned}$$

where the last inequality holds for the stepsize $\gamma \leq \frac{1}{\theta}$.

Rearranging gives

$$\left\|\nabla f(x^k)\right\|^2_{\overline{\mathbf{L}}^{-1}\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-1}} \leq \frac{2}{\gamma}\left(f(x^k) - \mathbb{E}\left[f(x^{k+1}) \mid x^k\right]\right),$$

which after averaging gives the desired result

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2_{\overline{\mathbf{L}}^{-1}\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-1}}\right] \leq \frac{2}{\gamma K}\sum_{k=0}^{K-1}\left(f(x^k) - \mathbb{E}\left[f(x^{k+1})\right]\right) = \frac{2\left(f(x^0) - \mathbb{E}\left[f(x^K)\right]\right)}{\gamma K}. \tag{35}$$

Now we show the result for the **iterates convergence** (14).

Expectation conditioned on x^k:

$$\begin{aligned}
\mathbb{E}\left[\|x^{k+1}-x^\star\|^2_{\overline{\mathbf{L}}}\right] &= \mathbb{E}\left[\|x^k-\gamma g^k-x^\star\|^2_{\overline{\mathbf{L}}}\right] \\
&= \|x^k-x^\star\|^2_{\overline{\mathbf{L}}} - 2\gamma\left\langle x^k-x^\star, \mathbb{E}\left[\overline{\mathbf{L}}\,\overline{\mathbf{B}}^k\right](x^k-x^\star)\right\rangle \\
&\qquad +\gamma^2\left\langle \mathbb{E}\left[\overline{\mathbf{B}}^k\,\overline{\mathbf{L}}\,\overline{\mathbf{B}}^k\right](x^k-x^\star), x^k-x^\star\right\rangle \\
&\overset{x^\star=0}{=} \|x^k-x^\star\|^2_{\overline{\mathbf{L}}} - 2\gamma\left\langle x^k-x^\star, \overline{\mathbf{W}}(x^k-x^\star)\right\rangle \\
&\qquad +\gamma^2\left\langle x^k-x^\star, \mathbb{E}\left[\overline{\mathbf{B}}^k\,\overline{\mathbf{L}}\,\overline{\mathbf{B}}^k\right](x^k-x^\star)\right\rangle \\
&\overset{(12)}{\leq} \|x^k-x^\star\|^2_{\overline{\mathbf{L}}} - 2\gamma\left\langle x^k-x^\star, \overline{\mathbf{W}}(x^k-x^\star)\right\rangle + \theta\gamma^2\left\langle x^k-x^\star, \overline{\mathbf{W}}(x^k-x^\star)\right\rangle \\
&= \|x^k-x^\star\|^2_{\overline{\mathbf{L}}} - 2\gamma\left(1-\theta\gamma/2\right)\left\|\overline{\mathbf{L}}^{\frac{1}{2}}(x^k-x^\star)\right\|^2_{\overline{\mathbf{L}}^{-\frac{1}{2}}\,\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-\frac{1}{2}}} \\
&\overset{\gamma\leq 1/\theta}{\leq} \|x^k-x^\star\|^2_{\overline{\mathbf{L}}} - \gamma\left\|\overline{\mathbf{L}}^{\frac{1}{2}}(x^k-x^\star)\right\|^2_{\overline{\mathbf{L}}^{-\frac{1}{2}}\,\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-\frac{1}{2}}} \\
&\leq \|x^k-x^\star\|^2_{\overline{\mathbf{L}}} - \gamma\lambda_{\min}\left(\overline{\mathbf{L}}^{-\frac{1}{2}}\,\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-\frac{1}{2}}\right)\left\|\overline{\mathbf{L}}^{\frac{1}{2}}(x^k-x^\star)\right\|^2 \\
&= \left(1-\gamma\lambda_{\min}\left(\overline{\mathbf{L}}^{-\frac{1}{2}}\,\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-\frac{1}{2}}\right)\right)\|x^k-x^\star\|^2_{\overline{\mathbf{L}}}.
\end{aligned}$$

After unrolling the recursion we obtain the convergence result

$$\mathbb{E}\left[\|x^{k+1}-x^\star\|^2_{\overline{\mathbf{L}}}\right] \leq \left(1-\gamma\lambda_{\min}\left(\overline{\mathbf{L}}^{-\frac{1}{2}}\,\overline{\mathbf{W}}\,\overline{\mathbf{L}}^{-\frac{1}{2}}\right)\right)^{k+1}\|x^0-x^\star\|^2_{\overline{\mathbf{L}}}.$$

□

B.3. Non-zero Solution

As a reminder, in the most general case, the problem has the form

$$f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x), \qquad f_i(x) \equiv \frac{1}{2}x^\top \mathbf{L}_i x - x^\top \mathrm{b}_i\,.$$

with the gradient estimator

$$g^k = \frac{1}{n}\sum_{i=1}^n \mathbf{C}_i^k \nabla f_i(\mathbf{C}_i^k x^k) = \frac{1}{n}\sum_{i=1}^n \mathbf{C}_i^k\left(\mathbf{L}_i\mathbf{C}_i^k x^k - \mathrm{b}_i\right) = \overline{\mathbf{B}}^k x^k - \frac{1}{n}\sum_{i=1}^n \mathbf{C}_i^k\,\mathrm{b}_i\,. \tag{36}$$

General Calculations for Estimator (16). In the heterogeneous case, the following sketch preconditioner is used

$$\tilde{\mathbf{C}}_i \coloneqq \sqrt{n/\left[\mathbf{L}_i\right]_{\pi_i,\pi_i}}\,e_{\pi_i}e_{\pi_i}^\top.$$

Then $\mathbb{E}\left[\overline{\mathbf{B}}^k\right] = \mathbf{I}$ (calculation was done as in Section B.1.1) and

$$\begin{aligned}
\mathbb{E}\left[\overline{\mathbf{C}\mathrm{b}}\right] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\tilde{\mathbf{C}}_i^k\,\mathrm{b}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\sqrt{n}[\mathbf{L}_i]^{-\frac{1}{2}}_{\pi_i,\pi_i}e_{\pi_i}e_{\pi_i}^\top\,\mathrm{b}_i\right] \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{n}\sum_{j=1}^n \sqrt{n}[\mathbf{L}_i]^{-\frac{1}{2}}_{j,j}e_j[\mathrm{b}_i]_j
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{n}\sqrt{n}\mathbf{D}_i^{-\frac{1}{2}}\,\mathrm{b}_i \\
&= \frac{1}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n}\mathbf{D}_i^{-\frac{1}{2}}\,\mathrm{b}_i \\
&= \frac{1}{\sqrt{n}}\underbrace{\overline{\mathbf{D}^{-\frac{1}{2}}\,\mathrm{b}}}_{\widetilde{\mathbf{D}\,\mathrm{b}}}
\end{aligned}
$$

B.3.1. Convergence analysis for heterogeneous case: proof of Theorem 4.1.

Here we formulate and further prove a more general version of Theorem 4.1, which is obtained as a special case of the next result for $c = 1/2$.

Theorem B.1. *Consider the method (2) with estimator (16) for a quadratic problem (10) with positive-definite matrix $\overline{\mathbf{L}} \succ 0$. Then, if for every $\mathbf{D}_i \coloneqq \mathrm{Diag}(\mathbf{L}_i)$ matrices $\mathbf{D}_i^{-\frac{1}{2}}$ exist, scaled permutation sketches $\mathbf{C}_i \coloneqq \sqrt{n}[\mathbf{L}_i^{-\frac{1}{2}}]_{\pi_i,\pi_i} e_{\pi_i} e_{\pi_i}^\top$ are used and heterogeneity is bounded as $\mathbb{E}\left[\left\|g^k - \mathbb{E}\left[g^k\right]\right\|_{\overline{\mathbf{L}}}^2\right] \leq \sigma^2$. Then, the step size is chosen as*

$$
0 < \gamma \leq \gamma_{c,\beta} \coloneqq \frac{1-c-\beta}{\beta + 1/2}, \tag{37}
$$

where $\gamma_{c,\beta} \in (0,1]$ for $\beta + c < 1$, the iterates satisfy

$$
\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|_{\overline{\mathbf{L}}^{-1}}^2\right] \leq \frac{f(x^0) - \mathbb{E}\left[f(x^K)\right]}{c\gamma K} + \left(\frac{1-\gamma}{c\beta} + \frac{\gamma}{2c}\right)\|h\|_{\overline{\mathbf{L}}}^2 + \frac{\gamma}{2c}\sigma^2. \tag{38}
$$

where $\overline{\mathbf{L}} = \frac{1}{n}\sum_{i=1}^n \mathbf{L}_i$, $h = \overline{\mathbf{L}}^{-1}\,\overline{\mathrm{b}} - \frac{1}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^n \mathbf{D}_i^{-\frac{1}{2}}\,\mathrm{b}_i$ and $\overline{\mathrm{b}} = \frac{1}{n}\sum_{i=1}^n \mathrm{b}_i$.

Proof. By using $\mathbf{L}$-smoothness

$$
\begin{aligned}
\mathbb{E}\left[f(x^{k+1}) \mid x^k\right] &\overset{(7)}{\leq} f(x^k) - \gamma\left\langle \nabla f(x^k), \mathbb{E}\left[g^k\right]\right\rangle + \frac{\gamma^2}{2}\mathbb{E}\left[\|g^k\|_{\overline{\mathbf{L}}}^2\right] \\
&\overset{(17),(25)}{=} f(x^k) - \gamma\left\langle \nabla f(x^k), \overline{\mathbf{L}}^{-1}\nabla f(x^k) + h\right\rangle \\
&\qquad + \frac{\gamma^2}{2}\left(\left\|\mathbb{E}\left[g^k\right]\right\|_{\overline{\mathbf{L}}}^2 + \mathbb{E}\left[\left\|g^k - \mathbb{E}\left[g^k\right]\right\|_{\overline{\mathbf{L}}}^2\right]\right) \\
&\overset{(17)}{=} f(x^k) - \gamma\left(\left\langle \nabla f(x^k), \overline{\mathbf{L}}^{-1}\nabla f(x^k)\right\rangle + \left\langle \nabla f(x^k), h\right\rangle\right) \\
&\qquad + \frac{\gamma^2}{2}\left(\left\|\overline{\mathbf{L}}^{-1}\nabla f(x^k) + h\right\|_{\overline{\mathbf{L}}}^2 + \mathbb{E}\left[\left\|g^k - \mathbb{E}\left[g^k\right]\right\|_{\overline{\mathbf{L}}}^2\right]\right) \\
&\overset{(23)}{=} f(x^k) - \gamma\left(\left\|\nabla f(x^k)\right\|_{\overline{\mathbf{L}}^{-1}}^2 + \left\langle \nabla f(x^k), h\right\rangle\right) + \frac{\gamma^2}{2}\mathbb{E}\left[\left\|g^k - \mathbb{E}\left[g^k\right]\right\|_{\overline{\mathbf{L}}}^2\right] \\
&\qquad + \frac{\gamma^2}{2}\left(\left\|\nabla f(x^k)\right\|_{\overline{\mathbf{L}}^{-1}}^2 + 2\left\langle \nabla f(x^k), h\right\rangle + \|h\|_{\overline{\mathbf{L}}}^2\right) \\
&\leq f(x^k) - \gamma\left(1 - \gamma/2\right)\left\|\nabla f(x^k)\right\|_{\overline{\mathbf{L}}^{-1}}^2 + \frac{\gamma^2}{2}\sigma^2 \\
&\qquad - \gamma\left(1-\gamma\right)\left\langle \nabla f(x^k), h\right\rangle + \frac{\gamma^2}{2}\|h\|_{\overline{\mathbf{L}}}^2,
\end{aligned}
$$

where the last inequality follows from the grouping of similar terms and bounded heterogeneity

$$
\mathbb{E}\left[\left\|g^k - \mathbb{E}\left[g^k\right]\right\|_{\overline{\mathbf{L}}}^2\right] = \mathbb{E}\left[\left\|g^k - \left(\overline{\mathbf{L}}^{-1}\nabla f(x^k) + h\right)\right\|_{\overline{\mathbf{L}}}^2\right] \tag{39}
$$

$$
= \mathbb{E}\left[\left\|\overline{\mathbf{B}}^k x^k - \overline{\mathbf{C}\mathrm{b}} - \left(x^k - \frac{1}{\sqrt{n}}\widetilde{\mathbf{D}\,\mathrm{b}}\right)\right\|_{\overline{\mathbf{L}}}^2\right] \leq \sigma^2. \tag{40}
$$

Next, using a Fenchel-Young inequality (27) for $\left\langle \nabla f(x^k), -h \right\rangle$ and $1-\gamma \geq 0$

$$\begin{aligned}
\mathbb{E}\left[f(x^{k+1}) \mid x^k\right] &\leq f(x^k) - \gamma\left(1-\gamma/2\right)\left\|\nabla f(x^k)\right\|^2_{\mathbf{L}^{-1}} + \frac{\gamma^2}{2}\left(\|h\|^2_{\mathbf{L}} + \sigma^2\right) \\
&\quad + \gamma\left(1-\gamma\right)\left[\beta\|\nabla f(x^k)\|^2_{\tilde{\mathbf{L}}^{-1}} + 0.25\beta^{-1}\|h\|^2_{\tilde{\mathbf{L}}}\right] \\
&\leq f(x^k) - \gamma\left(1-\gamma/2-\beta\left(1-\gamma\right)\right)\left\|\nabla f(x^k)\right\|^2_{\mathbf{L}^{-1}} \\
&\quad + \gamma\left\{\left(\beta^{-1}\left(1-\gamma\right)+\frac{\gamma}{2}\right)\|h\|^2_{\mathbf{L}} + \frac{\gamma}{2}\sigma^2\right\},
\end{aligned} \tag{41}$$

where in the last inequality we grouped similar terms and used the fact that $0.25 < 1$.

Now to guarantee that $1-\gamma/2-\beta(1-\gamma) \geq c > 0$, we choose the step size using

$$0 < \gamma \leq \gamma_{c,\beta} \coloneqq \frac{1-c-\beta}{\beta+1/2}, \tag{42}$$

where $\gamma_{c,\beta} > 0$ for $\beta + c < 1$. This means that β can not arbitrarily grow to diminish β^{-1}.
Then, after standard manipulations and unrolling the recursion

$$\gamma c\left\|\nabla f(x^k)\right\|^2_{\mathbf{L}^{-1}} \leq f(x^k) - \mathbb{E}\left[f(x^{k+1}) \mid x^k\right] + \gamma\left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\mathbf{L}} + \frac{\gamma^2}{2}\sigma^2 \tag{43}$$

we obtain

$$\frac{c}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2_{\mathbf{L}^{-1}}\right] \leq \frac{f(x^0)-\mathbb{E}\left[f(x^K)\right]}{\gamma K} + \left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\mathbf{L}} + \frac{\gamma}{2}\sigma^2. \tag{44}$$

□

B.3.2. Homogeneous case

The main difference compared to the result in the previous subsection is that the gradient estimator expression (20) holds deterministically (without expectation $\mathbb{E}$). That is why $g^k = \mathbb{E}\left[g^k\right]$ and heterogeneity term σ^2 equals to 0.

We provide the full statement and proof for the homogeneous result discussed in 4.2.

Theorem B.2. *Consider the method (2) with estimator (20) for a homogeneous quadratic problem (10) with positive-definite matrix $\mathbf{L}_i \equiv \mathbf{L} \succ 0$. Then if exists $\mathbf{D}^{-\frac{1}{2}}$ for $\mathbf{D} \coloneqq \mathrm{Diag}(\mathbf{L})$, scaled permutation sketch $\mathbf{C}_i' = \sqrt{n}e_{\pi_i}e_{\pi_i}^\top$ is used and the step size is chosen as*

$$0 < \gamma \leq \gamma_{c,\beta} \coloneqq \frac{1-c-\beta}{\beta+1/2}, \tag{45}$$

where $\gamma_{c,\beta} > 0$ for $\beta + c < 1$. Then the iterates satisfy

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2_{\tilde{\mathbf{L}}^{-1}}\right] \leq \frac{f(x^0)-\mathbb{E}\left[f(x^K)\right]}{c\gamma K} + \left(\frac{1-\gamma}{c\beta}+\frac{\gamma}{2c}\right)\|h\|^2_{\tilde{\mathbf{L}}}, \tag{46}$$

where $\tilde{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$, $h = \tilde{\mathbf{L}}^{-1}\tilde{\mathrm{b}} - \frac{1}{\sqrt{n}}\tilde{\mathrm{b}}$ and $\tilde{\mathrm{b}} = \mathbf{D}^{-\frac{1}{2}}\mathrm{b}$.

Proof. By using $\mathbf{L}$-smoothness

$$\begin{aligned}
\mathbb{E}\left[f(x^k-\gamma g^k) \mid x^k\right] &\overset{(7)}{\leq} f(x^k) - \left\langle \nabla f(x^k), \gamma\mathbb{E}\left[g^k\right]\right\rangle + \frac{\gamma^2}{2}\mathbb{E}\left[\left\|g^k\right\|^2_{\tilde{\mathbf{L}}}\right] \\
&\leq f(x^k) - \gamma\left\langle \nabla f(x^k), \tilde{\mathbf{L}}^{-1}\nabla f(x^k) + h\right\rangle + \frac{\gamma^2}{2}\left\|\tilde{\mathbf{L}}^{-1}\nabla f(x^k) + h\right\|^2_{\tilde{\mathbf{L}}} \\
&\overset{(23)}{=} f(x^k) - \gamma\left(\left\langle \nabla f(x^k), \tilde{\mathbf{L}}^{-1}\nabla f(x^k)\right\rangle + \left\langle \nabla f(x^k), h\right\rangle\right) \\
&\quad + \frac{\gamma^2}{2}\left(\left\|\nabla f(x^k)\right\|^2_{\tilde{\mathbf{L}}^{-1}} + 2\left\langle \nabla f(x^k), h\right\rangle + \|h\|^2_{\tilde{\mathbf{L}}}\right) \\
&= f(x^k) - \gamma\left(1-\gamma/2\right)\left\|\nabla f(x^k)\right\|^2_{\tilde{\mathbf{L}}^{-1}} + \frac{\gamma^2}{2}\|h\|^2_{\tilde{\mathbf{L}}} - \gamma\left(1-\gamma\right)\left\langle \nabla f(x^k), h\right\rangle
\end{aligned}$$

Next by using a Fenchel-Young inequality (27) for $\left\langle \nabla f(x^k), -h \right\rangle$ and $1-\gamma \geq 0$

$$\begin{aligned}
\mathbb{E}\left[f(x^{k+1}) \mid x^k\right] &\leq f(x^k) - \gamma\left(1-\gamma/2\right)\left\|\nabla f(x^k)\right\|^2_{\tilde{\mathbf{L}}^{-1}} + \frac{\gamma^2}{2}\|h\|^2_{\tilde{\mathbf{L}}} \\
&\quad + \gamma\left(1-\gamma\right)\left[\beta\|\nabla f(x^k)\|^2_{\tilde{\mathbf{L}}^{-1}} + 0.25\beta^{-1}\|h\|^2_{\tilde{\mathbf{L}}}\right] \\
&= f(x^k) - \gamma\left(1-\gamma/2-\beta(1-\gamma)\right)\left\|\nabla f(x^k)\right\|^2_{\tilde{\mathbf{L}}^{-1}} \\
&\quad + \gamma\left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\tilde{\mathbf{L}}}.
\end{aligned}$$

Now to guarantee that $1-\gamma/2-\beta(1-\gamma) \geq c > 0$ we choose the step size as

$$0 < \gamma \leq \gamma_{c,\beta} := \frac{1-c-\beta}{\beta+1/2}, \tag{47}$$

where $\gamma_{c,\beta} \geq 0$ for $\beta + c < 1$.
Then after standard manipulations and unrolling the recursion

$$\gamma c\left\|\nabla f(x^k)\right\|^2_{\tilde{\mathbf{L}}^{-1}} \leq f(x^k) - \mathbb{E}\left[f(x^{k+1}) \mid x^k\right] + \gamma\left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\tilde{\mathbf{L}}} \tag{48}$$

we obtain the formulated result

$$\frac{c}{K}\sum_{k=0}^{K-1}\mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2_{\tilde{\mathbf{L}}^{-1}}\right] \leq \frac{f(x^0)-\mathbb{E}\left[f(x^K)\right]}{\gamma K} + \left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\tilde{\mathbf{L}}}. \tag{49}$$

□

Remark B.3. **1)** The first term in the convergence upper bound (46) is minimized by maximizing product $c \cdot \gamma$, which motivates to choose $c > 0$ and $\gamma \leq 1$ as large as possible. Although due to the constraint on the step size (and $\beta > 0$)

$$0 < \gamma \leq \gamma_{c,\beta} := \frac{1-c-\beta}{\beta+1/2}, \tag{50}$$

constant $c \in (0,1)$. So, by maximizing c the value $\gamma_{c,\beta}$ becomes smaller, thus there is a trade-off.

2) The second term or the neighborhood size (multiplier in front of $\|h\|^2_{\tilde{\mathbf{L}}}$)

$$\Psi(\beta,\gamma) := \frac{\beta^{-1}\left(1-\gamma\right)+\gamma/2}{c} = \frac{\beta^{-1}\left(1-\gamma\right)+\gamma/2}{1-\gamma/2-\beta(1-\gamma)} \tag{51}$$

can be numerically minimized (e.g. by using WolframAlpha) with constraints $\gamma \in (0,1]$ and $\beta > 0$. The solution of such optimization problem is $\gamma^\star \approx 1$ and $\beta^\star \approx \xi \in \{3.992, 2.606, 2.613\}$. In fact, $\Psi(\beta^\star,\gamma^\star) \approx 0.5$.

Functional Gap Convergence. Note that for the quadratic optimization problem (10)

$$\left\|\nabla f(x^k)\right\|^2_{\tilde{\mathbf{L}}^{-1}} = \left\langle \tilde{\mathbf{L}}\,x^k - \tilde{\mathrm{b}}, \tilde{\mathbf{L}}^{-1}\left(\tilde{\mathbf{L}}\,x^k - \tilde{\mathrm{b}}\right)\right\rangle = 2\left(f(x^k) - f(x^\star)\right). \tag{52}$$

Then by rearranging and subtracting $f^\star := f(x^\star)$ from both sides of inequality (48) we obtain

$$\begin{aligned}
\mathbb{E}\left[f(x^{k+1}) \mid x^k\right] - f^\star &\leq f(x^k) - f^\star - \gamma c\left\|\nabla f(x^k)\right\|^2_{\tilde{\mathbf{L}}^{-1}} + \gamma\left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\tilde{\mathbf{L}}} \\
&\overset{(52)}{=} \left(f(x^k) - f^\star\right) - \gamma c \cdot 2\left(f(x^k) - f^\star\right) + \gamma\left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\tilde{\mathbf{L}}} \\
&= \left(1-2\gamma c\right)\left(f(x^k) - f^\star\right) + \gamma\left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\tilde{\mathbf{L}}}.
\end{aligned}$$

After unrolling the recursion

$$\begin{aligned}
\mathbb{E}\left[f(x^{k+1}) \mid x^k\right] - f^\star &\leq \left(1-2\gamma c\right)^k\left(f(x^0) - f^\star\right) + \gamma\left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\tilde{\mathbf{L}}}\sum_{i=0}^{k}\left(1-2\gamma c\right)^i \\
&\leq \left(1-2\gamma c\right)^k\left(f(x^0) - f^\star\right) + \frac{1}{2c}\left(\beta^{-1}\left(1-\gamma\right)+\gamma/2\right)\|h\|^2_{\tilde{\mathbf{L}}}.
\end{aligned}$$

This result is formalized in the following Theorem.

Theorem B.4. *Consider the method (2) with estimator (20) for a homogeneous quadratic problem (10) with positive-definite matrix $\mathbf{L}_i \equiv \mathbf{L} \succ 0$. Then if exists $\mathbf{D}^{-\frac{1}{2}}$ for $\mathbf{D} \coloneqq \mathrm{Diag}(\mathbf{L})$, scaled permutation sketch $\mathbf{C}_i' = \sqrt{n} e_{\pi_i} e_{\pi_i}^\top$ is used and the step size is chosen as*

$$0 < \gamma \leq \gamma_{c,\beta} \coloneqq \frac{1-c-\beta}{\beta + 1/2}, \tag{53}$$

where $\gamma_{c,\beta} > 0$ for $\beta + c < 1$. Then the iterates satisfy

$$\mathbb{E}\left[f(x^k)\right] - f^\star \leq (1-2\gamma c)^k \left(f(x^0) - f^\star\right) + \frac{1}{2c}\left(\beta^{-1}(1-\gamma) + \gamma/2\right) \|h\|_{\tilde{\mathbf{L}}}^2, \tag{54}$$

where $h = \tilde{\mathbf{L}}^{-1}\tilde{\mathrm{b}} - \frac{1}{\sqrt{n}}\tilde{\mathrm{b}}$ and $\tilde{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}, \tilde{\mathrm{b}} = \mathbf{D}^{-\frac{1}{2}}\mathrm{b}$.

This result shows that for a proper choice of the step size $\gamma = 1$ and constant $c = 1/2$, the functional gap can converge in basically one iteration to the neighborhood of size

$$\|h\|_{\tilde{\mathbf{L}}}^2 = \left\langle \tilde{\mathbf{L}}\left(\tilde{\mathbf{L}}^{-1}\tilde{\mathrm{b}} - \frac{1}{\sqrt{n}}\tilde{\mathrm{b}}\right), \tilde{\mathbf{L}}^{-1}\tilde{\mathrm{b}} - \frac{1}{\sqrt{n}}\tilde{\mathrm{b}}\right\rangle,$$

which equals zero if $\tilde{\mathbf{L}}^{-1}\tilde{\mathrm{b}} = \frac{1}{\sqrt{n}}\tilde{\mathrm{b}}$. This condition is the same as the condition we obtained at the end of Subsection 4.2 with asymptotic analysis of the iterates in the homogeneous case.

Discussion of the Trace. Consider a positive-definite $\mathbf{L} \succ 0$ such that $\exists \mathbf{D}^{-\frac{1}{2}}$. Thus $\tilde{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$ has only ones on the diagonal and $\mathrm{tr}(\tilde{\mathbf{L}}) = n$. Then

$$n \cdot \mathrm{tr}(\tilde{\mathbf{L}}^{-1}) = \mathrm{tr}(\tilde{\mathbf{L}})\mathrm{tr}(\tilde{\mathbf{L}}^{-1}) = (\lambda_1 + \cdots + \lambda_n)\left(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_n}\right) \geq n^2,$$

where the last inequality is due to the relation between harmonic and arithmetic means. Therefore $\mathrm{tr}(\tilde{\mathbf{L}}^{-1}) = \lambda_1^{-1} + \cdots + \lambda_n^{-1} \geq n$ and sum of $\tilde{\mathbf{L}}^{-1}$ eigenvalues has to be greater than n.

B.4. Generalization to $n \neq d$ Case.

Our results can be generalized in a similar way as in (Szlendak et al., 2022).

1) $d = qn$, for integer $q \geq 1$. Let $\pi = (\pi_1, \ldots, \pi_d)$ be a random permutation of $\{1, \ldots, d\}$. Then for each $i \in \{1, \ldots, n\}$ define

$$\mathbf{C}_i' \coloneqq \sqrt{n} \cdot \sum_{j=q(i-1)+1}^{qi} e_{\pi_j} e_{\pi_j}^\top. \tag{55}$$

Matrix $\mathbb{E}\left[\overline{\mathbf{B}}^k\right]$ for the homogeneous preconditioned case can be computed as follows:

$$\begin{aligned}
\mathbb{E}\left[\overline{\mathbf{B}}^k\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \mathbf{C}_i' \tilde{\mathbf{L}} \mathbf{C}_i'\right] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\sum_{j=q(i-1)+1}^{qi} n e_{\pi_j} e_{\pi_j}^\top \tilde{\mathbf{L}} e_{\pi_j} e_{\pi_j}^\top\right] \\
&= \sum_{i=1}^n \sum_{j=q(i-1)+1}^{qi} \mathbb{E}\left[e_{\pi_j} e_{\pi_j}^\top \tilde{\mathbf{L}} e_{\pi_j} e_{\pi_j}^\top\right] \\
&= \sum_{i=1}^n \sum_{j=q(i-1)+1}^{qi} \frac{1}{d}\sum_{l=1}^d e_l e_l^\top \tilde{\mathbf{L}} e_l e_l^\top \\
&= \sum_{i=1}^n \sum_{j=q(i-1)+1}^{qi} \frac{1}{d}\mathrm{Diag}(\tilde{\mathbf{L}})
\end{aligned}$$

$$
\begin{aligned}
&= n\frac{q}{d}\mathrm{Diag}(\tilde{\mathbf{L}}) \\
&= \mathrm{Diag}(\tilde{\mathbf{L}}) \\
&= \mathbf{I}.
\end{aligned}
$$

As for the linear term

$$
\begin{aligned}
\mathbb{E}\left[\mathbf{C}'\,\mathrm{b}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{C}'_i\,\tilde{\mathrm{b}}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sum_{j=q(i-1)+1}^{qi}\sqrt{n}e_{\pi_j}e_{\pi_j}^{\top}\,\tilde{\mathrm{b}}\right] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=q(i-1)+1}^{qi}\frac{1}{d}\mathbf{I}\,\tilde{\mathrm{b}} = \frac{\sqrt{n}q}{d}\mathbf{I}\,\tilde{\mathrm{b}} = \frac{1}{\sqrt{n}}\,\tilde{\mathrm{b}}\,.
\end{aligned}
$$

2) $n = qd$, for integer $q \geq 1$. Define the multiset $S := \{1,\ldots,1,2,\ldots,2,\ldots,d,\ldots,d\}$, where each number occurs precisely q times. Let $\pi = (\pi_1,\ldots,\pi_n)$ be a random permutation of S. Then for each $i \in \{1,\ldots,n\}$ define

$$
\mathbf{C}'_i := \sqrt{d}\cdot e_{\pi_i}e_{\pi_i}^{\top}. \tag{56}
$$

$$
\begin{aligned}
\mathbb{E}\left[\overline{\mathbf{B}}^k\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{C}'_i\,\tilde{\mathbf{L}}\,\mathbf{C}'_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[de_{\pi_i}e_{\pi_i}^{\top}\,\tilde{\mathbf{L}}\,e_{\pi_i}e_{\pi_i}^{\top}\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{d}\sum_{j=1}^{d}de_je_j^{\top}\,\tilde{\mathbf{L}}\,e_je_j^{\top} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Diag}(\tilde{\mathbf{L}}) = \mathbf{I}.
\end{aligned}
$$

The linear term

$$
\mathbb{E}\left[\mathbf{C}'\,\mathrm{b}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{C}'_i\,\tilde{\mathrm{b}}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\sqrt{d}e_{\pi_i}e_{\pi_i}^{\top}\,\tilde{\mathrm{b}}\right] = \frac{\sqrt{d}}{n}\sum_{i=1}^{n}\frac{1}{d}\mathbf{I}\,\tilde{\mathrm{b}} = \frac{1}{\sqrt{d}}\,\tilde{\mathrm{b}}\,.
$$

To sum up both cases, in a homogeneous preconditioned setting $\mathbb{E}\left[\overline{\mathbf{B}}^k\right] = \mathbf{I}$ and

$$
\mathbb{E}\left[\mathbf{C}'\,\mathrm{b}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{C}'_i\,\mathrm{b}\right] = \tilde{\mathrm{b}}\,/\sqrt{\min(n,d)}.
$$

Similar modifications and calculations can be performed for heterogeneous scenarios. The case when n does not divide d and vice versa is generalized using constructions from (Szlendak et al., 2022).

C. Generalization Beyond Quadratics

In this section our analysis is revisited for the more general case of smooth (non-convex) functions and class of compressors. While this result may not be as nuanced as for the quadratic model, it still does not require restrictive assumptions on gradient estimator, unlike prior works.

C.1. Preliminary Facts

Our convergence analysis relies on the following Lemma.

Lemma C.1 (Descent Lemma (Li et al., 2021))**.** *Suppose that function f is L-smooth (7) and let $x^{k+1} := x^k - \gamma g^k$. Then for any $g^k \in \mathbb{R}^d$ and $\gamma > 0$, we have*

$$
f(x^{k+1}) \leq f(x^k) - \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\|x^{k+1} - x^k\|^2 + \frac{\gamma}{2}\left\|g^k - \nabla f(x^k)\right\|^2, \tag{57}
$$

If a continuously differentiable function f is L-smooth (7), then for any $x, y \in \mathbb{R}^d$, it is satisfied

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{58}$$

Lemma C.2 (Khaled & Richtárik (2023))**.** *Let f be L-smooth (7) and $f^{\inf}$-lower bounded. Then for any $x \in \mathbb{R}^d$ we have:*

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f^{\inf}) \tag{59}$$

We also use introduced by Szlendak et al. (2022) notion of AB inequality for a collection of (correlated) compressors $\mathcal{C}_i$.

Definition C.3 (AB inequality)**.** A collection of random unbiased operators $\mathcal{C}_1, \ldots, \mathcal{C}_n : \mathbb{R}^d \to \mathbb{R}^d$ satisfies AB inequality if there exist constants $A, B \geq 0$ such that

$$\mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{C}_i(x_i) - \frac{1}{n}\sum_{i=1}^n x_i\right\|^2 \leq A\frac{1}{n}\sum_{i=1}^n \|x_i\|^2 - B\left\|\frac{1}{n}\sum_{i=1}^n x_i\right\|^2 \tag{60}$$

for all $x_1, \ldots, x_n \in \mathbb{R}^d$. For brevity: $\{\mathcal{C}_i\}_{i=1}^n \in \mathbb{C}(A, B)$.

This definition is handy for analyzing algorithms like IST as it allows to generalize analysis for the case when compressors $\mathcal{C}_1, \ldots, \mathcal{C}_n$ are dependent, which happens when the parameters of the model are randomly decomposed in a non-overlapping fashion. Moreover, AB inequality is satisfied for unbiased compressors (2.1): $\mathcal{C}_i \in \mathbb{U}(\omega_i)$ for all i.

Denote $w_i^k := \mathcal{C}_i^k(x^k)$. If compressors $\mathcal{C}_i^k$ satisfy the *perfect reconstruction* property $\frac{1}{n}\sum_{i=1}^n \mathcal{C}_i^k(x) = x$, which makes sense for the IST formulation, then method (5) can be reformulated in the following way

$$\begin{aligned} x^{k+1} &= \frac{1}{n}\sum_{i=1}^n \left(w_i^k - \gamma \mathcal{Q}_i^k\left(\nabla f_i(w_i^k)\right)\right) \\ &= \frac{1}{n}\sum_{i=1}^n \mathcal{C}_i^k(x^k) - \gamma\frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i^k\left(\nabla f_i(w_i^k)\right)) \\ &= x^k - \gamma \underbrace{\frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i^k\left(\nabla f_i(\mathcal{C}_i^k(x^k))\right)}_{g^k}. \end{aligned} \tag{61}$$

This reformulation makes the algorithm amenable to analysis using Lemma C.1.

C.2. Convergence Analysis

Theorem C.4. *Let f and f_i be L and L_i-smooth (7) respectively. Collection of unbiased compressors $\mathcal{C}_i^k, \mathcal{Q}_i^k$ satisfy AB-inequality (60) and property $\frac{1}{n}\sum_{i=1}^n \mathcal{C}_i^k(x) = x$ for every k. Then for step size chosen as*

$$\gamma \leq \min\left\{\frac{1}{L}, \frac{\sqrt{1 + {}^{2}/_{(AL_m K)}} - 1}{2}\right\}$$

the iterations of Algorithm (61) for any $K \geq 1$ satisfy

$$\min_{0\leq k\leq K-1} \left\|\nabla f(x^k)\right\|^2 \leq \frac{6\left(f(x^0) - f^{\inf}\right)}{\gamma K} + 2\left(1 + (1+\beta^{-1})A\right)\max_k \|x^k\|^2\overline{L_\omega^2} + 4AL_{\max}(1+\beta)\Delta, \tag{62}$$

where $\beta \leq \gamma$, $\overline{L_\omega^2} = \frac{1}{n}\sum_{i=1}^n L_i^2\omega_i$, $L_{\max} = \max_i L_i$, and $\Delta = \frac{1}{n}\sum_{i=1}^n \left(f^{\inf} - f_i^{\inf}\right)$.

Note that minimum over squared gradient norm $\left\|\nabla f(x^k)\right\|^2$ and maximum over model weights $\|x^k\|^2$ in (62) can be replaced with with weighted sums due to proof step (68). Let us contrast the obtained convergence result with what we have for the quadratic case in (19). The first term is basically the same and decreases as $1/K$ for constant step size γ. The other terms are different as they involve the norm of the iterates $\|x^k\|^2$ throughout training. However, they also can not be

decreased by diminishing the step size, which is similar to the quadratic case. Thus, we showed that for the more general case of smooth losses, IST can converge to the irreducible neighborhood of the stationary point. The obtained conclusions highlight the generalizability of our insights beyond the quadratic model.

Moreover, there is an additional part representing the heterogeneity of the distributed problem proportional to Δ. In contrast to the term involving $\|x^k\|^2$ it can be eliminated in the homogeneous case when functions have a shared "minimizer" $f_i^{\inf} = f^{\inf}$, which may hold in the overparametrized regime.

Contrasting to prior works. Khaled & Richtárik (2019) analyzed a similar method with unbiased compression in the single node and strongly convex setting. Yuan et al. (2022) extended their results to a finite-sum case with random sampling of one client. However, prior works suffer from a very strong condition on the sparsification variance $\omega \lesssim \mu/L$, which is not the case for our analysis. Another important difference is that convergence bounds in (Khaled & Richtárik, 2019) and (Yuan et al., 2022) depend on the norm of the optimal solution $x^\star = \arg\min f(x)$, which may not exist in the non-convex setting.

Proof. We start with the Descent Lemma (57)

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) - \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 - \left(\frac{1}{2\gamma} - \frac{L}{2}\right)\|x^{k+1} - x^k\|^2 + \frac{\gamma}{2}\left\|g^k - \nabla f(x^k)\right\|^2 \\ &\leq f(x^k) - \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 + \frac{\gamma}{2}\left\|g^k - \nabla f(x^k)\right\|^2, \end{aligned} \tag{63}$$

where the step size is chosen as $\gamma \leq 1/L$.

Next prove an auxiliary result needed to work with the last term of (63)

$$\begin{aligned} \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathcal{C}_i^k(x^k)) - \nabla f_i(x^k)\right\|^2 &\overset{(26)}{\leq} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\|\nabla f_i(\mathcal{C}_i^k(x^k)) - \nabla f_i(x^k)\right\|^2 \\ &\overset{(58)}{\leq} \frac{1}{n}\sum_{i=1}^n L_i^2\,\mathbf{E}\left\|\mathcal{C}_i^k(x^k) - x^k\right\|^2 \\ &\overset{(3)}{\leq} \frac{1}{n}\sum_{i=1}^n L_i^2\omega_i\|x^k\|^2 \\ &= \overline{L_\omega^2}\|x^k\|^2, \end{aligned} \tag{64}$$

for $\overline{L_\omega^2} := \frac{1}{n}\sum_{i=1}^n L_i^2\omega_i$.

Recall the expression for $g^k = \frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i^k\left(\nabla f_i(w_i^k)\right)$ and $w_i^k := \mathcal{C}_i^k(x^k)$. Now we can upper bound the last term of (63)

$$\begin{aligned} \mathbf{E}\left\|g^k - \nabla f(x^k)\right\|^2 &= \mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i^k\left(\nabla f_i(w_i^k)\right) - \nabla f_i(x^k)\right\|^2 \\ &\overset{(24)}{\leq} 2\,\mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n \mathcal{Q}_i^k\left(\nabla f_i(w_i^k)\right) - \nabla f_i(w_i^k)\right\|^2 + 2\,\mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(w_i^k) - \nabla f_i(x^k)\right\|^2 \\ &\overset{(60)}{\leq} 2\left[A\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\|\nabla f_i(w_i^k)\right\|^2 - B\,\mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(w_i^k)\right\|^2\right] \\ &\quad + 2\,\mathbf{E}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathcal{C}_i^k(x^k)) - \nabla f_i(x^k)\right\|^2 \\ &\overset{(64)}{\leq} 2A\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left\|\nabla f_i(\mathcal{C}_i^k(x^k)) \pm \nabla f_i(x^k)\right\|^2 + 2\frac{1}{n}\sum_{i=1}^n L_i^2\omega_i\|x^k\|^2 \\ &\overset{(24)}{\leq} 2A\frac{1}{n}\sum_{i=1}^n \left[\left(1+\beta^{-1}\right)\mathbf{E}\left\|\nabla f_i(\mathcal{C}_i^k(x^k)) - \nabla f_i(x^k)\right\|^2 + (1+\beta)\left\|\nabla f_i(x^k)\right\|^2\right] + 2\overline{L_\omega^2}\|x^k\|^2 \end{aligned}$$

$$
\begin{aligned}
&\overset{(64)}{\leq} \quad 2A\frac{1}{n}\sum_{i=1}^{n}\left[\left(1+\beta^{-1}\right)L_i^2\omega_i\|x^k\|^2+\left(1+\beta\right)\left\|\nabla f_i(x^k)\right\|^2\right]+2\overline{L_\omega^2}\|x^k\|^2 \\
&\overset{(59)}{\leq} \quad 2\left(1+A\left(1+\beta^{-1}\right)\right)\overline{L_\omega^2}\|x^k\|^2+2A\left(1+\beta\right)\frac{1}{n}\sum_{i=1}^{n}2L_i\left[f_i(x^k)-f_i^{\inf}\right] \\
&= \quad 2\left(1+A\left(1+\beta^{-1}\right)\right)\overline{L_\omega^2}\|x^k\|^2+4A\left(1+\beta\right)\frac{1}{n}\sum_{i=1}^{n}L_i\left[f_i(x^k)-f^{\inf}-f_i^{\inf}+f^{\inf}\right] \\
&\leq \quad 2\left(1+A\left(1+\beta^{-1}\right)\right)\overline{L_\omega^2}\|x^k\|^2+4A\left(1+\beta\right)L_{\max}\frac{1}{n}\sum_{i=1}^{n}\left[f_i(x^k)-f^{\inf}+f^{\inf}-f_i^{\inf}\right] \\
&= \quad 2\left(1+A\left(1+\beta^{-1}\right)\right)\overline{L_\omega^2}\|x^k\|^2+4A\left(1+\beta\right)L_m\Big[f(x^k)-f^{\inf}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(f^{\inf}-f_i^{\inf}\right)}_{\Delta}\Big],
\end{aligned}
$$

where $L_m = L_{\max} = \max L_i$.

Combined with (63) and denoting $\delta^k \coloneqq f(x^k) - f^{\inf}$ it leads to

$$
\begin{aligned}
\delta^{k+1} &\leq \delta^k - \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 + 2\gamma A\left(1+\beta\right)L_m\left(f(x^k)-f^{\inf}\right) \\
&\quad + \frac{\gamma}{2}\left[2\left(1+A\left(1+\beta^{-1}\right)\right)\overline{L_\omega^2}\|x^k\|^2+4A\left(1+\beta\right)L_m\Delta\right] \\
&= \left(1+2\gamma AL_m\left(1+\beta\right)\right)\delta^k - \frac{\gamma}{2}\left\|\nabla f(x^k)\right\|^2 \\
&\quad + \frac{\gamma}{2}\underbrace{\left[2\left(1+A\left(1+\beta^{-1}\right)\right)\overline{L_\omega^2}\|x^k\|^2+4A\left(1+\beta\right)L_m\Delta\right]}_{C^k}.
\end{aligned} \tag{65}
$$

After rearranging we obtain

$$
\left\|\nabla f(x^k)\right\|^2 \leq \frac{2}{\gamma}\big(1+\underbrace{2\gamma A\left(1+\beta\right)L_m}_{D}\big)\delta^k - \frac{2}{\gamma}\delta^{k+1} + C^k. \tag{66}
$$

Next by following technique by Stich (2019) we introduce an exponentially decaying weighting sequence

$$
w^k = \frac{w^{k-1}}{1+D} \leq \cdots \leq w^{-1}.
$$

Multiplying recursion (66) by w^k, we get

$$
\begin{aligned}
w^k\left\|\nabla f(x^k)\right\|^2 &\leq \frac{2w^k\left(1+D\right)}{\gamma}\delta^k - \frac{2w^k}{\gamma}\delta^{k+1} + w^kC^k \\
&= \frac{2w^{k-1}}{\gamma}\delta^k - \frac{2w^k}{\gamma}\delta^{k+1} + w^kC^k
\end{aligned}
$$

After summing up both sides for k from 0 to $K-1$, and telescoping terms

$$
\sum_{k=0}^{K-1} w^k\left\|\nabla f(x^k)\right\|^2 \leq \frac{2w^{-1}}{\gamma}\delta^0 - \frac{2w^{K-1}}{\gamma}\delta^K + \sum_{k=0}^{K-1} w^kC^k. \tag{67}
$$

Now define $W^K \coloneqq \sum_{k=0}^{K-1} w^k$ and divide both sides of (67) by W^K

$$
\min_{0\leq k\leq K-1}\left\|\nabla f(x^k)\right\|^2 \leq \frac{1}{W^k}\sum_{k=0}^{K-1} w^k\left\|\nabla f(x^k)\right\|^2 \leq \frac{2w^{-1}}{W^k\gamma}\delta^0 - \frac{2w^{K-1}}{W^k\gamma}\delta^K + \frac{1}{W^k}\sum_{k=0}^{K-1} w^kC^k. \tag{68}
$$

By using the fact that

$$W^k = \sum_{k=0}^{K-1} w^k \geq \sum_{k=0}^{K-1} \min_{0\leq i\leq K-1} w^i = Kw^{K-1} = \frac{Kw^{-1}}{(1+D)^K},$$

we obtain

$$\min_k \left\|\nabla f(x^k)\right\|^2 \leq 2\frac{\left(1+2\gamma A\left(1+\beta\right)L_m\right)^K}{\gamma K}\delta^0 + \max_k C^k. \tag{69}$$

Next to simplify the obtained upper bound we use the fact that $1+x \leq \exp(x)$

$$\left(1+2\gamma A\left(1+\beta\right)L_m\right)^K \leq \left(\exp\left(2\gamma A\left(1+\beta\right)L_m\right)\right)^K = \exp\left(2\gamma A\left(1+\beta\right)L_m K\right) \leq \exp(1) \leq 3,$$

where the second inequality for $\beta \leq \gamma$ holds if

$$\underbrace{2AL_mK}_{A'_K}\gamma\left(1+\gamma\right) \leq 1.$$

Then condition $A'_K\gamma + A'_K\gamma^2 - 1 \leq 0$ holds for

$$\gamma \leq \frac{\sqrt{1+4/A'_K}-1}{2} = \frac{\sqrt{1+{}^2/(AL_mK)}-1}{2}.$$

As a result (69) leads to

$$\min_k \left\|\nabla f(x^k)\right\|^2 \leq \frac{6\left(f(x^0)-f^{\inf}\right)}{\gamma K} + 2\left(1+A\left(1+\beta^{-1}\right)\right)\overline{L_\omega^2}\max_k\|x^k\|^2 + 4A\left(1+\beta\right)L_m\Delta,$$

where $\beta \leq \gamma$, $\Delta = \frac{1}{n}\sum_{i=1}^n\left(f^{\inf}-f_i^{\inf}\right)$, and $\overline{L_\omega^2} = \frac{1}{n}\sum_{i=1}^n L_i^2\omega_i$.

□

D. Comparison to Related Works

Overview of theory provided in the original IST work (Yuan et al., 2022). The authors consider the following method

$$x^{k+1} = \mathcal{C}(x^k) - \gamma\nabla f_{i_k}(\mathcal{C}(x^k)), \tag{70}$$

where $[\mathcal{C}(x)]_i = x_i \cdot \mathcal{B}e(p)$[4] is a Bernoulli sparsifier and i_k is sampled uniformly at random from $[n]$.

The analysis in (Yuan et al., 2022) relies on the assumptions

1. L_i-smoothness of individual losses f_i;
2. Q-Lipschitz continuity of f: $|f(x)-f(y)| \leq Q\|x-y\|$;
3. Error bound (or PŁ-condition): $\|\nabla f(x)\| \geq \mu\|x^\star - x\|$, where $x^\star$ is the global optimum;
4. Stochastic gradient variance: $\mathbb{E}\left[\|\nabla f_{i_k}(x)\|^2\right] \leq M + M_f\|\nabla f(x)\|^2$;
5. $\mathbb{E}\left[\nabla f_{i_k}(\mathcal{C}(x^k))\,|\,x^k\right] = \nabla f(x^k) + \varepsilon, \quad \|\varepsilon\| \leq B.$

Convergence result from Theorem 1 (Yuan et al., 2022) for step size $\gamma = {}^1/(2L_{\max})$:

$$\min_{k\in\{1,\ldots,K\}} \mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \frac{f(x^0)-f(x^\star)}{\alpha(K+1)} + \frac{1}{\alpha}\cdot\left(\frac{BQ}{2L_{\max}} + \frac{5L_{\max}\omega}{2}\|x^\star\|^2 + \frac{M}{4L_{\max}}\right), \tag{71}$$

[4] $\mathcal{B}_p(x) := \begin{cases} x/p & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$

where $\alpha := \frac{1}{2L_{\max}}\left(1 - \frac{M_f}{2}\right) - \frac{5\omega L_{\max}}{2\mu^2}$, $\omega := \frac{1}{p} - 1 < \frac{\mu^2}{10L_{\max}^2}$, and $L_{\max} := \max_i L_i$.

If Lipschitzness and Assumption 5 are replaced with *norm condition*:

$$\|\mathbb{E}\left[\nabla f_{i_k}(\mathcal{C}(x^k)) \,|\, x^k\right] - \nabla f(x^k)\| \leq \theta \|\nabla f(x^k)\| \tag{72}$$

they obtain the following (for step size $\gamma = 1/2L_{\max}$)

$$\min_{k\in\{1,\ldots,K\}} \mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2\right] \leq \frac{f(x^0) - f(x^\star)}{\alpha(K+1)} + \frac{1}{\alpha}\cdot\left(\frac{5L_{\max}\omega}{2}\|x^\star\|^2 + \frac{M}{4L_{\max}}\right), \tag{73}$$

where $\alpha = \frac{1}{2L_{\max}}\left(\frac{1}{2} - \theta - \frac{M_f}{2}\right) - \frac{5\omega L_{\max}}{2\mu^2}$ and $\omega = \frac{1}{p} - 1 < \frac{\mu^2}{5L_{\max}^2\left(\frac{1}{2}-\theta-\frac{M_f}{2}\right)}$.

Remark D.1. The original method (70) does not incorporate gradient sparsification, which can create a significant disparity between theory and practice. This is because the gradient computed at the compressed model, denoted as $\nabla f(\mathcal{C}(x))$, is not guaranteed to be sparse and representative of the submodel computations. Such modification of the method also significantly simplifies theoretical analysis, as using a single sketch (instead of **CLC**) allows for an unbiased gradient estimator.

Through our analysis of the IST gradient estimator in Equation (20), we discover that conditions—such as Assumption 5 and Inequality (72)—are not satisfied, even in the homogeneous setting for a simple quadratic problem. Furthermore, it is evident that such conditions are also not met for logistic loss. At the same time, in general, it is expected that insightful theory for general (non-)convex functions should yield appropriate results for quadratic problems. Additionally, it remains unclear whether the norm condition (72) is satisfied in practical scenarios. The situation is not straightforward—even for quadratic problems—as we show in the expression for σ^2 in Equation (39).

Masked Training (Mohtashami et al., 2022). The authors consider the following "Partial SGD" method

$$\begin{aligned} \hat{x}^k &= x^k + \delta x^k = x^k - (1-p)\odot x^k \\ x^{k+1} &= x^k - \gamma p \odot \nabla f(\hat{x}^k, \xi^k), \end{aligned} \tag{74}$$

where $\nabla f(x,\xi)$ is an unbiased stochastic gradient estimator of a L-smooth loss function f, $\odot$ is an element-wise product, and p is a binary sparsification mask.

Mohtashami et al. (2022) make the following "bounded perturbation" assumption

$$\max_k \frac{\|\delta x^k\|}{\max\left\{\|p^k \odot \nabla f(x^k)\|, \|p^k \odot \nabla f(\hat{x}^k)\|\right\}} \leq \frac{1}{2L}. \tag{75}$$

This inequality may not hold for a simple convex case. Consider a function $f(x) = \frac{1}{2}x^\top Ax$, for

$$A = \begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix}, \qquad x^0 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad p^0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{76}$$

Then condition (75) (at iteration $k = 0$) will be equivalent to

$$\frac{x_1}{cx_2} \leq \frac{1}{2a} \Leftrightarrow 2 \leq \frac{2a}{c} \leq \frac{x_2}{x_1},$$

which clearly does not hold for an arbitrary initialization x^0.

In addition, the convergence bound in Theorem 1 (Mohtashami et al., 2022) suggests choosing the step size as $\gamma_0\alpha^k$, where

$$\alpha^k = \min\left\{1, \frac{\left\langle p^k \odot \nabla f(x^k), p^k \odot \nabla f(\hat{x}^k)\right\rangle}{\|p^k \odot \nabla f(\hat{x}^k)\|^2}\right\} \tag{77}$$

is not guaranteed to be positive to the inner product $\left\langle p^k \odot \nabla f(x^k), p^k \odot \nabla f(\hat{x}^k)\right\rangle$, which may lead to non-convergence of the method.

Optimization with access to auxiliary information framework (Chayti & Karimireddy, 2024) suggests modeling training with compressed models via performing gradient steps with respect to function $h(x) \coloneqq \mathbb{E}_{\mathcal{M}}\left[f(1_{\mathcal{M}} \odot x)\right]$. This function allows access to a sparse version of the original model $f(x)$. They impose the following bounded Hessian dissimilarity assumption on h and f

$$\left\|\nabla^2 f(x) - \mathbb{E}_{\mathcal{M}}\left[\mathbf{D}_{\mathcal{M}} \nabla^2 f(1_{\mathcal{M}} \odot x) \mathbf{D}_{\mathcal{M}}\right]\right\|_2 \leq \delta, \tag{78}$$

where $1_{\mathcal{M}}$ and $\mathbf{D}_{\mathcal{M}} = \mathrm{Diag}(1_{\mathcal{M}})$ refer to a binary vector and matrix sparsification masks.

This approach relies on variance-reduction and requires gradient computations on the full model x, and thus it is not suitable for our problem setting.

Comparison to the work of Liao & Kyrillidis (2022). Next, we try our best to briefly and accurately represent some of the previous work's findings and comment on the differences.

The authors provide a *high probability* convergence analysis of a "Single Hidden-Layer Neural Network with ReLU activations" based on the Neural Tangent Kernel (NTK) framework. The network's first layer weights are initialized based on $\mathcal{N}(0, \kappa^2 \mathbf{I})$ and the weight vector of the second layer is initialized uniformly at random from $\{-1, 1\}$. In contrast, we do not make any assumptions on the initialized parameters x (in our notation).

The second differentiation is assumptions on the data. Liao & Kyrillidis (2022) assume that for every data point (a_j, y_j), it holds that $||a_j||^2 = 1$ and $|y_j| \leq C - 1$ for some constant $C \geq 1$. Moreover, for any $j \neq l$, it holds that the points a_i, a_l are not co-aligned, i.e., $a_i \neq \xi a_l$ for any $\xi \in \mathbb{R}$. In contrast, we do not make any assumptions about the data apart from the ones on matrices $\mathbf{L}_i$. In addition, analysis by Liao & Kyrillidis (2022) assumes that the number of hidden nodes is greater than a certain quantity and that NN's weights distance from initialization is uniformly bounded.

Liao & Kyrillidis (2022) consider a regression (MSE) loss function, a special case of quadratic loss and full gradients computation. They provide guarantees for IST under a "simplified assumption that every worker has full data access", which corresponds to the homogeneous setting in our terminology.

Comparison of IST and 3D Parallelism (Shoeybi et al., 2019). IST and 3D parallelism were introduced independently and concurrently in 2019. While sharing some conceptual similarities (combination of data and model parallelism), they have a different focus. The key distinction is the way how model parallelism is implemented. Namely, 3D parallelism combines *Pipeline* and *Tensor* parallelism.

Pipeline parallelism suggests splitting the model's layers across computing nodes which requires additional transmission of tensors for every forward and backward step. Tensor parallelism breaks the layers (MLP blocks and attention heads) into parts, which creates a need for additional synchronization between GPUs. Thus, 3D parallelism makes computation on nodes dependent on mutual communications, unlike IST.

Independent Subnetwork Training decomposes the model into smaller subnets independently trained in parallel. IST does not need synchronization during local updates and requires the transmission of fewer parameters, which decreases per-step communication costs.

3D parallelism introduces significant communication overhead, which greatly increases computational cost (Bian et al., 2023). The problem described is especially relevant for public compute clouds (such as Amazon EC2), which often suffer from slow interconnects. At the same time, IST is most beneficial for such a setup as it improves communication efficiency by design. Moreover, 3D parallelism is incompatible with a standard federated learning setting. At the same time, the IST-like approach is a viable technique, as every network is independent and can be trained on resource-constrained devices (Dun et al., 2023). In summary, both IST and 3D parallelism are viable approaches with pros and cons and are best suited for different scenarios.