Sharpened Lazy Incremental Quasi-Newton Method

Aakash Sunil Lahoti[*]
Carnegie Mellon University

Spandan Senapati[*]
University of Southern California

Ketan Rajawat
Indian Institute of Technology Kanpur

Alec Koppel
J.P. Morgan AI Research

Abstract

The problem of minimizing the sum of n functions in d dimensions is ubiquitous in machine learning and statistics. In many applications where the number of observations n is large, it is necessary to use incremental or stochastic methods, as their per-iteration cost is independent of n. Of these, Quasi-Newton (QN) methods strike a balance between the per-iteration cost and the convergence rate. Specifically, they exhibit a superlinear rate with $\mathcal{O}(d^2)$ cost in contrast to the linear rate of first-order methods with $\mathcal{O}(d)$ cost and the quadratic rate of second-order methods with $\mathcal{O}(d^3)$ cost. However, existing incremental methods have notable shortcomings: Incremental Quasi-Newton (IQN) only exhibits asymptotic superlinear convergence. In contrast, Incremental Greedy BFGS (IGS) offers explicit superlinear convergence but suffers from poor empirical performance and has a per-iteration cost of $\mathcal{O}(d^3)$. To address these issues, we introduce the Sharpened Lazy Incremental Quasi-Newton Method (SLIQN) that achieves the best of both worlds: an explicit superlinear convergence rate, and superior empirical performance at a per-iteration $\mathcal{O}(d^2)$ cost. SLIQN features two key changes: first, it incorporates a hybrid strategy of using both classic and greedy BFGS updates, allowing it to empirically outperform both IQN and IGS. Second, it employs a clever constant multiplicative factor along with a lazy propagation strategy, which enables it to have a cost of $\mathcal{O}(d^2)$. Additionally, our experiments demonstrate the superiority of SLIQN over other incremental and stochastic Quasi-Newton variants and establish its competitiveness with second-order incremental methods.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

We consider the finite sum minimization problem,

$$\boldsymbol{x}^\star = \arg\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}), \qquad (\mathcal{P})$$

where each f_i is μ-strongly convex, L-smooth and has a Lipschitz continuous Hessian. The canonical example of $(\mathcal{P})$ is the empirical risk minimization problem in supervised learning, where $\boldsymbol{x}$ denotes model parameters, n is the number of training samples, and f_i denotes the loss incurred by the i^{th} sample. Other instances of this problem arise in maximum likelihood estimation (Li et al. (2021, 2020)), control theory (Wu et al. (2018)), and unsupervised learning (Song and Ermon (2020)). In many applications, $(\mathcal{P})$ is both high-dimensional (large d) and data-intensive (large n).

When n is large, it becomes infeasible to process the entire dataset at every iteration, thus making classical algorithms such as gradient descent or Newton's method impractical. Consequently, stochastic and incremental variants of these algorithms have been widely adopted for such problems, because their per-iteration complexity is independent of n. While first-order methods like stochastic gradient descent (SGD) enjoy a low per-iteration complexity of $\mathcal{O}(d)$, their convergence rates, even with enhancements like variance reduction or acceleration (Defazio et al. (2014); Johnson and Zhang (2013)), remain linear at best. In contrast, second-order methods like Newton Incremental Method (NIM) by Rodomanov and Kropotov (2016), achieve a superlinear rate but at a $\mathcal{O}(d^3)$ cost, which is prohibitively large for high-dimensional problem settings.

[*]Equal Contribution

Stochastic and incremental Quasi-Newton (QN) methods strike a balance between the computational efficiency of SGD and the fast convergence rate of NIM. Specifically, the Incremental Quasi-Newton (IQN) method from Mokhtari et al. (2018), was the first QN method to achieve a superlinear convergence rate with a per-iteration complexity of $\mathcal{O}(d^2)$. However, the analysis presented in Mokhtari et al. (2018) was asymptotic and did not include an explicit rate of convergence. Typically, explicit rates are preferred over asymptotic ones as they enable a more fine-grained comparison among algorithms. For instance, with $\rho \in (0, 1)$, both $\mathcal{O}(\rho^{t^2})$ and $\mathcal{O}(\rho^{t\ln(t)})$ qualify as superlinear rates. However, the former, $\mathcal{O}(\rho^{t^2})$, is faster than the latter, $\mathcal{O}(\rho^{t\ln(t)})$. Furthermore, the mathematical expression of ρ helps determine the rate's dependence on problem parameters like condition number and dimension.

The Incremental Greedy BFGS (IGS) method by Gao et al. (2020) aimed to address this issue by incorporating the greedy updates, first introduced in Rodomanov and Nesterov (2021a), into the IQN framework. While IGS achieves an explicit superlinear convergence rate of $\mathcal{O}(e^{-t^2/n^2})$, it suffers from several major drawbacks: First, like NIM, it has a large per-iteration cost of $\mathcal{O}(d^3)$. This stems from IGS's Hessian updates not being low ran, which precludes an efficient evaluation of the Hessian inverse. Since this complexity mirrors that of NIM, it undermines the computational advantages of incorporating QN updates. Second, the empirical performance of IGS was not studied in Gao et al. (2020), and our experiments (ref. Figure 2) indicate that on multiple datasets, IGS severely underperforms compared to IQN. Finally, IGS lacked theoretical analysis to support its lemmas and theorems. In this light, we ask the following question:

Can we devise an incremental QN method with a per-iteration complexity of $\mathcal{O}(d^2)$, achieving the best-known incremental convergence rate of $\mathcal{O}(e^{-t^2/n^2})$, and demonstrating superior empirical performance?

We put forth the Sharpened Lazy IQN (SLIQN) method that meets all these objectives. SLIQN is inspired by the recent work of Jin et al. (2022), which showcased the superior performance of sharpened updates over greedy updates in the *non-incremental setting*. We first show that a direct incorporation of sharpened updates into the IQN framework does not work because the Hessian update matrices corresponding to sharpened updates are not low rank, and therefore the resulting Sharpened IQN (SIQN) method incurs a per-iteration cost of $\mathcal{O}(d^3)$. We then propose our novel Sharpened Lazy IQN (SLIQN) algorithm that overcomes this limitation by modifying the updates of SIQN using a clever constant multiplicative factor and incorporating a lazy propagation strategy. The resulting algorithm incurs a per-iteration complexity of $\mathcal{O}(d^2)$ and achieves a convergence rate of $\mathcal{O}(\rho^{t^2/n^2})$, where $\rho := 1 - \mu/dL$, which is the best-known rate in the incremental setting. We also establish an explicit linear rate of convergence of the Hessian approximation to the true Hessian. Moreover, in contrast to IGS, we provide a comprehensive theoretical analysis. Furthermore, we demonstrate the superior empirical performance of SLIQN as compared to IQN, IGS, and other state-of-the-art incremental and stochastic QN methods.* Notably, SLIQN demonstrates performance competitive to NIM, which is a second-order algorithm that utilizes the full Hessian information when taking the descent step.

2 RELATED WORK

In recent decades, several works have developed first-order, QN, and second-order methods for stochastic or incremental settings. Typically, the goal for first-order methods is to achieve a linear rate at a $\mathcal{O}(d)$ cost, while for QN and second-order methods, the goals are to achieve superlinear rates at costs of $\mathcal{O}(d^2)$ and $\mathcal{O}(d^3)$, respectively. These methods cater to different objectives: first-order methods are preferred for low-precision solutions, due to their lower computational cost, whereas higher-order methods are more effective for high-precision solutions, due to their faster rate.

Early works like Mokhtari and Ribeiro (2014, 2015); Byrd et al. (2016) were only successful in developing QN methods with sub-linear convergence guarantees. Subsequent works like Moritz et al. (2016); Chang et al. (2019); Derezinski (2023) employed various acceleration and variance reduction techniques to recover a linear rate. IQN by Mokhtari et al. (2018) was the first QN algorithm to achieve an asymptotic superlinear rate of convergence. IGS by Gao et al. (2020) employed greedy updates, introduced in Rodomanov and Nesterov (2021a), within the IQN framework to derive a explicit superlinear rate, albeit at a large $\mathcal{O}(d^3)$ per-iteration cost. Another recent work, Chen et al. (2022), put forth a QN style algorithm with a cost of $\mathcal{O}(d)$ for Generalized Linear Models (GLMs). However, the method only enjoys a linear rate of convergence, and is inefficient for general functions with a cost of $\mathcal{O}(d^3)$.

Other works have focused on developing first-order and second-order methods for stochastic or incremental settings. First order methods like Defazio et al. (2014); Johnson and Zhang (2013) have employed variance reduction techniques to derive methods with a linear rate of convergence. On the second-order front, recent works include the Newton Incremental method (NIM) by Rodomanov and Kropotov (2016) and Stochastic

*The code for the experiments is available on the repository: https://github.com/aakashlahoti/sliqn.

Newton (SN) by Kovalev et al. (2019). While SN and NIM are both Newton-like methods with a per-iteration complexity of $\mathcal{O}(d^3)$, NIM has a fast superlinear convergence rate while SN only enjoys a linear rate of convergence. The only setting under which SN has been shown to converge superlinearly is with a full batch of size n. We have consolidated the memory usage, computational cost, and convergence rates of principal-related algorithms in Table 1.

In a different line of work, Newton-LESS from Derezinski et al. (2021) utilized sketching algorithms like Newton Sketch Pilanci and Wainwright (2016, 2017) to attain a local linear convergence rate. Additional works such as Gonen et al. (2016); Liu et al. (2019) have used second-order information to accelerate SVRG, which is a first-order method. However, both methods were only able to attain an improved linear rate of convergence.

3 NOTATION AND PRELIMINARIES

Vectors (matrices) are denoted by lowercase (uppercase) bold alphabets. The i-th standard basis vector of $\mathbb{R}^d$ is denoted by $\boldsymbol{e}_i \in \mathbb{R}^d$ for $i \in [d] \coloneqq \{1, \ldots, d\}$. We define the index function $i_t \coloneqq 1+(t-1) \bmod n$. The symbol $\mathbf{0}$ denotes the all-zero matrix or vector, whose size can be inferred from the context. We use $\boldsymbol{X} \succeq \mathbf{0}$ and $\boldsymbol{X} \succ \mathbf{0}$ to denote that the symmetric matrix $\boldsymbol{X}$ is positive semi-definite and positive definite, respectively. Likewise, the notation $\boldsymbol{X} \succeq \boldsymbol{Y}$ ($\boldsymbol{X} \succ \boldsymbol{Y}$) denotes $\boldsymbol{X} - \boldsymbol{Y} \succeq \mathbf{0}$ ($\boldsymbol{X} - \boldsymbol{Y} \succ \mathbf{0}$). For vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we denote the inner product by $\langle \boldsymbol{u}, \boldsymbol{v} \rangle \coloneqq \boldsymbol{u}^\mathsf{T} \boldsymbol{v}$ and the Euclidean norm by $\|\boldsymbol{u}\| \coloneqq \sqrt{\langle \boldsymbol{u}, \boldsymbol{u} \rangle}$. For matrices $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d \times d}$, we define $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle \coloneqq \mathrm{Tr}(\boldsymbol{X}^\mathsf{T} \boldsymbol{Y})$, and we let $\|\boldsymbol{X}\|$ denote the spectral norm of the matrix. Given a convex function $f : \mathbb{R}^d \to \mathbb{R}$, we define the norm of the vector $\boldsymbol{y}$ with respect to $\nabla^2 f(\boldsymbol{x})$ as $\|\boldsymbol{y}\|_{\boldsymbol{x}} \coloneqq \sqrt{\langle \boldsymbol{y}, \nabla^2 f(\boldsymbol{x}) \boldsymbol{y} \rangle}$. For a function f, we denote μ as the strong convexity parameter, L as the smoothness parameter, $\tilde{L}$ as the Hessian Lipschitz continuity parameter, and $M = \tilde{L} \mu^{-\frac{3}{2}}$ as the strong self-concordance parameter.

3.1 Quasi-Newton (QN) Methods

We introduce QN methods as an iterative algorithm to optimize the problem ($\mathcal{P}$) with $n = 1$. At iteration $t \in \mathbb{Z}_+$, given the current iterate $\boldsymbol{x}^t$ and the positive definite Hessian approximation $\boldsymbol{B}^t$ of $\nabla^2 f(\boldsymbol{x}^t)$, the next iterate $\boldsymbol{x}^{t+1}$ is computed as,

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - (\boldsymbol{B}^t)^{-1} \nabla f(\boldsymbol{x}^t). \tag{1}$$

The Hessian approximation for the next iteration $\boldsymbol{B}^{t+1}$, is obtained by applying a constant rank update to $\boldsymbol{B}^t$. The precise update distinguishes the exact type of QN algorithm, such as BFGS, DFP, or SR1 (Nocedal and Wright (1999)). The efficiency of QN methods, $\mathcal{O}(d^2)$, over second-order methods, $\mathcal{O}(d^3)$, stems from the fact that the $\boldsymbol{B}^t$ update is low rank, which allows us to use Sherman-Morrison formula (Appendix A) to efficiently evaluate $(\boldsymbol{B}^{t+1})^{-1}$ from $(\boldsymbol{B}^t)^{-1}$ in $\mathcal{O}(d^2)$ cost.

Though in the remainder of the paper, we are primarily concerned with BFGS updates, all the follows can also be extended to the entire restricted Broyden class (Appendix H). Given a matrix $\boldsymbol{K}$ and its approximation $\boldsymbol{B}$, the generalized BFGS update refines this approximation along direction $\boldsymbol{u} \in \mathbb{R}^d$ as,

$$\mathrm{BFGS}(\boldsymbol{B}, \boldsymbol{K}, \boldsymbol{u}) = \boldsymbol{B}_+ \coloneqq \boldsymbol{B} - \frac{\boldsymbol{B}\boldsymbol{u}\boldsymbol{u}^\mathsf{T}\boldsymbol{B}}{\langle \boldsymbol{u}, \boldsymbol{B}\boldsymbol{u} \rangle} + \frac{\boldsymbol{K}\boldsymbol{u}\boldsymbol{u}^\mathsf{T}\boldsymbol{K}}{\langle \boldsymbol{u}, \boldsymbol{K}\boldsymbol{u} \rangle}. \tag{2}$$

Setting $\boldsymbol{K}^t = \int_0^1 \nabla^2 f(\boldsymbol{x}^t + \tau(\boldsymbol{x}^{t+1} - \boldsymbol{x}^t)) d\tau$, and $\boldsymbol{u}^t = \boldsymbol{s}^t \coloneqq \boldsymbol{x}^{t+1} - \boldsymbol{x}^t$ yields the classical BFGS update,

$$\begin{aligned} \boldsymbol{B}^{t+1} &= \mathrm{BFGS}(\boldsymbol{B}^t, \boldsymbol{K}^t, \boldsymbol{u}^t) \\ &= \boldsymbol{B}^t - \frac{\boldsymbol{B}^t \boldsymbol{s}^t (\boldsymbol{s}^t)^\mathsf{T} \boldsymbol{B}^t}{\langle \boldsymbol{s}^t, \boldsymbol{B}^t \boldsymbol{s}^t \rangle} + \frac{\boldsymbol{y}^t (\boldsymbol{y}^t)^\mathsf{T}}{\langle \boldsymbol{s}^t, \boldsymbol{y}^t \rangle}, \end{aligned} \tag{3}$$

where $\boldsymbol{y}_t \coloneqq \boldsymbol{K}^t \boldsymbol{s}^t = \nabla f(\boldsymbol{x}^{t+1}) - \nabla f(\boldsymbol{x}^t)$. This update seeks to approximate the Hessian along the Newton direction $\boldsymbol{s}^t$ and has been shown by Jin and Mokhtari (2022); Rodomanov and Nesterov (2021c,b) to achieve a superlinear convergence rate. Furthermore, since BFGS makes a rank 2 update to $\boldsymbol{B}^t$, we can use the Sherman-Morrison formula twice to evaluate $(\boldsymbol{B}^{t+1})^{-1}$ from $(\boldsymbol{B}^t)^{-1}$ in $\mathcal{O}(d^2)$ cost.

In contrast to classical BFGS update, the greedy BFGS update by Rodomanov and Nesterov (2021a) sets $\boldsymbol{K}^t = \nabla^2 f(\boldsymbol{x}^t)$, and defines the greedy vector,

$$\bar{\boldsymbol{u}}(\boldsymbol{B}^t, \boldsymbol{K}^t) \coloneqq \underset{\boldsymbol{u} \in \{\boldsymbol{e}_i\}_{i=1}^d}{\arg\max} \frac{\langle \boldsymbol{u}, \boldsymbol{B}^t \boldsymbol{u} \rangle}{\langle \boldsymbol{u}, \boldsymbol{K}^t \boldsymbol{u} \rangle}, \tag{4}$$

which results in $\boldsymbol{B}^{t+1} = \mathrm{BFGS}(\boldsymbol{B}^t, \boldsymbol{K}^t, \bar{\boldsymbol{u}}^t(\boldsymbol{B}^t, \boldsymbol{K}^t))$. Greedy BFGS, similar to classic BFGS, exhibits a superlinear rate of convergence. However, unlike classic BFGS, it can guarantee convergence in the σ sense. Specifically, the Hessian approximation error,

$$\sigma(\boldsymbol{B}^t, \boldsymbol{K}^t) \coloneqq \langle (\boldsymbol{K}^t)^{-1}, \boldsymbol{B}^t - \boldsymbol{K}^t \rangle, \tag{5}$$

decays linearly with t. In practice, a trade-off exists between the two updates. Greedy BFGS dedicates initial iterations to construct a reliable Hessian approximation. In contrast, classic BFGS gains an initial advantage because it updates along the Newton direction. Greedy BFGS only outperforms classic BFGS if it has enough time to build an accurate approximation before convergence. Sharpened BFGS proposed by Jin et al. (2022) incorporated both classic and greedy BFGS updates, to

Algorithm	Memory	Computation cost	Convergence Rate	Limit
SN	$\boldsymbol{\mathcal{O}(d^2)}$	$\mathcal{O}(d^3)$	Linear	**Non-asymptotic**
NIM	$\mathcal{O}(nd + d^2)$	$\mathcal{O}(d^3)$	**Superlinear**	**Non-asymptotic**
IQN	$\mathcal{O}(nd^2)$	$\boldsymbol{\mathcal{O}(d^2)}$	**Superlinear**	Asymptotic
IGS	$\mathcal{O}(nd^2)$	$\mathcal{O}(d^3)$	**Superlinear**	**Non-asymptotic**
SIQN(This work)	$\mathcal{O}(nd^2)$	$\mathcal{O}(d^3)$	**Superlinear**	**Non-asymptotic**
SLIQN(This work)	$\mathcal{O}(nd^2)$	$\boldsymbol{\mathcal{O}(d^2)}$	**Superlinear**	**Non-asymptotic**

Table 1: Comparison of the maximum memory requirement, computation cost (per iteration), convergence rate, and the limit of attainment of the convergence rate for different algorithms.

achieve the best of both updates: an explicit superlinear convergence of $\boldsymbol{x}^t$, a linear convergence of $\boldsymbol{B}^t$ and no initial "ramp-up" phase to build the approximation.

3.2 Incremental Quasi-Newton (IQN)

We now introduce IQN by Mokhtari et al. (2018). For each iteration $t \geq 1$, IQN maintains tuples of the form $(\boldsymbol{z}_i^t, \nabla f_i(\boldsymbol{z}_i^t), \boldsymbol{B}_i^t)$ for each index $i \in [n]$. Here $\boldsymbol{z}_i^t \in \mathbb{R}^d$ is the iterate corresponding to the function f_i, $\nabla f_i(\boldsymbol{z}_i^t)$ is the gradient of f_i at $\boldsymbol{z}_i^t$, and $\boldsymbol{B}_i^t$ is the positive definite Hessian approximation of $\nabla^2 f_i(\boldsymbol{z}_i^t)$.

IQN begins by constructing a second-order Taylor approximation g_i^t of each f_i, centered at $\boldsymbol{z}_i^{t-1}$ and using the Hessian approximation $\boldsymbol{B}_i^{t-1}$ as,

$$g_i^t(\boldsymbol{x}) := f_i(\boldsymbol{z}_i^{t-1}) + \langle \nabla f_i(\boldsymbol{z}_i^{t-1}), \boldsymbol{x} - \boldsymbol{z}_i^{t-1} \rangle + \frac{1}{2}\langle \boldsymbol{x} - \boldsymbol{z}_i^{t-1}, \boldsymbol{B}_i^{t-1}(\boldsymbol{x} - \boldsymbol{z}_i^{t-1}) \rangle. \quad (6)$$

The iterate $\boldsymbol{x}^t$ is then calculated as,

$$\boldsymbol{x}^t = \arg\min_{\boldsymbol{x}} \frac{1}{n}\sum_{i=1}^n g_i^t(\boldsymbol{x}) = (\bar{\boldsymbol{B}}^{t-1})^{-1}\left(\sum_{i=1}^n \boldsymbol{B}_i^{t-1}\boldsymbol{z}_i^{t-1} - \nabla f_i(\boldsymbol{z}_i^{t-1})\right), \quad (7)$$

where $\bar{\boldsymbol{B}}^{t-1} = \sum_{i=1}^n \boldsymbol{B}_i^{t-1}$. In every iteration t, IQN only updates the tuple whose index is given by $i_t = 1 + (t-1) \bmod n$, using the following scheme:

1. $\boldsymbol{z}_{i_t}^t = \boldsymbol{x}^t$, $\nabla f_{i_t}(\boldsymbol{z}_{i_t}^t) = \nabla f_{i_t}(\boldsymbol{x}^t)$.
2. $\boldsymbol{z}_j^t = \boldsymbol{z}_j^{t-1}$, $\nabla f_j(\boldsymbol{z}_j^t) = \nabla f_j(\boldsymbol{z}_j^{t-1})$, and $\boldsymbol{B}_j^t = \boldsymbol{B}_j^{t-1}$, for all $j \neq i_t$.
3. $\boldsymbol{B}_{i_t}^t = \text{BFGS}(\boldsymbol{B}_{i_t}^{t-1}, \boldsymbol{K}^t, \boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1})$, where $\boldsymbol{K}^t = \int_0^1 \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^{t-1} + \tau(\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}))d\tau$.

The per-iteration complexity of IQN: the cost of gradient evaluation in (1) is $\mathcal{O}(d)$, the no-operation step in (2) incurs no cost, and the BFGS step (3) has a $\mathcal{O}(d^2)$ cost as it is a constant rank update to $\bar{\boldsymbol{B}}_{i_t}^t$. To compute the iterate $\boldsymbol{x}^{t+1}$, we first evaluate the inverse of $\bar{\boldsymbol{B}}^t = \bar{\boldsymbol{B}}^{t-1} + \boldsymbol{B}_{i_t}^t - \boldsymbol{B}_{i_t}^{t-1}$ from $(\bar{\boldsymbol{B}}^{t-1})^{-1}$ using the Sherman-Morrison formula at a cost of $\mathcal{O}(d^2)$. Then, we calculate $\sum_{i=1}^n \boldsymbol{B}_i^t \boldsymbol{z}_i^t - \nabla f_i(\boldsymbol{z}_i^t)$ from the memoized value of $\sum_{i=1}^n \boldsymbol{B}_i^{t-1}\boldsymbol{z}_i^{t-1} - \nabla f_i(\boldsymbol{z}_i^{t-1})$ in $\mathcal{O}(d^2)$ cost. Therefore, the overall per-iteration cost of IQN is $\mathcal{O}(d^2)$. Please refer to Appendix C for details.

4 PROPOSED ALGORITHM

We first introduce the SIQN method, which incorporates the sharpened updates into the IQN framework. However, this direct incorporation results in an inefficient $\mathcal{O}(d^3)$ method. This is because, ensuring the positive semi-definiteness of the Hessian approximation with respect to the true Hessian results in Hessian updates which are not low rank. And consequently, the Hessian inversion incurs a large $\mathcal{O}(d^3)$ cost. We then propose the SLIQN algorithm that overcomes this problem by modifying the SIQN updates using a constant corrective multiplicative factor based on the theoretical analysis of SIQN. It then utilizes a novel lazy propagation strategy to implement this factor correction efficiently with a per-iteration cost of $\mathcal{O}(d^2)$.

4.1 Sharpened Incremental Quasi-Newton (SIQN)

Similar to IQN, SIQN also maintains tuples of the form $(\boldsymbol{z}_i^t, \nabla f_i(\boldsymbol{z}_i^t), \boldsymbol{B}_i^t)$ for each iteration $t \geq 1$ and each index $i \in [n]$. The iterate $\boldsymbol{x}^t$ is calculated as,

$$\boldsymbol{x}^t = (\bar{\boldsymbol{B}}^{t-1})^{-1}\left(\sum_{i=1}^n \boldsymbol{B}_i^{t-1}\boldsymbol{z}_i^{t-1} - \nabla f_i(\boldsymbol{z}_i^{t-1})\right), \quad (8)$$

where $\bar{\boldsymbol{B}}^{t-1} = \sum_{i=1}^n \boldsymbol{B}_i^{t-1}$. The tuples are updated in a deterministic cyclic order as follows:

1. $\boldsymbol{z}_{i_t}^t = \boldsymbol{x}^t$, $\nabla f_{i_t}(\boldsymbol{z}_{i_t}^t) = \nabla f_{i_t}(\boldsymbol{x}^t)$.
2. $\boldsymbol{z}_j^t = \boldsymbol{z}_j^{t-1}$, $\nabla f_j(\boldsymbol{z}_j^t) = \nabla f_j(\boldsymbol{z}_j^{t-1})$, and $\boldsymbol{B}_j^t = \boldsymbol{B}_j^{t-1}$, for all $j \neq i_t$.

To update $\boldsymbol{B}_{i_t}^t$, SIQN first performs the classic BFGS update followed by the greedy update,

Algorithm 1 Sharpened Incremental Quasi-Newton

1: **Initialize:** $\{\boldsymbol{z}_i^0 = \boldsymbol{x}^0\}_{i=1}^n, \{\boldsymbol{B}_i^0\}_{i=1}^n$ such that for all $i \in [n]$, $\boldsymbol{B}_i^0 \succeq \nabla^2 f_i(\boldsymbol{z}_i)^0$
2: **while** *not converged:*
3: Set current index $i_t = (t-1) \mod n + 1$;
4: Update $\boldsymbol{x}^t$ as per (8);
5: Update $\boldsymbol{z}_{i_t}^t = \boldsymbol{x}^t$ and $\boldsymbol{B}_{i_t}^t$ as per (3);
6: Update the tuples with index $j \neq i_t$ as $\boldsymbol{z}_j^t = \boldsymbol{z}_j^{t-1}, \boldsymbol{B}_j^t = \boldsymbol{B}_j^{t-1}$;
7: Increment the iteration counter t;
8: **end while**

3. $\boldsymbol{Q}^t = \text{BFGS}((1+\beta_t)^2\boldsymbol{B}_{i_t}^{t-1}, (1+\beta_t)\boldsymbol{K}^t, \boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}),$

$$\boldsymbol{B}_{i_t}^t = \text{BFGS}(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t), \bar{\boldsymbol{u}}(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))),$$

where, $\boldsymbol{K}^t \coloneqq \int_0^1 \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^{t-1} + \tau(\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}))d\tau$ and the scaling factor $\beta_t \coloneqq \frac{M}{2}\left\|\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}\right\|_{\boldsymbol{z}_{i_t}^{t-1}}$ ensures that before the classical BFGS update, the Hessian approximation is positive semi-definite with respect to the corresponding Hessian, $(1+\beta_t)^2\boldsymbol{B}_{i_t}^{t-1} \succeq (1+\beta_t)\boldsymbol{K}^t$. For technical details, please refer to Lemma E.1. The pseudo-code for SIQN is provided in Algorithm 1.

We now consider the per-iteration complexity of SIQN: the cost of gradient evaluation is $\mathcal{O}(d)$, the no-operation step incurs no cost and the BFGS step can be computed in $\mathcal{O}(d^2)$ cost. To compute $\boldsymbol{x}^{t+1}$, we need to calculate the inverse of $\bar{\boldsymbol{B}}^t = \bar{\boldsymbol{B}}^{t-1} + \boldsymbol{B}_{i_t}^t - \boldsymbol{B}_{i_t}^{t-1}$,

$$\begin{aligned}
\boldsymbol{B}_{i_t}^t - \boldsymbol{B}_{i_t}^{t-1} &= \text{BFGS}(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t), \\
&\qquad \bar{\boldsymbol{u}}(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))) - \boldsymbol{B}_{i_t}^{t-1}, \\
&= \boldsymbol{Q}^t - \frac{\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t(\bar{\boldsymbol{u}}^t)^\mathsf{T}\boldsymbol{Q}^t}{\langle \bar{\boldsymbol{u}}^t, \boldsymbol{Q}^t\bar{\boldsymbol{u}}^t\rangle} \\
&\quad + \frac{\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t(\bar{\boldsymbol{u}}^t)^\mathsf{T}\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)}{\langle \bar{\boldsymbol{u}}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t\rangle} - \boldsymbol{B}_{i_t}^{t-1}, \\
&= ((1+\beta_t)^2 - 1)\boldsymbol{B}_{i_t}^{t-1} \\
&\quad - (1+\beta_t)^2\frac{\boldsymbol{B}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\boldsymbol{s}_{i_t}^{t\,\mathsf{T}}\boldsymbol{B}_{i_t}^{t-1}}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{B}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\rangle} \\
&\quad + (1+\beta_t)\frac{\boldsymbol{y}_{i_t}^t\boldsymbol{y}_{i_t}^{t\,\mathsf{T}}}{\langle \boldsymbol{y}_{i_t}^{t\,\mathsf{T}}, \boldsymbol{s}_{i_t}^t\rangle} \\
&\quad - \frac{\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t(\bar{\boldsymbol{u}}^t)^\mathsf{T}\boldsymbol{Q}^t}{\langle \bar{\boldsymbol{u}}^t, \boldsymbol{Q}^t\bar{\boldsymbol{u}}^t\rangle} \\
&\quad + \frac{\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t(\bar{\boldsymbol{u}}^t)^\mathsf{T}\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)}{\langle \bar{\boldsymbol{u}}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t\rangle}, \qquad (9)
\end{aligned}$$

where $\bar{\boldsymbol{u}}^t = \bar{\boldsymbol{u}}(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))$, $\boldsymbol{s}_{i_t}^t = \boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}$ and $\boldsymbol{y}_{i_t}^t = \nabla f_{i_t}(\boldsymbol{z}_{i_t}^t) - \nabla f_{i_t}(\boldsymbol{z}_{i_t}^{t-1})$. Observe that the expression (9) is generally not a matrix of constant rank, since the rank of $((1+\beta_t)^2-1)\boldsymbol{B}_{i_t}^{t-1}$ may be as large as d. Therefore, it is not possible to compute the inverse of $\bar{\boldsymbol{B}}^t$ by using the inverse of $\bar{\boldsymbol{B}}^{t-1}$ in $\mathcal{O}(d^2)$ cost.

4.2 Sharpened Lazy Incremental Quasi-Newton (SLIQN)

We now present the SLIQN method as a solution to the aforementioned issues. We observe that the primary reason for the update of $\bar{\boldsymbol{B}}^t$ in SIQN not being low-rank, is the presence of the scaling factor β_t. To resolve this issue, we begin by noting that the convergence analysis of SIQN (ref. Appendix E) indicates that there exists a factor $\alpha_{\lceil t/n-1\rceil}$, such that $\beta_t \leq \alpha_{\lceil t/n-1\rceil}$ and using $\alpha_{\lceil t/n-1\rceil}$ instead of β_t preserves the convergence properties of SIQN. Since $\alpha_{\lceil t/n-1\rceil}$ is constant throughout an epoch, rather than multiplying $\alpha_{\lceil t/n-1\rceil}$ to $\boldsymbol{B}_{i_t}^{t-1}$ at iteration t, we instead pre-multiply $\alpha_{\lceil t/n-1\rceil}$ to each Hessian approximation at the start of every epoch. This enables us to compute the inverse of $\bar{\boldsymbol{B}}^t$ trivially by dividing the old inverse by $\alpha_{\lceil t/n-1\rceil}$. However, this pre-multiplication step is a $\mathcal{O}(nd^2)$ operation and it undermines the utility of incremental algorithms. To address this issue, we employ a lazy propagation strategy, wherein we scale the individual Hessian approximations just before they are updated in their respective iterations, but treat all memoized quantities as if the approximations are already scaled. These key changes enable SLIQN to achieve an $\mathcal{O}(d^2)$ per-iteration cost along with an explicit superlinear rate.

In what follows, we will denote the Hessian approximations by $\{\boldsymbol{D}_i^t\}_{i=1}^n$ instead of $\{\boldsymbol{B}_i^t\}_{i=1}^n$ to distinguish SLIQN updates from SIQN updates. We now formally present the SLIQN algorithm.

Initialization: At $t = 0$, we initialize the iterates $\{\boldsymbol{z}_i^0\}_{i=1}^n$ as $\boldsymbol{z}_i^0 = \boldsymbol{x}^0$ and their corresponding hessian approximations $\{\boldsymbol{D}_i^0\}_{i=1}^n$ as $\boldsymbol{D}_i^0 = (1+\alpha_0)^2\boldsymbol{I}_i^0$, where $\boldsymbol{x}_0, \alpha_0, \{\boldsymbol{I}_i^0\}_{i=1}^n$ satisfy the the premise of Lemma 2.

Iterative Updates: For each iteration $t \geq 1$, we set the iterate $\boldsymbol{x}^t$ as,

$$\boldsymbol{x}^t = \left(\bar{\boldsymbol{D}}^{t-1}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{D}_i^{t-1}\boldsymbol{z}_i^{t-1} - \nabla f_i(\boldsymbol{z}_i^{t-1})\right), \quad (10)$$

where $\bar{\boldsymbol{D}}^{t-1} = \sum_{i=1}^n \boldsymbol{D}_i^{t-1}$. The scheme to update each tuple is as follows:

1. $\boldsymbol{z}_{i_t}^t = \boldsymbol{x}^t$, $\nabla f_{i_t}(\boldsymbol{z}_{i_t}^t) = \nabla f_{i_t}(\boldsymbol{x}^t)$.
2. $\boldsymbol{z}_i^t = \boldsymbol{z}_i^{t-1}, \boldsymbol{D}_i^t = \omega_t\boldsymbol{D}_i^{t-1}$, $\forall i \in [n]\backslash\{i_t\}$, where $\omega_t \coloneqq (1+\alpha_{\lceil t/n\rceil})^2$ if $t \bmod n = 0$ and 1 otherwise.
3. $\boldsymbol{Q}^t = \text{BFGS}(\boldsymbol{D}_{i_t}^{t-1}, (1+\alpha_{\lceil t/n-1\rceil})\boldsymbol{K}^t, \boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1})$,

$$\boldsymbol{D}_{i_t}^t = \omega_t\text{BFGS}(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t), \bar{\boldsymbol{u}}^t(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))).$$

Algorithm 2 provides the pseudo-code for SLIQN. We now consider its per-iteration complexity. Observe that in the $j+1^{\text{th}}$ epoch, the updates (2) and (3) are carried out differently for $t \bmod n = 0$ and $t \bmod n \neq 0$. Specifically, for $t \in \{nj+1, \ldots, nj+n-1\}$, we can carry out the update (10) in $\mathcal{O}(d^2)$ cost using Sherman-Morrison formula (ref. Appendix F.1) and can compute other iterative updates in $\mathcal{O}(d^2)$ cost as they consist of a constant number of matrix-vector multiplications. However, for $t = nj+n$, each $\boldsymbol{D}_i^t$ is multiplied with the scaling factor of $(1+\alpha_{j+1})^2$, which incurs a large $\mathcal{O}(nd^2)$ overhead. Instead, we implement this step by lazily scaling the Hessian approximations at the iteration in which they are updated while treating all memoized quantities as if the approximations are already scaled. The details of the lazy strategy are provided in Appendix F.2

Algorithm 2 Sharpened Lazy IQN

1: **Initialization:** Initialize the iterates $\{\boldsymbol{z}_i^0\}_{i=1}^n$ as $\boldsymbol{z}_i^0 = \boldsymbol{x}^0$ and their corresponding hessian approximations $\{\boldsymbol{D}_i^0\}_{i=1}^n$ as $\boldsymbol{D}_i^0 = (1+\alpha_0)^2 \boldsymbol{I}_i^0$
2: **while** *not converged:*
3: Update $\boldsymbol{x}^t$ as per (10);
4: Update $\boldsymbol{z}_{i_t}^t$ as $\boldsymbol{z}_{i_t}^t = \boldsymbol{x}^t$;
5: Update $\boldsymbol{Q}^t$ and $\boldsymbol{D}_{i_t}^t$ as per (3);
6: Update the tuples with index $i \neq i_t$ as per (2);
7: Increment the iteration counter t;
8: **end while**

Remark 1 *We remark that both IQN and SLIQN exhibit a per-iteration cost of $\mathcal{O}(d^2)$ which is in contrast to the $\mathcal{O}(nd^2)$ cost for QN methods. However, this efficiency comes with an increased memory cost of $\mathcal{O}(nd^2)$. To address this issue we propose a pipelining scheme in Appendix D, in which we leverage a much larger disk to augment the main memory by prefetching the data and processing the updates in parallel.*

5 THEORETICAL ANALYSIS OF ALGORITHM 2

5.1 Assumptions

We analyze SLIQN under the assumptions of smoothness, strong convexity, and Lipschitz continuity of the Hessian. These assumptions are commonly used in the analysis of Quasi-Newton methods.

A1 ***(Strong convexity and smoothness)*** *The functions $\{f_i\}_{i=1}^n$ are μ-strongly convex and L-smooth, that is $\frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 \leq f_i(\boldsymbol{y}) - f_i(\boldsymbol{x}) - \langle \nabla f_i(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle$, and $f_i(\boldsymbol{y}) - f_i(\boldsymbol{x}) - \langle \nabla f_i(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle \leq \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2$ hold for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, and for all $i \in [n]$.*

A2 ***(Lipschitz continuous Hessian)*** *The Hessians $\{\nabla^2 f_i\}_{i=1}^n$ are $\tilde{L}$-Lipschitz continuous, that is, $\left\|\nabla^2 f_i(\boldsymbol{y}) - \nabla^2 f_i(\boldsymbol{x})\right\| \leq \tilde{L}\|\boldsymbol{y}-\boldsymbol{x}\|$, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, and for all $i \in [n]$.*

The above assumptions also imply that the functions $\{f_i\}_{i=1}^n$ are M-strongly self-concordant, which is defined as, $\nabla^2 f_i(\boldsymbol{y}) - \nabla^2 f_i(\boldsymbol{x}) \preceq M\|\boldsymbol{y}-\boldsymbol{x}\|_{\boldsymbol{z}} \nabla^2 f_i(\boldsymbol{w})$, $\forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w} \in \mathbb{R}^d$ with $M \coloneqq \tilde{L}\mu^{-\frac{3}{2}}$ (Rodomanov and Nesterov, 2021a, Ex 4.1).

5.2 Convergence Lemmas

We establish the convergence guarantees in three steps: Lemma 1 establishes a one-step inequality that bounds the residual $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\|$ in terms of the previous residuals $\left\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\right\|$ and the norm error in the Hessian approximation $\left\|\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\right\|$, $\forall i \in [n]$. Lemma 2 uses the result of Lemma 1 to inductively show that both the residual $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\|$ and the Hessian approximation error in the σ sense, i.e., $\sigma(\boldsymbol{D}_{i_t}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))$, decrease linearly with $\lceil t/n \rceil$. Using the result of Lemma 1 and Lemma 2, Lemma 3 establishes a mean-superlinear convergence result. We finally show in Theorem 1 that the residuals can be upper bounded by a superlinearly convergent sequence.

Lemma 1 *If Assumptions **A1** and **A2** hold, the sequence of iterates generated by Algorithm 2 satisfy*

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\| \leq \frac{\tilde{L}\Gamma^{t-1}}{2}\sum_{i=1}^n \left\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\right\|^2 + \Gamma^{t-1}\sum_{i=1}^n \left\|\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\right\| \left\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\right\|, \quad (11)$$

for all $t \geq 1$, where $\Gamma^t \coloneqq \left\|\left(\sum_{i=1}^n \boldsymbol{D}_i^t\right)^{-1}\right\|$.

The proof of this result can be found in Appendix G.1. It is important to note that our bound in (11) differs from a similar result presented in (Mokhtari et al., 2018, Lemma 2), where they utilize $\left\|\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{x}^\star)\right\|$ instead of $\left\|\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\right\|$. This modification helps connect the approximation error in the norm sense, i.e., $\left\|\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\right\|$ with the error in the σ sense, i.e., $\sigma(\boldsymbol{D}_i^{t-1}, \nabla^2 f_i(\boldsymbol{z}_i^{t-1}))$. This connection is crucial for quantifying the improvements achieved by the greedy updates.

Lemma 2 *If Assumptions **A1** and **A2** hold, for any ρ such that $0 < \rho < 1 - \frac{\mu}{dL}$, there exist positive constants ϵ and σ_0 such that if $\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| \leq \epsilon$ and $\sigma(\boldsymbol{I}_i^0, \nabla^2 f_i(\boldsymbol{x}^0)) \leq \sigma_0$ for all $i \in [n]$, the sequence of iterates generated by Algorithm 2 satisfy*

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\| \leq \rho^{\lceil \frac{t}{n} \rceil}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|. \quad (12)$$

Further, it holds that

$$\sigma(\omega_t^{-1}\boldsymbol{D}_{i_t}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)) \leq \big(1-\frac{\mu}{dL}\big)^{\lceil \frac{t}{n} \rceil}\delta, \quad (13)$$

where $\delta := e^{\frac{4M\sqrt{L}\epsilon}{1-\rho}}\big(\sigma_0 + \epsilon\frac{4Md\sqrt{L}}{1-\frac{\rho}{1-\frac{\mu}{dL}}}\big), M = \tilde{L}/\mu^{\frac{3}{2}}$, $\omega_t = (1+\alpha_{\lceil t/n \rceil})^2$ *if* t *is a multiple of* n *and* 1 *otherwise, and the sequence* $\{\alpha_k\}$ *is defined as* $\alpha_k := M\sqrt{L}\epsilon\rho^k, \forall k \geq 0$.

The proof of Lemma 2 can be found in Appendix G.2. Under the hood, the proof uses induction to show that if the initialized $\boldsymbol{x}^0$ is close to $\boldsymbol{x}^\star$ and $\boldsymbol{D}_i^0$ is close to $\nabla^2 f_i(\boldsymbol{x}^0)$, then the iterate $\boldsymbol{x}^t$ converges linearly to $\boldsymbol{x}^\star$ and $\boldsymbol{D}_{i_t}^t$ converges linearly to $\nabla^2 f_i(\boldsymbol{z}_{i_t}^t)$. Our result is stronger than (Mokhtari et al., 2018, Lemma 3) as we establish the linear convergence of $\boldsymbol{D}_i^t$, whereas Mokhtari et al. (2018) were only able to establish that $\big\|\boldsymbol{D}_i^t - \nabla^2 f_i(\boldsymbol{x}^\star)\big\|$ does not grow with t. Moreover, Mokhtari et al. (2018) did not guarantee convergence of $\boldsymbol{D}_i^t$. Also, there is no equivalent convergence result presented in the analysis of IGS.

Lemma 3 *If Assumptions* ***A1****,* ***A2*** *hold, then there exist positive constants* ϵ *and* σ_0 *such that if* $\big\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\big\| \leq \epsilon$ *and* $\sigma(\boldsymbol{I}_i^0, \nabla^2 f_i(\boldsymbol{x}^0)) \leq \sigma_0$ *for all* $i \in [n]$*, the sequence of iterates generated by Algorithm 2 satisfy*

$$\big\|\boldsymbol{x}^t - \boldsymbol{x}^\star\big\| \leq \big(1-\frac{\mu}{dL}\big)^{\lceil \frac{t}{n} \rceil}\frac{1}{n}\sum_{i=1}^{n}\big\|\boldsymbol{x}^{t-i} - \boldsymbol{x}^\star\big\|.$$

The proof of Lemma 3 can be found in Appendix G.3. The main idea behind the proof is to substitute the linear convergence results, specifically (12) and (13) from Lemma 2, back into the result from Lemma 1. By doing so, the first term on the right-hand side of (11) converges quadratically, while the second term converges superlinearly, which proves the result.

Our result is markedly different from (Mokhtari et al., 2018, Theorem 6) which proves asymptotic superlinear convergence of IQN. Their proof is based on a variant of the Dennis-Moré theorem to show that superlinear convergence kicks asymptotically. Since their proof is existential, unlike Lemma 3, it cannot be used to derive an explicit rate of superlinear convergence.

Theorem 1 *If Assumptions* ***A1****,* ***A2*** *hold, then there exist positive constants* ϵ *and* σ_0 *such that if* $\big\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\big\| \leq \epsilon$ *and* $\sigma(\boldsymbol{I}_i^0, \nabla^2 f_i(\boldsymbol{x}^0)) \leq \sigma_0$ *for all* $i \in [n]$*, and for the sequence of iterates generated by Algorithm 2, there exists a sequence* $\{\zeta^k\}$, $k \geq 1$ *such that* $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\| \leq \zeta^{\lfloor \frac{t-1}{n} \rfloor}$ *for all* $t \geq 1$ *and* $\{\zeta^k\}$ *satisfies,*

$$\zeta^k \leq \epsilon\big(1-\frac{\mu}{dL}\big)^{\frac{(k+2)(k+1)}{2}}. \quad (14)$$

The proof of Theorem 1 involves the construction of a sequence that provides an upper bound on the residual $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\|$, and can be found in Appendix G.4.

Remark 2 *It is instructive to compare our convergence rate with that of IGS (Gao et al., 2020, Theorem 3). According to their result, the rate is given as* $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\| \leq \big(1-\frac{\mu}{dL}\big)^{\frac{k(k+1)}{2}} r^{k_0}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|$, $\forall t \geq 1$, $r \in (0,1), k = \big\lfloor \frac{t-1}{n} \big\rfloor + 1 - k_0$ *and* k_0 *is a constant such that* $\big(1-\frac{\mu}{dL}\big)^{k_0} D \leq 1$*. The parameter* D *depends on the underlying objective function. Observe that their superlinear rate only takes effect after* $\big\lfloor \frac{t-1}{n} \big\rfloor \geq k_0 - 1$*, and* k_0 *could potentially be large, though it is not possible to infer the bounds on* k_0 *from Gao et al. (2020). In contrast, our convergence rate guarantees superlinear convergence right from the first iteration.*

6 NUMERICAL EXPERIMENTS

6.1 Quadratic Function Minimization

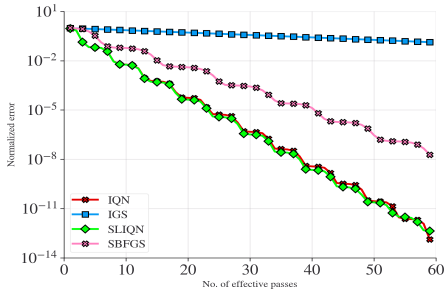
We begin with a comparative analysis of the empirical performance of SLIQN, IQN, Sharpened BFGS (SBFGS) Jin et al. (2022), and IGS on a synthetic quadratic minimization problem.

Problem Definition: We consider the function $f(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\big(\frac{1}{2}\langle \boldsymbol{x}, \boldsymbol{A}_i\boldsymbol{x}\rangle + \langle \boldsymbol{b}_i, \boldsymbol{x}\rangle\big)$, where $\boldsymbol{A}_i \succ \boldsymbol{0}$, and $\boldsymbol{b}_i \in \mathbb{R}^d$, $\forall i \in [n]$. The detailed generation scheme for $\boldsymbol{A}_i, \boldsymbol{b}_i$ can be found in Appendix I.1.

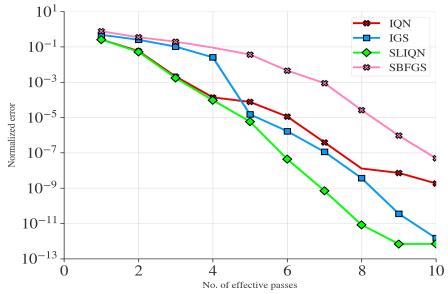
Experiments and Inference: We study the performance of the algorithms on two extreme cases: $d \gg n$ (Fig. 1a), and $n \gg d$ (Fig. 1b). In each case, we plot the normalized error $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\| / \big\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\big\|$ against the number of effective passes or epochs. We see that in both these cases, SLIQN outperforms IGS, IQN, and SBFGS. We also observe that in the case where $n \gg d$, IGS outperforms IQN, whereas in the case where $d \gg n$, IQN surpasses IGS. This is because IGS devotes the initial $\mathcal{O}(d)$ iterations to constructing a precise Hessian approximation, after which its fast convergence phase kicks in. On the other hand, since IQN takes the descent step in the Newton direction, its Hessian approximation is never precise and therefore its normalized error decreases at more or less a "consistent" rate. SLIQN combines the strengths of both IQN and IGS: during the initial iterations when its Hessian approximation is not accurate enough, the classical BFGS updates are responsible for the progress made. In the later iterations, when the Hessian has been sufficiently well approximated, its fast convergence phase kicks in.

6.2 Regularized Logistic Regression

We now compare the performance of SLIQN against IQN, IGS, NIM, and SN on the regularized logistic



(a) $n = 20, d = 500$



(b) $n = 50000, d = 10$

Figure 1: Normalized error vs. number of effective passes for the quadratic minimization problem

regression task given by

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N}\big(y_i \log(1+e^{-\langle \boldsymbol{x}, \boldsymbol{z}_i\rangle}) + (1-y_i)\log(1+e^{\langle \boldsymbol{x}, \boldsymbol{z}_i\rangle})\big) + \frac{\lambda}{2}\|\boldsymbol{x}\|^p, \quad (15)$$

where $\{\boldsymbol{z}_i\}_{i=1}^N$ are the training samples, $\{y_i\}_{i=1}^N$ are their corresponding binary labels and p is set at 2.1. It is easy to observe that $f(\boldsymbol{x})$ is smooth, strongly convex and has a Lipschitz continuous Hessian, thereby satisfying Assumptions **A1**, **A2**. We compare the algorithms across 9 datasets with a large variation in the values of n and d. Each algorithm is initialization with the same iterate and the regularization parameter is set as $\lambda = 1/N$. For SN, we set the mini-batch size $\tau = N$ since that is the regime it works best in. Please refer to Appendix I for a complete experimental setup. We observe in Figure 2 that SLIQN outperforms IGS, IQN, and SN on each of the 9 datasets from LIBSVM by Chang and Lin (2011). This supports our claim that SLIQN offers the best of both, IQN and IGS. Furthermore, we observe that while NIM outperforms SLIQN, their performance remains comparable. It is important to note that NIM utilizes the full Hessian information for the descent step and is an $\mathcal{O}(d^3)$ algorithm, while SLIQN has a per-iteration complexity of $\mathcal{O}(d^2)$. Thus, these results underscore the superiority of SLIQN over other incremental QN style methods.

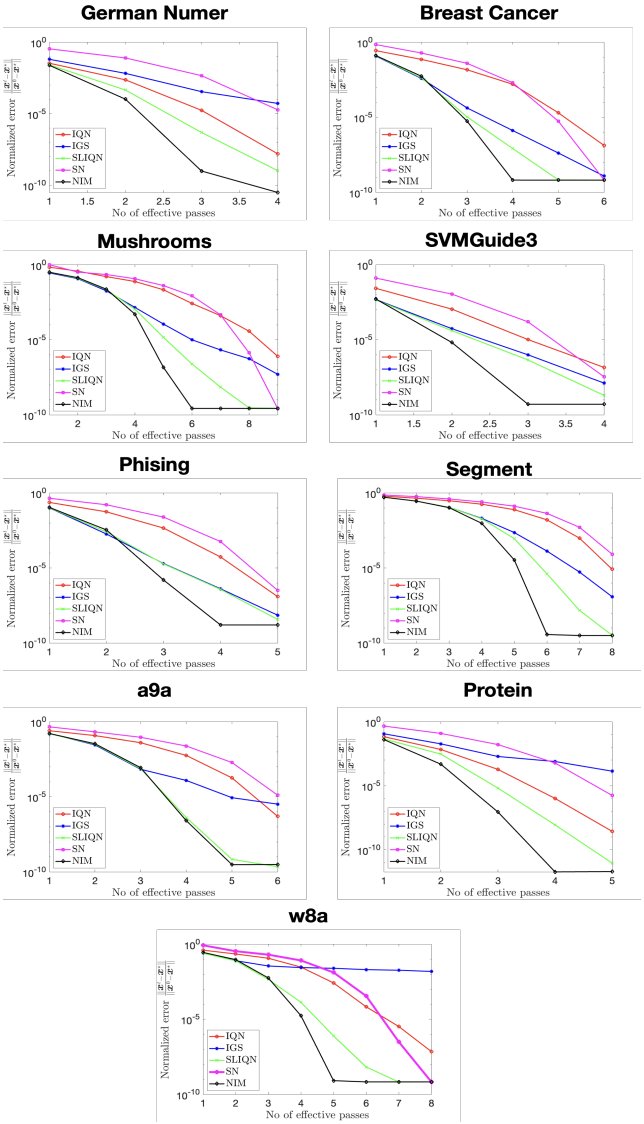


Figure 2: Normalized error vs. number of effective passes for regularized logistic loss minimization

7 CONCLUSION AND FUTURE WORK

We introduced the SLIQN method for minimizing finite-sum problems. SLIQN enjoys the best known incremental rate of $\mathcal{O}((1-\frac{\mu}{dL})^{t^2/n^2})$, has a $\mathcal{O}(d^2)$ per-iteration cost, an explicit superlinear convergence rate, and exhibits a superior empirical performance compared to several other incremental and stochastic QN methods. The key novelty is the construction of modified update rules using a clever multiplicative factor and a lazy propagation strategy. We back up our empirical results with a comprehensive theory that explains the superior performance of SLIQN. The convergence rate of SLIQN is locally superlinear; analyzing the global convergence of the proposed algorithm remains as a future work.

References

Byrd, R. H., Hansen, S. L., Nocedal, J., and Singer, Y. (2016). A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3).

Chang, D., Sun, S., and Zhang, C. (2019). An accelerated linearly convergent stochastic l-bfgs algorithm. *IEEE transactions on neural networks and learning systems*, 30(11):3338–3346.

Chen, J., Yuan, R., Garrigos, G., and Gower, R. M. (2022). San: Stochastic average newton algorithm for minimizing finite sums. In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 279–318. PMLR.

Defazio, A., Bach, F. R., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Neural Information Processing Systems*.

Derezinski, M. (2023). Stochastic variance-reduced newton: Accelerating finite-sum minimization with large batches. In *OPT 2023: Optimization for Machine Learning*.

Derezinski, M., Lacotte, J., Pilanci, M., and Mahoney, M. W. (2021). Newton-less: Sparsification without trade-offs for the sketched newton update. *Advances in Neural Information Processing Systems*, 34:2835–2847.

Gao, Z., Koppel, A., and Ribeiro, A. (2020). Incremental greedy bfgs: An incremental quasi-newton method with explicit superlinear rate. In *Adv. Neural Inf. Process. Syst. 12th OPT Workshop Optim. Mach. Learn.*

Gonen, A., Orabona, F., and Shalev-Shwartz, S. (2016). Solving ridge regression using sketched preconditioned svrg. In *International conference on machine learning*, pages 1397–1405. PMLR.

Jin, Q., Koppel, A., Rajawat, K., and Mokhtari, A. (2022). Sharpened quasi-newton methods: Faster superlinear rate and larger local convergence neighborhood. In *International Conference on Machine Learning*, pages 10228–10250. PMLR.

Jin, Q. and Mokhtari, A. (2022). Non-asymptotic superlinear convergence of standard quasi-newton methods. *Mathematical Programming*, pages 1–49.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26.

Kovalev, D., Mishchenko, K., and Richtárik, P. (2019). Stochastic newton and cubic newton methods with simple local linear-quadratic rates. *arXiv preprint arXiv:1912.01597*.

Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., and Lu, C. (2021). Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11025–11034.

Li, J., Zhou, P., Xiong, C., Socher, R., and Hoi, S. C. H. (2020). Prototypical contrastive learning of unsupervised representations. *CoRR*, abs/2005.04966.

Liu, Y., Feng, F., and Yin, W. (2019). Acceleration of svrg and katyusha x by inexact preconditioning. In *International Conference on Machine Learning*, pages 4003–4012. PMLR.

Mokhtari, A., Eisen, M., and Ribeiro, A. (2018). Iqn: An incremental quasi-newton method with local superlinear convergence rate. *SIAM Journal on Optimization*, 28(2):1670–1698.

Mokhtari, A. and Ribeiro, A. (2014). Res: Regularized stochastic bfgs algorithm. *IEEE Transactions on Signal Processing*, 62(23):6089–6104.

Mokhtari, A. and Ribeiro, A. (2015). Global convergence of online limited memory bfgs. *The Journal of Machine Learning Research*, 16(1):3151–3181.

Moritz, P., Nishihara, R., and Jordan, M. (2016). A linearly-convergent stochastic l-bfgs algorithm. In *Artificial Intelligence and Statistics*, pages 249–258. PMLR.

Nocedal, J. and Wright, S. J. (1999). *Numerical optimization*. Springer-Verlag, New York, NY, 2 edition.

Pilanci, M. and Wainwright, M. J. (2016). Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1):1842–1879.

Pilanci, M. and Wainwright, M. J. (2017). Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245.

Rodomanov, A. and Kropotov, D. (2016). A superlinearly-convergent proximal newton-type method for the optimization of finite sums. In *International Conference on Machine Learning*, pages 2597–2605. PMLR.

Rodomanov, A. and Nesterov, Y. (2021a). Greedy quasi-newton methods with explicit superlinear convergence. *SIAM Journal on Optimization*, 31(1):785–811.

Rodomanov, A. and Nesterov, Y. (2021b). New results on superlinear convergence of classical quasi-newton methods. *Journal of optimization theory and applications*, 188:744–769.

Rodomanov, A. and Nesterov, Y. (2021c). Rates of superlinear convergence for classical quasi-newton methods. *Mathematical Programming*, pages 1–32.

Song, Y. and Ermon, S. (2020). Improved techniques for training score-based generative models. *ArXiv*, abs/2006.09011.

Wu, J., Huang, W., Huang, J., and Zhang, T. (2018). Error compensated quantized sgd and its applications to large-scale distributed optimization. In *International Conference on Machine Learning*, pages 5325–5333. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, De-anonymized GitHub link for camera ready version has been provided]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [The error measure of *normalized error* is well defined. SLIQN is a deterministic algorithm, so error bars are not applicable.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [We have used a personal computer for all experiments]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

SUPPLEMENTARY MATERIAL

A ESTABLISHED RESULTS

Lemma A.1 (Banach's Lemma) *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be a matrix such that its norm satisfies $\|\boldsymbol{A}\| < 1$. Then the matrix $(\boldsymbol{I}+\boldsymbol{A})$ is invertible and*

$$\frac{1}{1+\|\boldsymbol{A}\|} < \left\|(\boldsymbol{I}+\boldsymbol{A})^{-1}\right\| < \frac{1}{1-\|\boldsymbol{A}\|}.$$

Proposition A.1 (Sherman-Morrison Formula) *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be an invertible matrix. Then, for all vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we have*

$$(\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^\mathsf{T})^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\mathsf{T}\boldsymbol{A}^{-1}}{1+\langle \boldsymbol{v}, \boldsymbol{A}^{-1}\boldsymbol{u}\rangle}. \tag{16}$$

Lemma A.2 (Lemma 2.1, Lemma 2.2 Rodomanov and Nesterov (2021c)) *Consider positive definite matrices $\boldsymbol{A}, \boldsymbol{G} \in \mathbb{R}^{d\times d}$ and suppose $\boldsymbol{G}_+ := \mathrm{BFGS}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u})$, where $\boldsymbol{u} \neq \boldsymbol{0}$. Then, the following results hold:*

1. *For any constants $\xi, \eta \geq 1$, we have*

$$\frac{1}{\xi}\boldsymbol{A} \preceq \boldsymbol{G} \preceq \eta\boldsymbol{A} \implies \frac{1}{\xi} \preceq \boldsymbol{G}_+ \preceq \eta\boldsymbol{A}.$$

2. *If $\boldsymbol{A} \preceq \boldsymbol{G}$, then we have*

$$\sigma(\boldsymbol{A}, \boldsymbol{G}) \geq \sigma(\boldsymbol{A}, \boldsymbol{G}_+).$$

Lemma A.3 (Lemma 4.2 Rodomanov and Nesterov (2021a)) *Suppose an objective function $f(\boldsymbol{x})$ is strongly self-concordant with constant $M > 0$. Consider $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $r := \|\boldsymbol{y}-\boldsymbol{x}\|_{\boldsymbol{x}}$, and $\boldsymbol{K} := \int_0^1 \nabla^2 f(\boldsymbol{x} + \tau(\boldsymbol{y}-\boldsymbol{x}))d\tau$. Then, we have that*

$$\begin{aligned}
\frac{\nabla^2 f(\boldsymbol{x})}{1+Mr} \preceq \nabla^2 f(\boldsymbol{y}) &\preceq (1+Mr)\nabla^2 f(\boldsymbol{x}),\\
\frac{\nabla^2 f(\boldsymbol{x})}{1+\frac{Mr}{2}} \preceq \boldsymbol{K} &\preceq (1+\frac{Mr}{2})\nabla^2 f(\boldsymbol{x}),\\
\frac{\nabla^2 f(\boldsymbol{y})}{1+\frac{Mr}{2}} \preceq \boldsymbol{K} &\preceq (1+\frac{Mr}{2})\nabla^2 f(\boldsymbol{y}).
\end{aligned}$$

Lemma A.4 (Theorem 2.5 Rodomanov and Nesterov (2021a)) *Consider positive definite matrices $\boldsymbol{A}, \boldsymbol{G} \in \mathbb{R}^{d\times d}$ such that $\boldsymbol{A} \preceq \boldsymbol{G}$ and $\mu\boldsymbol{I} \preceq \boldsymbol{A} \preceq L\boldsymbol{I}$ for constants $\mu, L > 0$. Suppose $\boldsymbol{G}_+ := \mathrm{BFGS}(\boldsymbol{G}, \boldsymbol{A}, \bar{\boldsymbol{u}}(\boldsymbol{G}, \boldsymbol{A}))$, where $\bar{\boldsymbol{u}}(\boldsymbol{G}, \boldsymbol{A})$ (4) is the greedy vector of $\boldsymbol{G}$ with respect to $\boldsymbol{A}$. Then, the following holds:*

$$\sigma(\boldsymbol{G}_+, \boldsymbol{A}) \leq \big(1 - \frac{\mu}{dL}\big)\sigma(\boldsymbol{G}, \boldsymbol{A}).$$

B SUPPORTING LEMMAS

Lemma B.1 *For all positive definite matrices $\boldsymbol{A}, \boldsymbol{G} \in \mathbb{R}^{d\times d}$, if $\boldsymbol{A} \preceq L\boldsymbol{I}$ and $\boldsymbol{A} \preceq \boldsymbol{G}$, then*

$$\|\boldsymbol{G}-\boldsymbol{A}\| \leq L\sigma(\boldsymbol{G},\boldsymbol{A}).$$

Proof: For any positive definite matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, let $\lambda_{\max}(\boldsymbol{X})$ denote its maximum eigenvalue and let $\sum_{i=1}^{d}\lambda_i(\boldsymbol{X})$ denote the sum of all of its eigenvalues. We can bound $\frac{1}{L}\|\boldsymbol{G}-\boldsymbol{A}\|$ as

$$\tfrac{1}{L}\|\boldsymbol{G}-\boldsymbol{A}\| \overset{\text{def}}{=} \tfrac{1}{L}\lambda_{\max}(\boldsymbol{G}-\boldsymbol{A}) \leq \tfrac{1}{L}\sum_{i=1}^{d}\lambda_i(\boldsymbol{G}-\boldsymbol{A}) \overset{\text{def}}{=} \tfrac{1}{L}\mathrm{Tr}(\boldsymbol{G}-\boldsymbol{A}).$$

Recall from the premise, we have $\boldsymbol{A} \preceq L\boldsymbol{I}$, which implies that $\boldsymbol{I} \preceq L\boldsymbol{A}^{-1}$. Therefore,

$$\sigma(\boldsymbol{G},\boldsymbol{A}) = \langle \boldsymbol{A}^{-1}, \boldsymbol{G}-\boldsymbol{A}\rangle \geq \tfrac{1}{L}\langle \boldsymbol{I}, \boldsymbol{G}-\boldsymbol{A}\rangle = \tfrac{1}{L}\mathrm{Tr}(\boldsymbol{G}-\boldsymbol{A}) \geq \tfrac{1}{L}\|\boldsymbol{G}-\boldsymbol{A}\|.$$

This completes the proof. □

Lemma B.2 *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a real valued function that is μ-strongly convex, L-smooth, and M-strongly self-concordant. Let $\boldsymbol{x}, \boldsymbol{x}_+ \in \mathbb{R}^d\backslash\{\boldsymbol{0}\}$ and $\boldsymbol{B}$ be a matrix such that $\boldsymbol{B} \succeq \nabla^2 f(\boldsymbol{x})$. Define the constant $r := \|\boldsymbol{x}_+ - \boldsymbol{x}\|_{\boldsymbol{x}}$ and the matrix $\boldsymbol{P} := (1+\frac{Mr}{2})^2\boldsymbol{B}$. Consider the following BFGS updates:*

$$\boldsymbol{Q} := \mathrm{BFGS}(\boldsymbol{P}, (1+\tfrac{Mr}{2})\boldsymbol{K}, \boldsymbol{x}_+ - \boldsymbol{x}),$$
$$\boldsymbol{B}_+ := \mathrm{BFGS}(\boldsymbol{Q}, \nabla^2 f(\boldsymbol{x}_+), \bar{\boldsymbol{u}}(\boldsymbol{Q}, \nabla^2 f(\boldsymbol{x}_+))).$$

Here, the matrix $\boldsymbol{K} := \int_0^1 \nabla^2 f(\boldsymbol{x}+\tau(\boldsymbol{x}_+-\boldsymbol{x}))d\tau$ and the vector $\bar{\boldsymbol{u}}(\boldsymbol{Q}, \nabla^2 f(\boldsymbol{x}_+))$ is the greedy vector 4. Then, $\boldsymbol{B}_+ \succeq \nabla^2 f(\boldsymbol{x}_+)$ and

$$\sigma(\boldsymbol{B}_+, \nabla^2 f(\boldsymbol{x}_+)) \leq \big(1-\tfrac{\mu}{dL}\big)\bigg(\big(1+\tfrac{Mr}{2}\big)^4\sigma(\boldsymbol{B}, \nabla^2 f(\boldsymbol{x})) + d\big(1+\tfrac{Mr}{2}\big)^4 - d\bigg).$$

Proof: We begin by analyzing the first BFGS update. Since $\boldsymbol{B} \succeq \nabla^2 f(\boldsymbol{x})$, we have the following:

$$\boldsymbol{B} \succeq \nabla^2 f(\boldsymbol{x}) \implies \boldsymbol{P} \succeq \big(1+\frac{Mr}{2}\big)^2 f(\boldsymbol{x}) \overset{Lem.A.3}{\succeq} \big(1+\frac{Mr}{2}\big)\boldsymbol{K}.$$

Since $\boldsymbol{P} \succeq \big(1+\frac{Mr}{2}\big)\boldsymbol{K}$, the metric $\sigma\big(\boldsymbol{P}, (1+\frac{Mr}{2})\boldsymbol{K}\big)$ is well defined. Applying Lemma A.2, we obtain

$$\boldsymbol{Q} = \mathrm{BFGS}(\boldsymbol{P}, \big(1+\frac{Mr}{2}\big)\boldsymbol{K}, \boldsymbol{x}_+ - \boldsymbol{x}) \succeq \big(1+\frac{Mr}{2}\big)\boldsymbol{K}.$$

Applying Lemma A.3 to relate $\boldsymbol{K}$ and $\nabla^2 f(\boldsymbol{x}_+)$, we obtain

$$\boldsymbol{Q} \succeq \big(1+\frac{Mr}{2}\big)\boldsymbol{K} \overset{Lem.A.3}{\succeq} \nabla^2 f(\boldsymbol{x}_+).$$

We now begin analyzing the second BFGS update. Since $\boldsymbol{Q} \succeq \nabla^2 f(\boldsymbol{x}_+)$, applying Lemma A.2, we obtain

$$\boldsymbol{B}_+ = \mathrm{BFGS}(\boldsymbol{Q}, \nabla^2 f(\boldsymbol{x}_+), \bar{\boldsymbol{u}}(\boldsymbol{Q}, \nabla^2 f(\boldsymbol{x}_+)) \succeq \nabla^2 f(\boldsymbol{x}_+),$$

which completes the proof of the first part. Applying Lemma A.4, we obtain

$$\sigma(\boldsymbol{B}_+, \nabla^2 f(\boldsymbol{x}_+)) \leq \big(1-\tfrac{\mu}{dL}\big)\,\sigma(\boldsymbol{Q}, \nabla^2 f(\boldsymbol{x}_+)).$$

Define $c := \big(1-\frac{\mu}{dL}\big)$ for brevity. We are now ready to show the second result. Observe that

$$\begin{aligned}
\sigma(\boldsymbol{B}_+, \nabla^2 f(\boldsymbol{x}_+)) &\leq (1-c)\sigma(\boldsymbol{Q}, \nabla^2 f(\boldsymbol{x}_+)),\\
&= (1-c)(\langle \nabla^2 f(\boldsymbol{x}_+)^{-1}, \boldsymbol{Q}\rangle - d),\\
&\overset{(a)}{\leq} (1-c)\big(\big(1+\tfrac{Mr}{2}\big)\langle \boldsymbol{K}^{-1}, \boldsymbol{Q}\rangle - d\big),\\
&\overset{(b)}{=} (1-c)\big(\big(1+\tfrac{Mr}{2}\big)^2\langle \tilde{\boldsymbol{K}}^{-1}, \boldsymbol{Q}\rangle - d\big),
\end{aligned} \tag{17}$$

where (a) follows since

$$\big(1+\tfrac{Mr}{2}\big)\nabla^2 f(\boldsymbol{x}_+) \overset{Lem.A.3}{\succeq} \boldsymbol{K} \iff \nabla^2 f(\boldsymbol{x}_+)^{-1} \preceq \big(1+\frac{Mr}{2}\big)\boldsymbol{K}^{-1}.$$

In (b), we have defined $\tilde{\boldsymbol{K}} \coloneqq \big(1+\tfrac{Mr}{2}\big)\boldsymbol{K}$. Recall that we have already established $\boldsymbol{P} \succeq \tilde{\boldsymbol{K}}$. Applying Lemma A.2, we obtain

$$\sigma(\boldsymbol{Q}, \tilde{\boldsymbol{K}}) \leq \sigma(\boldsymbol{P}, \tilde{\boldsymbol{K}}) \iff \sigma(\boldsymbol{Q}, \boldsymbol{K}) \leq \sigma(\boldsymbol{P}, \boldsymbol{K}).$$

Continuing from (17), we obtain

$$\begin{aligned}
\sigma(\boldsymbol{B}_+, \nabla^2 f(\boldsymbol{x}_+)) &\leq (1-c)\big(\big(1+\tfrac{Mr}{2}\big)^2\langle \tilde{\boldsymbol{K}}^{-1}, \boldsymbol{P}\rangle - d\big),\\
&= (1-c)\big(\big(1+\tfrac{Mr}{2}\big)^4\langle \tilde{\boldsymbol{K}}^{-1}, \boldsymbol{B}\rangle - d\big),\\
&= (1-c)\big(\big(1+\tfrac{Mr}{2}\big)^3\langle \boldsymbol{K}^{-1}, \boldsymbol{B}\rangle - d\big),\\
&\overset{(a)}{\leq} (1-c)\big(\big(1+\tfrac{Mr}{2}\big)^4\langle \nabla^2 f(\boldsymbol{x})^{-1}, \boldsymbol{B}\rangle - d\big),
\end{aligned} \tag{18}$$

where (a) follows from $\boldsymbol{K}^{-1} \preceq \big(1+\tfrac{Mr}{2}\big)\nabla^2 f(\boldsymbol{x})^{-1}$ ($\because$ Lemma A.3). Finally, setting $\langle \nabla^2 f(\boldsymbol{x})^{-1}, \boldsymbol{B}\rangle = \sigma(\boldsymbol{B}, \nabla^2 f(\boldsymbol{x})) + d$ completes the proof. □

Corollary B.1 *Under the notation established in Lemma B.2, let $\alpha \in \mathbb{R}_+$ be an upper bound on r. Then, we have the following:*

$$\sigma(\boldsymbol{B}_+, \nabla^2 f(\boldsymbol{x}_+)) \leq (1-\tfrac{\mu}{dL})e^{2M\alpha}\big(\sigma(\boldsymbol{B}, \nabla^2 f(\boldsymbol{x})) + 2Md\alpha\big).$$

Proof: From Lemma B.2, we have

$$\begin{aligned}
\sigma(\boldsymbol{B}_+, \nabla^2 f(\boldsymbol{x}_+)) &\leq \big(1-\tfrac{\mu}{dL}\big)\big(1+\tfrac{M\alpha}{2}\big)^4\big(\sigma(\boldsymbol{B}, \nabla^2 f(\boldsymbol{x})) + d\big(1-\tfrac{1}{\big(1+\frac{M\alpha}{2}\big)^4}\big)\big),\\
&\overset{(a)}{\leq} \big(1-\tfrac{\mu}{dL}\big)e^{2M\alpha}\big(\sigma(\boldsymbol{B}, \nabla^2 f(\boldsymbol{x})) + d(1-e^{-2M\alpha})\big),\\
&\overset{(b)}{\leq} \big(1-\tfrac{\mu}{dL}\big)e^{2M\alpha}\big(\sigma(\boldsymbol{B}, \nabla^2 f(\boldsymbol{x})) + 2Md\alpha\big),
\end{aligned}$$

where (a) follows by the inequality $1+x \leq e^x, \forall x \in \mathbb{R}$ and (b) follows from the inequality $1-e^{-x} \leq x, \forall x > 0$. This completes the proof. □

Lemma B.3 *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a real valued function that is μ-strongly convex, L-smooth, and M-strongly self-concordant. Let $\tilde{\boldsymbol{x}} \in \mathbb{R}^d$ be some fixed vector and $0 \leq \gamma < 1$ be some fixed constant such that the sequence $\{\boldsymbol{x}^k\}$, for all $k \in [T]$ satisfies*

$$\left\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\right\| \leq \gamma^k \left\|\boldsymbol{x}^0 - \tilde{\boldsymbol{x}}\right\|.$$

Define the constant $r_k \coloneqq \left\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\right\|_{\boldsymbol{x}^{k-1}}$ for every k. Let $\boldsymbol{B}^0$ be a matrix such that it satisfies $\boldsymbol{B}^0 \succeq \nabla^2 f(\boldsymbol{x}^0)$. Consider the following BFGS updates:

$$\begin{aligned}
\boldsymbol{Q}^k &\coloneqq \mathrm{BFGS}(\boldsymbol{P}^{k-1}, (1+\tfrac{Mr_k}{2})\boldsymbol{K}^k, \boldsymbol{x}^k - \boldsymbol{x}^{k-1}),\\
\boldsymbol{B}^k &\coloneqq \mathrm{BFGS}(\boldsymbol{Q}^k, \nabla^2 f(\boldsymbol{x}^k), \bar{\boldsymbol{u}}(\boldsymbol{Q}^k, \nabla^2 f(\boldsymbol{x}^k))),
\end{aligned}$$

where $\boldsymbol{P}^{k-1} \coloneqq \big(1+\tfrac{Mr_k}{2}\big)^2\boldsymbol{B}^{k-1}$, $\boldsymbol{K}^k \coloneqq \int_0^1 \nabla^2 f(\boldsymbol{x}^{k-1} + \tau(\boldsymbol{x}^k - \boldsymbol{x}^{k-1}))d\tau$, and $\bar{\boldsymbol{u}}(\boldsymbol{Q}^k, \nabla^2 f(\boldsymbol{x}^k))$ is the greedy vector 4. Then, the following holds for all $k \in [T]$:

$$\sigma(\boldsymbol{B}^k, \nabla^2 f(\boldsymbol{x}^k)) \leq (1-\tfrac{\mu}{dL})^k e^{\frac{4M\sqrt{L}||\boldsymbol{x}^0 - \tilde{\boldsymbol{x}}||}{1-\gamma}}\left(\sigma(\boldsymbol{B}^0, \nabla^2 f(\boldsymbol{x}^0)) + \left\|\boldsymbol{x}^0 - \tilde{\boldsymbol{x}}\right\| \tfrac{4Md\sqrt{L}}{1-(1-\frac{\mu}{dL})^{-1}\gamma}\right). \tag{19}$$

Proof: From Lemma B.2, it can be shown that $\boldsymbol{B}^k \succeq \nabla^2 f(\boldsymbol{x}^k)$ for $k = 1, \ldots, T$. Therefore, $\sigma(\boldsymbol{B}^k, \nabla^2 f(\boldsymbol{x}^k))$ is well defined. We introduce the notation $\sigma_k := \sigma(\boldsymbol{B}^k, \nabla^2 f(\boldsymbol{x}^k))$, $d_k := \left\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\right\|$, and $c := \frac{\mu}{dL}$ for simplicity.

To apply Corollary B.1, we need an upper bound on r_k. This is trivial via the Triangle inequality

$$\begin{aligned} r_k = \left\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\right\|_{\boldsymbol{x}^{k-1}} &\overset{\Delta}{\leq} \left\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\right\|_{\boldsymbol{x}^{k-1}} + \left\|\boldsymbol{x}^{k-1} - \tilde{\boldsymbol{x}}\right\|_{\boldsymbol{x}^{k-1}} \\ &\overset{(a)}{\leq} \sqrt{L}\big(\left\|\boldsymbol{x}^k - \tilde{\boldsymbol{x}}\right\| + \left\|\boldsymbol{x}^{k-1} - \tilde{\boldsymbol{x}}\right\|\big) \leq 2\sqrt{L}\gamma^{k-1}\left\|\boldsymbol{x}^0 - \tilde{\boldsymbol{x}}\right\|, \end{aligned}$$

where (a) follows since $\nabla^2 f(\boldsymbol{x}) \preceq LI$. Therefore $\alpha_k = 2\sqrt{L}\gamma^{k-1}\left\|\boldsymbol{x}^0 - \tilde{\boldsymbol{x}}\right\|$ is an upper bound on r_k. Applying Corollary B.1, we obtain

$$\begin{aligned} \sigma_k &\leq (1-c)e^{2M\alpha_k}(\sigma_{k-1} + 2Md\alpha_k) \\ &\leq (1-c)e^{4M\sqrt{L}d_0\gamma^{k-1}}(\sigma_{k-1} + 4Md\sqrt{L}d_0\gamma^{k-1}). \end{aligned} \tag{20}$$

We solve the recursion (20) as follows:

$$\begin{aligned} \sigma_k &\leq (1-c)e^{4M\sqrt{L}d_0\gamma^{k-1}}(\sigma_{k-1} + 4Md\sqrt{L}d_0\gamma^{k-1}), \\ &\leq (1-c)^2 e^{4M\sqrt{L}d_0(\gamma^{k-1}+\gamma^{k-2})}\sigma_{k-2} + 4Md\sqrt{L}d_0\big(\gamma^{k-2}(1-c)^2 e^{4M\sqrt{L}d_0(\gamma^{k-1}+\gamma^{k-2})} \\ &\qquad + \gamma^{k-1}(1-c)e^{4M\sqrt{L}d_0\gamma^{k-1}}\big), \\ &\leq (1-c)^k e^{4M\sqrt{L}d_0\sum_{j=0}^{k-1}\gamma^j}\sigma_0 + 4Md\sqrt{L}d_0\big(\sum_{j=0}^{k-1}\gamma^j(1-c)^{k-j}e^{4M\sqrt{L}d_0\sum_{i=1}^{k-j}\gamma^{k-i}}\big), \\ &\leq (1-c)^k e^{4M\sqrt{L}d_0\sum_{j=0}^{\infty}\gamma^j}\sigma_0 + 4Md\sqrt{L}d_0 e^{4M\sqrt{L}d_0\sum_{i=0}^{\infty}\gamma^i}\sum_{j=0}^{k-1}\gamma^j(1-c)^{k-j}, \\ &\leq (1-c)^k e^{\frac{4M\sqrt{L}d_0}{1-\gamma}}\big(\sigma_0 + \tfrac{4Md\sqrt{L}d_0}{1-\frac{\gamma}{1-c}}\big). \end{aligned}$$

This completes the proof. □

Remark 3 *It can be concluded from the proof of Lemma B.3 that redefining $r_k := 2\sqrt{L}\gamma^{k-1}d_0$ (which is an upper bound on $\left\|\boldsymbol{x}^k - \boldsymbol{x}^{k-1}\right\|_{\boldsymbol{x}^{k-1}}$), the results of Lemma B.3 remain unchanged.*

C EFFICIENT IMPLEMENTATION OF IQN

Recall that, that at time t, IQN updates

$$\boldsymbol{B}_{i_t}^t = \text{BFGS}(\boldsymbol{B}_{i_t}^{t-1}, \boldsymbol{K}^t, \boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}).$$

Define $\boldsymbol{\phi}^t := \sum_{i=1}^n \boldsymbol{B}_i^t \boldsymbol{z}_i^t - \sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^t)$. At time t, since IQN only updates the tuple with index i_t, we have

$$\boldsymbol{\phi}^t = \boldsymbol{\phi}^{t-1} + \big(\boldsymbol{B}_{i_t}^t \boldsymbol{z}_{i_t}^t - \boldsymbol{B}_{i_t}^{t-1} \boldsymbol{z}_{i_t}^{t-1}\big) - \big(\nabla f_{i_t}(\boldsymbol{z}_{i_t}^t) - \nabla f_{i_t}(\boldsymbol{z}_{i_t}^{t-1})\big). \tag{21}$$

Therefore, given access to $\boldsymbol{\phi}^{t-1}$, we can compute $\boldsymbol{\phi}^t$ in $\mathcal{O}(d^2)$ time. This updating scheme can be implemented iteratively, where we only evaluate $\boldsymbol{\phi}^0$ explicitly and evaluate $\boldsymbol{\phi}^t$, for all $t \geq 1$ by (21). It only remains to compute $\big(\bar{\boldsymbol{B}}^t\big)^{-1}$, where $\bar{\boldsymbol{B}}^t := \big(\sum_{i=1}^n \boldsymbol{B}_i^t\big)$. This can be done by applying the Sherman-Morrison formula (16) twice to the matrix on the right (22).

$$\bar{\boldsymbol{B}}^t = \bar{\boldsymbol{B}}^{t-1} + \boldsymbol{B}_{i_t}^t - \boldsymbol{B}_{i_t}^{t-1} = \bar{\boldsymbol{B}}^{t-1} + \frac{\boldsymbol{y}_{i_t}^t \boldsymbol{y}_{i_t}^{t\,\mathsf{T}}}{\langle \boldsymbol{y}_{i_t}^t, \boldsymbol{s}_{i_t}^t \rangle} - \frac{\boldsymbol{B}_{i_t}^{t-1} \boldsymbol{s}_{i_t}^t \boldsymbol{s}_{i_t}^{t\,\mathsf{T}} \boldsymbol{B}_{i_t}^{t-1}}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{B}_{i_t}^{t-1} \boldsymbol{s}_{i_t}^t \rangle}, \tag{22}$$

where $\boldsymbol{y}_{i_t}^t = \boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}$, $\boldsymbol{s}_{i_t}^t = \nabla f_{i_t}(\boldsymbol{z}_{i_t}^t) - \nabla f_{i_t}(\boldsymbol{z}_{i_t}^{t-1})$. Applying(16) twice, we get (23) and (24)

$$(\bar{\boldsymbol{B}}^t)^{-1} = \boldsymbol{Z}^t + \frac{\boldsymbol{Z}^t (\boldsymbol{B}_{i_t}^{t-1} \boldsymbol{s}_{i_t}^t)(\boldsymbol{B}_{i_t}^{t-1} \boldsymbol{s}_{i_t}^t)^{\mathsf{T}} \boldsymbol{Z}^t}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{B}_{i_t}^{t-1} \boldsymbol{s}_{i_t}^t \rangle - \langle (\boldsymbol{B}_{i_t}^{t-1} \boldsymbol{s}_{i_t}^t), \boldsymbol{Z}^t (\boldsymbol{B}_{i_t}^{t-1} \boldsymbol{s}_{i_t}^t) \rangle}, \tag{23}$$

where $\boldsymbol{Z}^t$ is given by

$$\boldsymbol{Z}^t = (\bar{\boldsymbol{B}}^{t-1})^{-1} - \frac{(\bar{\boldsymbol{B}}^{t-1})^{-1} \boldsymbol{y}_{i_t}^t \boldsymbol{y}_{i_t}^{t\,\mathsf{T}} (\bar{\boldsymbol{B}}^{t-1})^{-1}}{\langle \boldsymbol{y}_{i_t}^t, \boldsymbol{s}_{i_t}^t \rangle + \langle \boldsymbol{y}_{i_t}^t, (\bar{\boldsymbol{B}}^{t-1})^{-1} \boldsymbol{y}_{i_t}^t \rangle}. \tag{24}$$

By iteratively implementing this scheme, we only need to compute $(\bar{\boldsymbol{B}}^0)^{-1}$ explicitly and $(\bar{\boldsymbol{B}}^t)^{-1}$, for all $t \geq 1$ by the Sherman-Morrison formula.

D LOW MEMORY IMPLEMENTATION OF INCREMENTAL METHODS

In this section, we illustrate how incremental methods can be effectively implemented for large-scale real-world scenarios which are characterized by substantial memory requirements of $\mathcal{O}(nd^2)$. Our solution leverages the disk, which offers significantly larger storage capacity compared to main memory but comes with an increased load-store latency. To mitigate this latency issue, we employ a pipelining scheme. In this scheme, we partition the data into blocks and simultaneously run compute operations on one block while performing load-store operations on the blocks adjacent to it. This parallelization effectively extends the main memory capacity to the available the disk size, all the while avoiding its larger latency.

Formally, let the available main memory capacity be g GB, the number of data samples be n, the dimensionality of the data be d, and the space requirement for each sample be s. We assume that the disk is sufficiently large to store the data for all samples, that it the size of the disk is greater than ns. We divide the data into $m = \frac{2ns}{g}$ blocks, denoted as b_i for $i \in [m]$. This choice of m ensures that two blocks can be accommodated in memory simultaneously. The processing proceeds as follows:

1. We assume that the memory holds blocks b_1 and b_2, along with the global memoized quantities for SLIQN. All data blocks are also stored on the disk.
2. At iteration $t = 1$, we process the block b_1 by executing the corresponding algorithm updates on it.
3. At any iteration $t > 1$, we execute the algorithm updates on $b_{i\%n+1}$ while concurrently storing the already processed block b_i back into the disk and loading the block $b_{(i+1)\%n+1}$ into memory to be processed next.

In practice, modern disks have access speeds of around 500 MBps, making processing the bottleneck in this parallel architecture, rather than disk access. For example, in our implementation of SLIQN with $g = 1200$ MB, $n = 20,000$, $d = 123$, $s = 0.1117$ MB, and $m = 4$, we observed that processing one block took 7.8 seconds, while the load-store operation required only 3.8 seconds.

E CONVERGENCE ANALYSIS OF SIQN

In this section, we provide the convergence analysis of SIQN, as the convergence analysis motivates the replacement of β_t with $\alpha_{\lceil \frac{t}{n}-1 \rceil}$. We begin by showing that at each time t, the matrix $\boldsymbol{Q}^t$ obtained after the first BFGS update satisfies $\boldsymbol{Q}^t \succeq \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)$. Using this, we show that the updated Hessian approximation satisfies $\boldsymbol{B}_{i_t}^t \succeq \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)$. These observations are essential in order ensure that $\sigma(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))$ and $\sigma(\boldsymbol{B}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))$ are well defined.

Lemma E.1 *For all $t \geq 1$ the following hold true:*

$$\boldsymbol{Q}^t \succeq \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t), \tag{25}$$

$$\boldsymbol{B}_i^t \succeq \nabla^2 f_i(\boldsymbol{z}_i^t), \tag{26}$$

for all $i \in [n]$.

Proof: The proof follows by induction on t.

Base case: At $t = 0$, (26) holds due to the initialization.

Induction Hypothesis (IH): Assume that (25) and (26) hold for $t = m-1$, for some $m \geq 1$. We prove that (25) and (26) hold for $t = m$ in the Induction step.

Induction step: Since $\boldsymbol{B}_{i_m}^{m-1} \succeq \nabla^2 f_{i_m}(\boldsymbol{z}_{i_m}^{m-1})$, we have the following:

$$\big(1+\beta_m\big)^2 \boldsymbol{B}_{i_m}^{m-1} \succeq \big(1+\beta_m\big)^2 \nabla^2 f_{i_m}(\boldsymbol{z}_{i_m}^{m-1}) \overset{Lem.A.3}{\succeq} \big(1+\beta_m\big)\boldsymbol{K}^m,$$

where recall that $\beta_m = \frac{M}{2}\left\|\boldsymbol{z}_{i_m}^m - \boldsymbol{z}_{i_m}^{m-1}\right\|_{\boldsymbol{z}_{i_m}^{m-1}}$. Applying Lemma A.2, we obtain

$$\boldsymbol{Q}^m = \mathrm{BFGS}(\big(1+\beta_m\big)^2 \boldsymbol{B}_{i_m}^{m-1}, \big(1+\beta_m\big)\boldsymbol{K}^m, \boldsymbol{z}_{i_m}^m - \boldsymbol{z}_{i_m}^{m-1}) \overset{Lem.A.2}{\succeq} \big(1+\beta_m\big)\boldsymbol{K}^m.$$

Applying Lemma A.3 to relate $\boldsymbol{K}^m$ and $\nabla^2 f_{i_m}(\boldsymbol{z}_{i_m}^m)$, we obtain

$$\boldsymbol{Q}^m \succeq \big(1+\beta_m\big)\boldsymbol{K}^m \overset{Lem.A.3}{\succeq} \nabla^2 f_{i_m}(\boldsymbol{z}_{i_m}^m).$$

Therefore, (25) holds for $t = m$. Since $\boldsymbol{Q}^m \succeq \nabla^2 f_{i_m}(\boldsymbol{z}_{i_m}^m)$, applying Lemma A.2, we obtain

$$\boldsymbol{B}_{i_m}^m = \mathrm{BFGS}(\boldsymbol{Q}^m, \nabla^2 f_{i_m}(\boldsymbol{z}_{i_m}^m), \bar{\boldsymbol{u}}(\boldsymbol{Q}^m, \nabla^2 f_{i_m}(\boldsymbol{z}_{i_m}^m)) \overset{Lem.A.2}{\succeq} \nabla^2 f_{i_m}(\boldsymbol{z}_{i_m}^m).$$

Therefore, (26) holds for $t = m$. This completes the induction step. The proof is hence complete by induction. □

Key steps: We establish the convergence guarantees in three steps: Lemma E.2 establishes a one-step inequality that bounds the residual $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\|$ in terms of the previous residuals $\left\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\right\|$ and the norm error in the Hessian approximation $\left\|\boldsymbol{B}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\right\|$, for all $i \in [n]$. Lemma E.3 uses the result of Lemma E.2 to inductively show that both the residual $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\|$ and the Hessian approximation error $\sigma(\boldsymbol{B}_{i_t}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))$, decrease linearly with $\lceil t/n \rceil$. Using the result of Lemma E.2 and Lemma E.3, Lemma E.4 establishes a mean-superlinear convergence result. We finally show in Theorem 1 that the residuals can be upper bounded by a superlinearly convergent sequence.

We now present our one-step inequality. A similar inequality with $\boldsymbol{B}_i^{t-1}$ replaced with $\boldsymbol{D}_i^{t-1}$ also appears in Lemma 1 (for SLIQN). Since the proofs are identical, we refer Appendix G.1 directly for the proof of Lemma E.2.

Lemma E.2 *If Assumptions **A1** and **A2** hold, the sequence of iterates generated by SIQN satisfy*

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\| \leq \tfrac{\tilde{L}\Gamma^{t-1}}{2}\sum_{i=1}^n \left\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\right\|^2 + \Gamma^{t-1}\sum_{i=1}^n \left\|\boldsymbol{B}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\right\| \left\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\right\|, \tag{27}$$

for all $t \geq 1$, where $\Gamma^t := \left\|\left(\sum_{i=1}^n \boldsymbol{B}_i^t\right)^{-1}\right\|$.

Proof: Refer Appendix G.1. □

Lemma E.3 *If Assumptions **A1** and **A2** hold, for any ρ such that $0 < \rho < 1 - \frac{\mu}{dL}$, there exist positive constants ϵ^{siqn} and σ_0^{siqn} such that if $\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| \leq \epsilon^{siqn}$ and $\sigma(\boldsymbol{B}_i^0, \nabla^2 f_i(\boldsymbol{x}^0)) \leq \sigma_0^{siqn}$ for all $i \in [n]$, the sequence of iterates generated by SIQN satisfy*

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\| \leq \rho^{\lceil \frac{t}{n} \rceil} \left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|. \tag{28}$$

Further, it holds that

$$\sigma(\boldsymbol{B}_{i_t}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)) \leq \left(1 - \tfrac{\mu}{dL}\right)^{\lceil \frac{t}{n} \rceil} \delta^{siqn}, \tag{29}$$

where $\delta^{siqn} := e^{\frac{4M\sqrt{L}\epsilon^{siqn}}{1-\rho}} \left(\sigma_0^{siqn} + \epsilon^{siqn} \frac{4Md\sqrt{L}}{1 - \frac{\rho}{1-\frac{\mu}{dL}}}\right)$, $M = \tilde{L}/\mu^{\frac{3}{2}}$.

Proof: For a given ρ that satisfies $0 < \rho < 1 - \frac{\mu}{dL}$, let the variables $\epsilon^{siqn}, \sigma_0^{siqn}$ be chosen to satisfy

$$\frac{\frac{\tilde{L}\epsilon^{siqn}}{2} + L\delta^{siqn}}{\mu} \leq \frac{\rho}{1+\rho} < 1. \tag{30}$$

Remark 4 *Indeed there exists positive constants $\epsilon^{siqn}, \sigma_0^{siqn}$ that satisfy (30). Recall from the premise that δ is a function of $\epsilon^{siqn}, \sigma_0^{siqn}$, and we can define the left-hand-side of (30) as a function $g(\epsilon^{siqn}, \sigma_0^{siqn})$ as*

$$g(\epsilon^{siqn}, \sigma_0^{siqn}) := \frac{\frac{\tilde{L}\epsilon^{siqn}}{2} + L\delta^{siqn}}{\mu} = \frac{\frac{\tilde{L}\epsilon^{siqn}}{2} + Le^{\frac{4M\sqrt{L}\epsilon^{siqn}}{1-\rho}} \left(\sigma_0^{siqn} + \epsilon^{siqn} \frac{4Md\sqrt{L}}{1 - \frac{\rho}{1-\frac{\mu}{dL}}}\right)}{\mu}.$$

Fix $\sigma_0^{siqn} = \frac{\mu}{2L}(\frac{\rho}{1+\rho}) > 0$. It is easy to see that $g(\epsilon^{siqn}, \sigma_0^{siqn})$ is continuous and monotonically increasing in ϵ^{siqn}. Also, note that $g(0, \frac{\mu}{2L}(\frac{\rho}{1+\rho})) = \frac{\rho}{2(1+\rho)}$ and $\lim_{\epsilon \to \infty} g(\epsilon, \frac{\mu}{2L}(\frac{\rho}{1+\rho})) = \infty$. We can therefore apply the Intermediate Value Theorem (IVT) to guarantee that there exists $\epsilon > 0$ such that $g(\epsilon, \frac{\mu}{2L}(\frac{\rho}{1+\rho})) \leq \frac{\rho}{1+\rho}$.

Base case: At $t = 1$, applying Lemma E.2, we have

$$\left\|\boldsymbol{x}^1 - \boldsymbol{x}^\star\right\| \leq \frac{\tilde{L}\Gamma^0}{2} \sum_{i=1}^n \left\|\boldsymbol{z}_i^0 - \boldsymbol{x}^\star\right\|^2 + \Gamma^0 \sum_{i=1}^n \left\|\boldsymbol{B}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\| \left\|\boldsymbol{z}_i^0 - \boldsymbol{x}^\star\right\|.$$

Since $\boldsymbol{B}_i^0 \succeq \nabla^2 f_i(\boldsymbol{z}_i^0)$ and $\nabla^2 f_i(\boldsymbol{z}_i^0) \preceq L\boldsymbol{I}$, applying Lemma B.1, we have $\sigma(\boldsymbol{B}_i^0, \nabla^2 f_i(\boldsymbol{z}_i^0)) \geq \frac{1}{L}\left\|\boldsymbol{B}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\|$. This gives

$$\left\|\boldsymbol{x}^1 - \boldsymbol{x}^\star\right\| \leq n\Gamma^0 \left(\frac{\tilde{L}\epsilon^{siqn}}{2} + L\sigma_0^{siqn}\right) \left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|,$$

where we have used $\boldsymbol{z}_i^0 = \boldsymbol{x}^0$ and $\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| \leq \epsilon^{siqn}, \sigma(\boldsymbol{B}_i^0, \nabla^2 f_i(\boldsymbol{z}_i^0)) \leq \sigma_0^{siqn}$.

We now bound Γ^0. Define $\boldsymbol{X}^0 := \frac{1}{n}\sum_{i=1}^n \boldsymbol{B}_i^0, \boldsymbol{Y}^0 := \frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(\boldsymbol{z}_i^0)$. We have the following:

$$\frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{B}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\| \overset{\Delta}{\geq} \left\|\boldsymbol{X}^0 - \boldsymbol{Y}^0\right\| = \left\|(\boldsymbol{Y}^0)((\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I})\right\| \overset{(a)}{\geq} \mu \left\|(\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I}\right\|,$$

where (a) follows since $\boldsymbol{Y}^0 = \frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(\boldsymbol{z}_i^0) \succeq \mu\boldsymbol{I}$ ($\because$ Assumption **A1**). This gives us

$$\left\|(\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I}\right\| \leq \frac{L\sigma_0^{siqn}}{\mu} \leq \frac{L\delta^{siqn}}{\mu} \overset{(30)}{<} \frac{\rho}{1+\rho}.$$

We can now upper bound Γ^0 using Banach's Lemma A.1. Consider the matrix $(\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I}$ and note that it satisfies the requirement of A.1 from the above result. Therefore, we have

$$\left\|(\boldsymbol{X}^0)^{-1}\boldsymbol{Y}^0\right\| = \left\|\left(\boldsymbol{I} + ((\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0) - \boldsymbol{I})\right)^{-1}\right\| \overset{Lem.A.1}{\leq} \frac{1}{1 - \left\|(\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I}\right\|} \leq 1 + \rho.$$

Recall that $\mu \boldsymbol{I} \preceq \boldsymbol{Y}^0$. Using this and the previous result, we can upper bound $\left\|(\boldsymbol{X}^0)^{-1}\right\|$ as $\mu \left\|(\boldsymbol{X}^0)^{-1}\right\| \leq \left\|(\boldsymbol{X}^0)^{-1}\boldsymbol{Y}^0\right\| \leq 1+\rho$. This gives us, $\Gamma^0 = \frac{1}{n}\left\|(\boldsymbol{X}^0)^{-1}\right\| \leq \frac{1+\rho}{n\mu}$. Therefore, we have the following bound on $\left\|\boldsymbol{x}^1 - \boldsymbol{x}^\star\right\|$:

$$\left\|\boldsymbol{x}^1 - \boldsymbol{x}^\star\right\| \leq \frac{\frac{\tilde{L}\epsilon^{siqn}}{2} + L\sigma_0^{siqn}}{\mu}(1+\rho)\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| \leq \frac{\frac{\tilde{L}\epsilon^{siqn}}{2} + L\delta^{siqn}}{\mu}(1+\rho)\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| \overset{(30)}{\leq} \rho\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|.$$

By the updates performed by Algorithm 1, we have $\boldsymbol{z}_1^1 = \boldsymbol{x}^1$ and $\boldsymbol{z}_i^1 = \boldsymbol{x}^0$ for $i \neq 1$. Applying Lemma B.3, we obtain

$$\sigma(\boldsymbol{B}_1^1, \nabla^2 f_1(\boldsymbol{z}_1^1)) \leq \big(1-c\big)e^{\frac{4M\sqrt{L}\epsilon^{siqn}}{1-\rho}}\big(\underbrace{\sigma(\boldsymbol{B}_1^0, \nabla^2 f_1(\boldsymbol{z}_1^0))}_{\leq \sigma_0^{siqn}} + \epsilon^{siqn}\frac{4Md\sqrt{L}}{1-\frac{\rho}{1-c}}\big) \leq (1-c)\delta^{siqn}.$$

This completes the proof for $t = 1$.

Induction Hypothesis (IH): Let (28) and (29) hold for $t \in [jn+m]$, where $j \geq 0, 0 \leq m < n$.

Induction step: We then prove that (28) and (29) also hold for $t = jn+m+1$. Recall that the tuples are updated in a deterministic cyclic order, and at the current time t, we are in the j^{th} cycle and have updated the m^{th} tuple. Therefore, it is easy to note that $\boldsymbol{z}_i^{jn+m} = \boldsymbol{x}^{jn+i}$, for all $i \in [m]$, which refer to the tuples updated in this cycle, and $\boldsymbol{z}_i^{jn+m} = \boldsymbol{x}^{jn-n+i}$ for all $i \in [n]\backslash[m]$, which refer to the tuples updated in the previous cycle. From the induction hypothesis, we have

$$\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| \leq \begin{cases} \rho^{\lceil \frac{jn+i}{n} \rceil}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| & i \in [m], \\ \rho^{\lceil \frac{(j-1)n+i)}{n} \rceil}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| & i \in [n]\backslash[m]. \end{cases} \tag{31}$$

Step 1

Applying Lemma E.2 on updating $\boldsymbol{z}_{m+1}$, we have

$$\begin{aligned}
\left\|\boldsymbol{z}_{m+1}^{jn+m+1} - \boldsymbol{x}^\star\right\| &\leq \frac{\tilde{L}\Gamma^{jn+m}}{2}\sum_{i=1}^{n}\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\|^2 + \\
&\qquad \Gamma^{jn+m}\sum_{i=1}^{n}\left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\|, \\
&= \frac{\tilde{L}\Gamma^{jn+m}}{2}\Big(\sum_{i=1}^{m}\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\|^2 + \sum_{i=m+1}^{n}\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\|^2\Big) + \\
&\qquad \Gamma^{jn+m}\Big(\sum_{i=1}^{m}\left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\|\Big) + \\
&\qquad \Gamma^{jn+m}\Big(\sum_{i=m+1}^{n}\left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\|\Big).
\end{aligned}$$

From the induction hypothesis, we have

$$\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| \leq \begin{cases} \rho^{\lceil \frac{jn+i}{n} \rceil}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| & i \in [m], \\ \rho^{\lceil \frac{(j-1)n+i)}{n} \rceil}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| & i \in [n]\backslash[m], \end{cases} \tag{32}$$

Since $\boldsymbol{B}_i^{jn+m} \succeq \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})$ (Lemma E.1), applying Lemma B.1 and the induction hypothesis for σ, we have the following bound on $\left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|$:

$$\left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \overset{Lem.B.1}{\leq} L\sigma^{siqn}(\boldsymbol{B}_i^{jn+m}, \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})) \leq \begin{cases} L\delta^{siqn}(1-c)^{\lceil \frac{jn+i}{n} \rceil} & i \in [m], \\ L\delta^{siqn}(1-c)^{\lceil \frac{(j-1)n+i}{n} \rceil} & i \in [n]\backslash[m]. \end{cases} \tag{33}$$

Therefore, the bound on $\left\|\boldsymbol{z}_{m+1}^{jn+m+1}-\boldsymbol{x}^\star\right\|$ simplifies to

$$\left\|\boldsymbol{z}_{m+1}^{jn+m+1}-\boldsymbol{x}^\star\right\| \overset{(32)(33)}{\leq} \Gamma^{jn+m}\big(m\frac{\tilde{L}\epsilon^{siqn}}{2}\rho^{2j+2}+(n-m)\frac{\tilde{L}\epsilon^{siqn}}{2}\rho^{2j}+mL\delta^{siqn}(1-c)^{j+1}\rho^{j+1} \tag{34}$$

$$+(n-m)L\delta^{siqn}(1-c)^j\rho^j\big)\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|, \tag{35}$$

$$\leq \Gamma^{jn+m}\rho^j(n\frac{\tilde{L}\epsilon^{siqn}}{2}\underbrace{\rho^j}_{\leq 1}+mL\delta^{siqn}\underbrace{(1-c)^{j+1}\rho}_{\leq 1}+(n-m)L\delta^{siqn}\underbrace{(1-c)^j}_{\leq 1})\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|,$$

$$\leq \Gamma^{jn+m}\rho^j\big(n\frac{\tilde{L}\epsilon^{siqn}}{2}+nL\delta^{siqn}\big)\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|. \tag{36}$$

We now bound Γ^{jn+m}. Define $\boldsymbol{X}^{jn+m}\coloneqq\frac{1}{n}\sum_{i=1}^n\boldsymbol{B}_i^{jn+m}$ and $\boldsymbol{Y}^{jn+m}\coloneqq\frac{1}{n}\sum_{i=1}^n\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})$. We follow the same recipe to bound Γ^0 in the base case. Observe that

$$\frac{1}{n}\sum_{i=1}^n\left\|\boldsymbol{B}_i^{jn+m}-\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|\overset{\Delta}{\geq}\left\|\boldsymbol{X}^{jn+m}-\boldsymbol{Y}^{jn+m}\right\|\overset{(a)}{\geq}\mu\left\|(\boldsymbol{Y}^{jn+m})^{-1}\boldsymbol{X}^{jn+m}-\boldsymbol{I}\right\|.$$

The inequality (a) follows from Assumption **A1** which implies that $\mu\boldsymbol{I}\preceq\frac{1}{n}\sum_{i=1}^n\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})=\boldsymbol{Y}^{jn+m}$. By tracking steps (35)-(36), we can establish that

$$\frac{1}{n}\sum_{i=1}^n\left\|\boldsymbol{B}_i^{jn+m}-\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|\leq L\delta^{siqn}.$$

From the above two chain of inequalities, we deduce

$$\|(\boldsymbol{Y}^{jn+m})^{-1}\boldsymbol{X}^{jn+m}-\boldsymbol{I}\|\leq\tfrac{L\delta^{siqn}}{\mu}\overset{(30)}{<}\frac{\rho}{1+\rho}.$$

We can now upper bound Γ^{jn+m} using Banach's Lemma by exactly following the procedure laid out in the base case. We get that $\Gamma^{jn+m}=\|\sum_{i=1}^n\boldsymbol{B}_i^{jn+m})^{-1}\|\leq\frac{1+\rho}{n\mu}$. Therefore, we obtain

$$\left\|\boldsymbol{z}_{m+1}^{jn+m+1}-\boldsymbol{x}^\star\right\|\leq\frac{\frac{\tilde{L}\epsilon^{siqn}}{2}+L\delta^{siqn}}{\mu}(1+\rho)\rho^j\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|\overset{(30)}{\leq}\rho^{j+1}\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|. \tag{37}$$

Since $\boldsymbol{z}_{m+1}^{jn+m+1}=\boldsymbol{x}^{jn+m+1}$, (28) holds for $t=jn+m+1$.

Step 2

Next, we prove that $\sigma(\boldsymbol{B}_{m+1}^{jn+m+1},\nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m+1}))\leq(1-c)^{j+1}\delta^{siqn}$. We define the sequence $\{\boldsymbol{y}_k\}$, for $k=0,\ldots,j+1$, such that $\{\boldsymbol{y}_k\}=\{\boldsymbol{x}^0,\boldsymbol{z}_{m+1}^{m+1},\ldots,\boldsymbol{z}_{m+1}^{jn+m+1}\}$. The sequence $\{\boldsymbol{y}_k\}_{k=1}^{j+1}$ comprises of the updated value of $\boldsymbol{z}_{m+1}$ till the current cycle j. Since $\{\boldsymbol{y}_k\}$ comes about from the application of BFGS updates as described in the statement of Lemma B.3, therefore $\{\boldsymbol{y}_k\}$ satisfies the conditions of Lemma B.3. This implies that

$$\|\boldsymbol{y}_k-\boldsymbol{x}^\star\|\leq\rho^{\lceil\frac{(k-1)n+m+1}{n}\rceil}\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|=\rho^k\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|,$$

for $k\in[j+1]$. Since $\boldsymbol{y}_{j+1}=\boldsymbol{z}_{m+1}^{jn+m+1}$, we have proved (29) for $t=jn+m+1$. The proof is hence complete via induction. □

Corollary E.1 *If Assumptions **A1** and **A2** hold, the following holds true for all $t\geq 1$:*

$$\left\|\boldsymbol{z}_{i_t}^t-\boldsymbol{z}_{i_t}^{t-1}\right\|_{\boldsymbol{z}_{i_t}^{t-1}}\leq 2U_t, \tag{38}$$

where $U_t\coloneqq\sqrt{L}\rho^{\lceil\frac{t}{n}-1\rceil}\epsilon^{siqn}$.

Proof: We can bound $\left\|\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}\right\|_{\boldsymbol{z}_{i_t}^{t-1}}$ in the following manner:

$$\begin{aligned}\left\|\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}\right\|_{\boldsymbol{z}_{i_t}^{t-1}} &\overset{(a)}{\leq} \sqrt{L}\left\|\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}\right\| \overset{\Delta}{\leq} \sqrt{L}(\left\|\boldsymbol{z}_{i_t}^t - \boldsymbol{x}^\star\right\| + \left\|\boldsymbol{z}_{i_t}^{t-1} - \boldsymbol{x}^\star\right\|),\\ &\leq \sqrt{L}\big(\rho^{\lceil \frac{t}{n}\rceil} + \rho^{\lceil \frac{t-n}{n}\rceil}\big)\epsilon^{siqn} \leq 2\sqrt{L}\rho^{\lceil \frac{t}{n}-1\rceil}\epsilon^{siqn},\end{aligned}$$

where (a) follows since $\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^{t-1}) \preceq L\boldsymbol{I}$ ($\because$ Assumption **A1**). Therefore, the correction term $\beta_t = \frac{M}{2}\left\|\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}\right\|_{\boldsymbol{z}_{i_t}^{t-1}} \leq M\sqrt{L}\rho^{\lceil \frac{t}{n}-1\rceil}\epsilon^{siqn} = 2U_t$, which establishes (38). □

Remark 5 (β_t can be bounded by a quantity that remains constant in a cycle) *Recall that the correction factor $\beta_t = \frac{M}{2}\left\|\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}\right\|_{\boldsymbol{z}_{i_t}^{t-1}}$ in SIQN was introduced to ensure that $(1+\beta_t)^2\boldsymbol{B}_{i_t}{}^{t-1} \succeq (1+\beta_t)\boldsymbol{K}^t$ (we formalized this in Lemma E.1). Intuitively, executing SIQN with a higher correction factor $\beta_t^{new} = MU_t \geq \frac{M}{2}\left\|\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}\right\|_{\boldsymbol{z}_{i_t}^{t-1}}$ (follows from Corollary E.1) which remains constant in a cycle, it should continue to hold that $(1+\beta_t^{new})^2\boldsymbol{B}_{i_t}^{t-1} \succeq (1+\beta_t^{new})\boldsymbol{K}^t$. We skip the proof for the sake of brevity as it is similar to the analysis SIQN.*

Next, we present our mean superlinear convergence result for the iterates of SIQN. The main idea behind the proof is to substitute the linear convergence results, specifically (28) and (29) from Lemma E.3, back into the result from Lemma E.2. By doing so, the first term on the right-hand side of (27) converges quadratically, while the second term converges superlinearly. This combination leads to the desired result.

Lemma E.4 *If Assumptions **A1** and **A2** hold, for any ρ such that $0 < \rho < 1 - \frac{\mu}{dL}$, there exist positive constants ϵ^{siqn} and σ_0^{siqn} such that if $\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| \leq \epsilon^{siqn}$ and $\sigma(\boldsymbol{B}_i^0, \nabla^2 f_i(\boldsymbol{x}^0)) \leq \sigma_0^{siqn}$ for all $i \in [n]$, the sequence of iterates produced by SIQN satisfy*

$$\left\|\boldsymbol{x}^t - \boldsymbol{x}^\star\right\| \leq \big(1 - \tfrac{\mu}{dL}\big)^{\lceil \frac{t}{n}\rceil}\frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{x}^{t-i} - \boldsymbol{x}^\star\right\|.$$

Proof: Let $t = jn + m + 1$, where $0 \leq j$ and $0 \leq m < n$. From the proof of Lemma E.3, we have an uniform upper bound on $\Gamma^{jn+m} = \left\|\big(\sum_{i=1}^n \boldsymbol{B}_i^{jn+m}\big)^{-1}\right\|$, given by

$$\Gamma^{jn+m} \leq \frac{1+\rho}{n\mu},$$

and the following upper bound on $\left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|$:

$$\left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \overset{Lem.B.1}{\leq} L\sigma(\boldsymbol{B}_i^{jn+m}, \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})) \leq \begin{cases} L\delta^{siqn}(1-c)^{\lceil \frac{jn+i}{n}\rceil} & i \in [m],\\ L\delta^{siqn}(1-c)^{\lceil \frac{(j-1)n+i}{n}\rceil} & i \in [n]\backslash[m].\end{cases}$$

This gives $\left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \leq L\delta^{siqn}(1-c)^j$, for all $i \in [n]$.

Further, from LemmaE.3, we also have $\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| \leq \rho^{j+1}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|$, for all $i \in [n]$ and $\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| \leq \rho^j\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|$, for all $i \in [n]\backslash[m]$. This clearly implies $\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| \leq \rho^j\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|$, for $i = 1, \ldots, n$.

Applying Lemma E.2 at $t = jn + m + 1$ and upper bounding Γ^{jn+m}, we obtain

$$\begin{aligned}\left\|\boldsymbol{x}^{jn+m+1} - \boldsymbol{x}^\star\right\| &\leq \frac{1+\rho}{n\mu}\sum_{i=1}^n \big(\frac{\tilde{L}}{2}\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| + \left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|\big)\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\|,\\ &\overset{(a)}{\leq} \frac{1+\rho}{n\mu}\big(\frac{\tilde{L}\epsilon}{2} + L\delta^{siqn}\big)(1-c)^j\sum_{i=1}^n \left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\|,\\ &\leq \rho(1-c)^j\frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| \overset{(b)}{\leq} (1-c)^{\lceil \frac{t}{n}\rceil}\frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{x}^{t-i} - \boldsymbol{x}^\star\right\|,\end{aligned}$$

where (a) follows from the bounds $\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| \leq \rho^j \left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|$ and $\left\|\boldsymbol{B}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \leq L\delta^{siqn}(1-c)^j$ discussed above and $\rho < 1-c$. Also (b) follows since $\boldsymbol{z}_i^{jn+m} = \boldsymbol{x}^{jn+i}$, for all $i \in [m]$ and $\boldsymbol{z}_i^{jn+m} = \boldsymbol{x}^{jn-n+i}$, for all $i \in [n]\backslash[m]$, which implies $\sum_{i=1}^n \left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| = \sum_{i=1}^n \left\|\boldsymbol{x}^{t-i} - \boldsymbol{x}^\star\right\|$. This completes the proof. □

The mean superlinear convergence result of Lemma E.4 ultimately gives a superlinear rate for SIQN. Note that an identical result as Lemma E.4 is given by Theorem 1(SLIQN). Therefore, we directly provide the proof of Lemma E.4 in Appendix G.4 while proving Theorem 1 for SLIQN.

Lemma E.5 *If Assumptions **A1** and **A2** hold, for any ρ such that $0 < \rho < 1 - \frac{\mu}{dL}$, there exist positive constants ϵ^{siqn} and σ_0^{siqn} such that if $\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| \leq \epsilon^{siqn}$ and $\sigma(\boldsymbol{B}_i^0, \nabla^2 f_i(\boldsymbol{x}^0)) \leq \sigma_0^{siqn}$ for all $i \in [n]$, for the sequence of iterates $\{\boldsymbol{x}^t\}$ generated by SIQN, there exists a sequence $\{\zeta^k\}$ such that $\|\boldsymbol{x}^t - \boldsymbol{x}^\star\| \leq \zeta^{\lfloor \frac{t-1}{n} \rfloor}$ for all $t \geq 1$ and the sequence $\{\zeta^k\}$ satisfies*

$$\zeta^k \leq \epsilon\big(1 - \tfrac{\mu}{dL}\big)^{\frac{(k+2)(k+1)}{2}}, \tag{39}$$

forall $k \geq 0$.

Proof: Refer Appendix G.4. □

F EFFICIENT IMPLEMENTATION OF SLIQN

We begin by showing that the update of $\boldsymbol{x}^t$ in the SLIQN algorithm (Algorithm 2) can be carried out in $\mathcal{O}(d^2)$ cost.

F.1 Carrying out (10) in Algorithm 2 in $\mathcal{O}(d^2)$ cost

We begin by defining the following variables that track the summands in (10)

$$\bar{\boldsymbol{D}}^t := \sum_{i=1}^n \boldsymbol{D}_i^t, \boldsymbol{\phi}^t := \sum_{i=1}^n \boldsymbol{D}_i^t \boldsymbol{z}_i^t, \boldsymbol{g}^t := \sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^t). \tag{40}$$

The update (10) (for time $t+1$) can be expressed in terms of the defined variables as

$$\boldsymbol{x}^{t+1} = \big(\bar{\boldsymbol{D}}^t\big)^{-1}\big(\boldsymbol{\phi}^t - \boldsymbol{g}^t\big). \tag{41}$$

From the updates performed by Algorithm 2 at time t, we have

$$\boldsymbol{g}^t = \boldsymbol{g}^{t-1} + \big(\nabla f_{i_t}(\boldsymbol{z}_{i_t}^t) - \nabla f_{i_t}(\boldsymbol{z}_{i_t}^{t-1})\big). \tag{42}$$

Further, we can express $\boldsymbol{\phi}^t$ in terms of $\boldsymbol{\phi}^{t-1}$ as follows:

$$\boldsymbol{\phi}^t = \sum_{i=1, i\neq i_t}^n \omega_t \boldsymbol{D}_i^{t-1}\boldsymbol{z}_i^{t-1} + \boldsymbol{D}_{i_t}^t \boldsymbol{z}_{i_t}^t = \omega_t \boldsymbol{\phi}^{t-1} - \omega_t \boldsymbol{D}_{i_t}^{t-1}\boldsymbol{z}_{i_t}^{t-1} + \boldsymbol{D}_{i_t}^t \boldsymbol{z}_{i_t}^t. \tag{43}$$

This updating scheme can be implemented iteratively, where we only evaluate $\boldsymbol{\phi}^0, \boldsymbol{g}^0$ explicitly and evaluate $\boldsymbol{\phi}^t$ in $\mathcal{O}(d^2)$ cost, for all $t \geq 1$ by (43) and $\boldsymbol{g}^t$ in $\mathcal{O}(d)$ cost, for all $t \geq 1$ by (42).

Next, we demonstrate the method for updating $(\bar{\boldsymbol{D}}^t)^{-1}$. We begin by expressing $\big(\bar{\boldsymbol{D}}^t\big)^{-1}$ in terms of $\big(\bar{\boldsymbol{D}}^{t-1}\big)^{-1}$ in the following manner:

$$\begin{aligned}(\bar{\boldsymbol{D}}^t)^{-1} &= \Bigg(\sum_{i=1, i\neq i_t}^{n-1} \omega_t \boldsymbol{D}_i^{t-1} + \boldsymbol{D}_{i_t}^t\Bigg)^{-1} = \big(\omega_t\big(\bar{\boldsymbol{D}}^{t-1} - \boldsymbol{D}_{i_t}^{t-1}\big) + \boldsymbol{D}_{i_t}^t\big)^{-1}, \\ &= \omega_t^{-1}\big(\bar{\boldsymbol{D}}^{t-1} + \omega_t^{-1}\boldsymbol{D}_{i_t}^t - \boldsymbol{D}_{i_t}^{t-1}\big)^{-1}.\end{aligned} \tag{44}$$

Expanding the BFGS update (3), we can express $\omega_t^{-1}\boldsymbol{D}_{i_t}^t$ as

$$\omega_t^{-1}\boldsymbol{D}_{i_t}^t \overset{(3)}{=} \boldsymbol{Q}^t - \frac{\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t(\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t)^\mathsf{T}}{\langle \bar{\boldsymbol{u}}^t, \boldsymbol{Q}^t\bar{\boldsymbol{u}}^t\rangle} + \frac{\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t(\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t)^\mathsf{T}}{\langle \bar{\boldsymbol{u}}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t\rangle}, \tag{45}$$

where we have used the shorthand $\bar{\boldsymbol{u}}^t$ for $\bar{\boldsymbol{u}}(\boldsymbol{Q}_t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))$. Further, $\boldsymbol{Q}^t$ can be expressed as

$$\boldsymbol{Q}^t \overset{(3)}{=} \boldsymbol{D}_{i_t}^{t-1} + \frac{\boldsymbol{y}_{i_t}^t \boldsymbol{y}_{i_t}^{t\,\mathsf{T}}}{\langle \boldsymbol{y}_{i_t}^t, \boldsymbol{s}_{i_t}^t\rangle} - \frac{\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t(\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t)^\mathsf{T}}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\rangle}, \tag{46}$$

where $\boldsymbol{s}_{i_t}^t = \boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}, \boldsymbol{y}_{i_t}^t = (1+\alpha_{\lceil t/n-1\rceil})\boldsymbol{K}^t\boldsymbol{s}_{i_t}^t = (1+\alpha_{\lceil t/n-1\rceil})(\nabla f_{i_t}(\boldsymbol{z}_{i_t}^t) - \nabla f_{i_t}(\boldsymbol{z}_{i_t}^{t-1}))$.
Adding (45) and (46), we obtain

$$\begin{aligned}\omega_t^{-1}\boldsymbol{D}_{i_t}^t - \boldsymbol{D}_{i_t}^{t-1} = \frac{\boldsymbol{y}_{i_t}^t \boldsymbol{y}_{i_t}^{t\,\mathsf{T}}}{\langle \boldsymbol{y}_{i_t}^t, \boldsymbol{s}_{i_t}^t\rangle} - \frac{\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t(\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t)^\mathsf{T}}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\rangle} - \frac{\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t(\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t)^\mathsf{T}}{\langle \bar{\boldsymbol{u}}^t, \boldsymbol{Q}^t\bar{\boldsymbol{u}}^t\rangle} + \\ \frac{\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t(\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t)^\mathsf{T}}{\langle \bar{\boldsymbol{u}}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t\rangle}.\end{aligned} \tag{47}$$

Continuing from (44), we obtain

$$\left(\bar{\boldsymbol{D}}^t\right)^{-1} = \omega_t^{-1}\Bigg(\bar{\boldsymbol{D}}^{t-1} + \frac{\boldsymbol{y}_{i_t}^t \boldsymbol{y}_{i_t}^{t\,\mathsf{T}}}{\langle \boldsymbol{y}_{i_t}^t, \boldsymbol{s}_{i_t}^t\rangle} - \frac{\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t(\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t)^{\mathsf{T}}}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\rangle} - \frac{\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t(\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t)^{\mathsf{T}}}{\langle \bar{\boldsymbol{u}}^t, \boldsymbol{Q}^t\bar{\boldsymbol{u}}^t\rangle} + \frac{\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t(\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t)^{\mathsf{T}}}{\langle \bar{\boldsymbol{u}}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t\rangle}\Bigg)^{-1}.$$

Next, we define the following matrix:

$$\boldsymbol{\psi}_1 := \bar{\boldsymbol{D}}^{t-1} + \frac{\boldsymbol{y}_{i_t}^t \boldsymbol{y}_{i_t}^{t\,\mathsf{T}}}{\langle \boldsymbol{y}_{i_t}^t, \boldsymbol{s}_{i_t}^t\rangle} - \frac{\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t(\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t)^{\mathsf{T}}}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\rangle} - \frac{\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t(\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t)^{\mathsf{T}}}{\langle \bar{\boldsymbol{u}}^t, \boldsymbol{Q}^t\bar{\boldsymbol{u}}^t\rangle}.$$

Expressing $\left(\bar{\boldsymbol{D}}^t\right)^{-1}$ in terms of $\boldsymbol{\psi}_1^{-1}$ we get the following:

$$\begin{aligned}\left(\bar{\boldsymbol{D}}^t\right)^{-1} &= \omega_t^{-1}\left(\boldsymbol{\psi}_1 + \frac{\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t(\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t)^{\mathsf{T}}}{\langle \bar{\boldsymbol{u}}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t\rangle}\right)^{-1}, \\ &\overset{(16)}{=} \omega_t^{-1}\left(\boldsymbol{\psi}^{-1} - \frac{\boldsymbol{\psi}_1^{-1}\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t(\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t)^{\mathsf{T}}\boldsymbol{\psi}_1^{-1}}{\langle \bar{\boldsymbol{u}}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t\rangle + \langle \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t, \boldsymbol{\psi}_1^{-1}\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t)\bar{\boldsymbol{u}}^t\rangle}\right). \end{aligned} \tag{48}$$

Define the following matrix:

$$\boldsymbol{\psi}_2 := \bar{\boldsymbol{D}}^{t-1} + \frac{\boldsymbol{y}_{i_t}^t \boldsymbol{y}_{i_t}^{t\,\mathsf{T}}}{\langle \boldsymbol{y}_{i_t}^t, \boldsymbol{s}_{i_t}^t\rangle} - \frac{\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t(\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t)^{\mathsf{T}}}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\rangle} \implies \boldsymbol{\psi}_1 = \boldsymbol{\psi}_2 - \frac{\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t(\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t)^{\mathsf{T}}}{\langle \bar{\boldsymbol{u}}^t, \boldsymbol{Q}^t\bar{\boldsymbol{u}}^t\rangle}.$$

Expressing $\boldsymbol{\psi}_1^{-1}$ in terms of $\boldsymbol{\psi}_2^{-1}$ we get

$$\boldsymbol{\psi}_1^{-1} \overset{(16)}{=} \boldsymbol{\psi}_2^{-1} + \frac{\boldsymbol{\psi}_2^{-1}\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t(\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t)^{\mathsf{T}}\boldsymbol{\psi}_2^{-1}}{\langle \bar{\boldsymbol{u}}^t, \boldsymbol{Q}^t\bar{\boldsymbol{u}}^t\rangle - \langle \boldsymbol{Q}^t\bar{\boldsymbol{u}}^t, \boldsymbol{\psi}_2^{-1}\boldsymbol{Q}^t\bar{\boldsymbol{u}}^t\rangle}. \tag{49}$$

Define the following matrix:

$$\boldsymbol{\psi}_3 := \bar{\boldsymbol{D}}^{t-1} + \frac{\boldsymbol{y}_{i_t}^t \boldsymbol{y}_{i_t}^{t\,\mathsf{T}}}{\langle \boldsymbol{y}_{i_t}^t, \boldsymbol{s}_{i_t}^t\rangle} \implies \boldsymbol{\psi}_2 = \boldsymbol{\psi}_3 - \frac{\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t(\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t)^{\mathsf{T}}}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\rangle}.$$

Expressing $\boldsymbol{\psi}_2^{-1}$ in terms of $\boldsymbol{\psi}_3^{-1}$ we get the following:

$$\boldsymbol{\psi}_2^{-1} \overset{(16)}{=} \boldsymbol{\psi}_3^{-1} + \frac{\boldsymbol{\psi}_3^{-1}\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t(\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t)^{\mathsf{T}}\boldsymbol{\psi}_3^{-1}}{\langle \boldsymbol{s}_{i_t}^t, \boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\rangle - \langle \boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t, \boldsymbol{\psi}_3^{-1}\boldsymbol{D}_{i_t}^{t-1}\boldsymbol{s}_{i_t}^t\rangle}. \tag{50}$$

Finally, using (16) to evaluate $\boldsymbol{\psi}_3^{-1}$, we obtain

$$\boldsymbol{\psi}_3^{-1} = \left(\bar{\boldsymbol{D}}^{t-1}\right)^{-1} - \frac{\left(\bar{\boldsymbol{D}}^{t-1}\right)^{-1}\boldsymbol{y}_{i_t}^t \boldsymbol{y}_{i_t}^{t\,\mathsf{T}}\left(\bar{\boldsymbol{D}}^{t-1}\right)^{-1}}{\langle \boldsymbol{y}_{i_t}^t, \boldsymbol{s}_{i_t}^t\rangle + \langle \boldsymbol{y}_{i_t}^t, \left(\bar{\boldsymbol{D}}^{t-1}\right)^{-1}\boldsymbol{y}_{i_t}^t\rangle}. \tag{51}$$

Therefore, given access to $\left(\bar{\boldsymbol{D}}^{t-1}\right)^{-1}$, we can evaluate ψ_3^{-1} in $\mathcal{O}(d^2)$ time. Similarly, given access to $\boldsymbol{\psi}_3^{-1}$, we can evaluate $\boldsymbol{\psi}_2^{-1}$ in $\mathcal{O}(d^2)$ time. Continuing similarly, we can evaluate $\boldsymbol{\psi}_1^{-1}$ and $\left(\bar{\boldsymbol{D}}^t\right)^{-1}$ in $\mathcal{O}(d^2)$ time. This scheme can be enumerated iteratively where we only compute $\left(\bar{\boldsymbol{D}}^0\right)^{-1}$ explicitly and evaluate $\left(\bar{\boldsymbol{D}}^t\right)^{-1}, \forall t \geq 1$ by the steps (51), (50), (49), and (48). Therefore, the update (10) in Algorithm 2 can be performed in $\mathcal{O}(d^2)$ time.

F.2 Efficient Implementation of Algorithm 2 in $\mathcal{O}(d^2)$ cost

By lazily carrying out the scaling of the $\boldsymbol{D}$'s, i.e., only scaling when they are used, we can improve the per-iteration complexity of SLIQN to $\mathcal{O}(d^2)$. The resulting algorithm is specified by the following pseudo code:

Algorithm 3 Sharpened Lazy Incremental Quasi-Newton (SLIQN)

Function {Sherman-Morrison} $\{\boldsymbol{A}^{-1}, \boldsymbol{u}, \boldsymbol{v}\}$
  return $\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}}{1+\langle \boldsymbol{v}, \boldsymbol{A}^{-1}\boldsymbol{u}\rangle}$
EndFunction

Initialize: Initialize $\{\boldsymbol{z}_i, \boldsymbol{D}_i\}_{i=1}^n$ similar to Algorithm F.2;
Evaluate $\bar{\boldsymbol{D}} := \big(\sum_{i=1}^n \boldsymbol{D}_i\big)^{-1}, \boldsymbol{\phi} := \sum_{i=1}^n \boldsymbol{D}_i \boldsymbol{z}_i$, and $\boldsymbol{g} := \sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i)$;
Maintain running auxiliary variables $\boldsymbol{x}, \bar{\boldsymbol{u}}, \boldsymbol{y}, \boldsymbol{s}, \boldsymbol{Q}, \boldsymbol{D}^{\text{old}}, \boldsymbol{K}$; // $\boldsymbol{x}$ keeps a track of $\boldsymbol{x}^t$, $\bar{\boldsymbol{u}}$ keeps a track of the greedy vector, $\boldsymbol{y}$ keeps a track of $\nabla f_{i_t}(\boldsymbol{z}_{i_t}^t) - \nabla f_{i_t}(\boldsymbol{z}_{i_t}^{t-1})$, $\boldsymbol{s}$ keeps a track of $\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}$, whereas $\boldsymbol{Q}, \boldsymbol{D}^{\text{OLD}}, \boldsymbol{K}$ keep track of the intermediate matrices
while *not converged:*
  Current index to be updated is $i_t \leftarrow (t-1) \mod n+1$;
  Update $\boldsymbol{x}$ as $\boldsymbol{x} \leftarrow \big(\bar{\boldsymbol{D}}\big)\big(\boldsymbol{\phi} - \boldsymbol{g}\big)$;
  Update $\boldsymbol{s} \leftarrow \boldsymbol{x} - \boldsymbol{z}_{i_t}$;
  Update $\boldsymbol{y} \leftarrow \nabla f_{i_t}(\boldsymbol{x}^t) - \nabla f_{i_t}(\boldsymbol{z}_{i_t})$;
  Compute ω_t; // $\omega_t = (1+\alpha_{\lceil t/n \rceil})^2$ if $t \mod n = 0$ and 1 otherwise
  Update $\boldsymbol{Q}$ as $\boldsymbol{Q} \leftarrow \text{BFGS}((1+\alpha_{\lceil t/n-1 \rceil})^2 \boldsymbol{D}_{i_t}, (1+\alpha_{\lceil t/n-1 \rceil})\boldsymbol{K}, \boldsymbol{s})$, where // Lazy Step

$$\boldsymbol{K} \leftarrow \int_0^1 \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t} + \tau(\boldsymbol{x}^t - \boldsymbol{z}_{i_t}))d\tau.$$

  Update $\bar{\boldsymbol{u}}$ as $\bar{\boldsymbol{u}} \leftarrow \arg\max_{\boldsymbol{u} \in \{\boldsymbol{e}_i\}_{i=1}^d} \frac{\langle \boldsymbol{u}, \boldsymbol{Q}\boldsymbol{u}\rangle}{\langle \boldsymbol{u}, \nabla^2 f_{i_t}(\boldsymbol{x}^t)\boldsymbol{u}\rangle}$;
  Update $\boldsymbol{D}^{\text{old}}$ as $\boldsymbol{D}^{\text{old}} \leftarrow \boldsymbol{D}_{i_t}$;
  Update $\boldsymbol{D}_{i_t}$ as $\boldsymbol{D}_{i_t} \leftarrow \text{BFGS}(\boldsymbol{Q}, \nabla^2 f_{i_t}(\boldsymbol{x}^t), \bar{\boldsymbol{u}})$;
  Update $\boldsymbol{\phi}$ as $\boldsymbol{\phi} \leftarrow \omega_t(\boldsymbol{\phi} - \boldsymbol{D}^{\text{old}}\boldsymbol{z}_{i_t}) + \boldsymbol{D}_{i_t}\boldsymbol{x}^t$;
  Update $\boldsymbol{g}$ as $\boldsymbol{g} \leftarrow \boldsymbol{g} + \big(\nabla f_{i_t}(\boldsymbol{x}^t) - \nabla f_{i_t}(\boldsymbol{z}_{i_t})\big)$;
  Update $\bar{\boldsymbol{D}}$ as $\bar{\boldsymbol{D}} \leftarrow \text{Sherman-Morrison}(\bar{\boldsymbol{D}}, \boldsymbol{y}, \frac{1}{\langle \boldsymbol{y}, \boldsymbol{s}\rangle}\boldsymbol{y})$;
  Update $\bar{\boldsymbol{D}}$ as $\bar{\boldsymbol{D}} \leftarrow \text{Sherman-Morrison}(\bar{\boldsymbol{D}}, -\boldsymbol{D}^{\text{old}}\boldsymbol{s}, \frac{1}{\langle \boldsymbol{s}, \boldsymbol{D}^{\text{old}}\boldsymbol{s}\rangle}\boldsymbol{D}^{\text{old}}\boldsymbol{s})$;
  Update $\bar{\boldsymbol{D}}$ as $\bar{\boldsymbol{D}} \leftarrow \text{Sherman-Morrison}(\bar{\boldsymbol{D}}, -\boldsymbol{Q}\bar{\boldsymbol{u}}, \frac{1}{\langle \bar{\boldsymbol{u}}, \boldsymbol{Q}\bar{\boldsymbol{u}}\rangle}\boldsymbol{Q}\bar{\boldsymbol{u}})$;
  Update $\bar{\boldsymbol{D}}$ as $\bar{\boldsymbol{D}} \leftarrow \omega_t^{-1}\text{Sherman-Morrison}(\bar{\boldsymbol{D}}, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t})\bar{\boldsymbol{u}}, \frac{1}{\langle \bar{\boldsymbol{u}}, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t})\bar{\boldsymbol{u}}\rangle}\nabla^2 f_{i_t}(\boldsymbol{z}_{i_t})\bar{\boldsymbol{u}})$;
  Update $\boldsymbol{z}_{i_t}$ as $\boldsymbol{z}_{i_t} \leftarrow \boldsymbol{x}$;
Increment the iteration counter t;
end while

G CONVERGENCE ANALYSIS OF SLIQN

G.1 Proof of Lemma 1

For all $t \geq 0$, we define $\boldsymbol{H}^t := \big(\sum_{i=1}^n \boldsymbol{D}_i^t\big)^{-1}$. From the update for $\boldsymbol{x}^t$ (8), we have

$$
\begin{aligned}
\boldsymbol{x}^t - \boldsymbol{x}^* &= \boldsymbol{H}^{t-1}\Big(\sum_{i=1}^n \boldsymbol{D}_i^{t-1}\boldsymbol{z}_i^{t-1} - \sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^{t-1})\Big) - \boldsymbol{x}^*, \\
&\overset{(a)}{=} \boldsymbol{H}^{t-1}\Big(\sum_{i=1}^n \boldsymbol{D}_i^{t-1}(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star) - \sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^{t-1})\Big), \\
&\overset{(b)}{=} \boldsymbol{H}^{t-1}\Big(\sum_{i=1}^n \boldsymbol{D}_i^{t-1}(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star) - \sum_{i=1}^n (\nabla f_i(\boldsymbol{z}_i^{t-1}) - \nabla f_i(\boldsymbol{x}^\star))\Big), \\
&\overset{(c)}{=} \boldsymbol{H}^{t-1}\Big(\sum_{i=1}^n \boldsymbol{D}_i^{t-1}(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star) - \sum_{i=1}^n \int_0^1 \nabla^2 f(\boldsymbol{x}^* + (\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)v)(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)dv\Big), \\
&\overset{(d)}{=} \boldsymbol{H}^{t-1}\Big(\sum_{i=1}^n \big(\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\big)(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star) + \\
&\qquad\qquad \sum_{i=1}^n \int_0^1 (\nabla^2 f_i(\boldsymbol{z}_i^{t-1}) - \nabla^2 f(\boldsymbol{x}^* + (\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)v))(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)dv\Big).
\end{aligned}
$$

The equality (a) follows from the definition $\boldsymbol{H}^{t-1} = \big(\sum_{i=1}^n \boldsymbol{D}_i^{t-1}\big)^{-1}$. The equality (b) uses the fact that $\nabla f(\boldsymbol{x}^*) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(\boldsymbol{x}^\star) = \boldsymbol{0}$. The equality (c) follows from the Fundamental Theorem of Calculus, and the equality (d) follows by adding and subtracting $\sum_{i=1}^n \nabla^2 f_i(\boldsymbol{z}_i^{t-1})(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)$. Taking norm on both sides and applying the Triangle inequality, we obtain

$$
\begin{aligned}
\big\|\boldsymbol{x}^t - \boldsymbol{x}^*\big\| &\overset{\Delta}{\leq} \big\|\boldsymbol{H}^{t-1}\big\| \Bigg(\sum_{i=1}^n \big\|\big(\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\big)\big\| \big\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\big\| + \\
&\qquad \sum_{i=1}^n \Big\|\int_0^1 (\nabla^2 f_i(\boldsymbol{z}_i^{t-1}) - \nabla^2 f(\boldsymbol{x}^* + (\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)v))(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)dv\Big\|\Bigg), \\
&\overset{(a)}{\leq} \big\|\boldsymbol{H}^{t-1}\big\| \Bigg(\sum_{i=1}^n \big\|\big(\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\big)\big\| \big\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\big\| + \\
&\qquad \sum_{i=1}^n \int_0^1 \big\|(\nabla^2 f_i(\boldsymbol{z}_i^{t-1}) - \nabla^2 f(\boldsymbol{x}^* + (\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)v))(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)\big\| dv\Bigg), \\
&\leq \big\|\boldsymbol{H}^{t-1}\big\| \Bigg(\sum_{i=1}^n \big\|\big(\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\big)\big\| \big\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\big\| + \\
&\qquad \sum_{i=1}^n \int_0^1 \big\|(\nabla^2 f_i(\boldsymbol{z}_i^{t-1}) - \nabla^2 f(\boldsymbol{x}^* + (\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)v))\big\| \big\|(\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)\big\| dv\Bigg), \\
&\overset{(b)}{\leq} \big\|\boldsymbol{H}^{t-1}\big\| \Bigg(\sum_{i=1}^n \big\|\big(\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\big)\big\| \big\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\big\| + \tilde{L}\int_0^1 (1-v)dv \sum_{i=1}^n \big\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^*\big\|^2\Bigg), \\
&\leq \big\|\boldsymbol{H}^{t-1}\big\| \Bigg(\sum_{i=1}^n \big\|\big(\boldsymbol{D}_i^{t-1} - \nabla^2 f_i(\boldsymbol{z}_i^{t-1})\big)\big\| \big\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\big\| + \frac{\tilde{L}}{2}\sum_{i=1}^n \big\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^*\big\|^2\Bigg).
\end{aligned}
$$

The inequality (a) follows from the known result that if $g\colon \mathbb{R} \to \mathbb{R}^d$ is a continuous function, then $\Big\|\int_0^1 g(v)\,dv\Big\| \leq \int_0^1 \|g(v)\|\,dv$, and the inequality (b) follows from the assumption that the Hessian of f_i is $\tilde{L}-$ Lipschitz (**A2**), i.e. $\big\|\nabla^2 f_i(\boldsymbol{z}_i^{t-1}) - \nabla^2 f(\boldsymbol{x}^* + (\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star)v)\big\| \leq \tilde{L}(1-v)\big\|\boldsymbol{z}_i^{t-1} - \boldsymbol{x}^\star\big\|$. This completes the proof.

G.2 Proof of Lemma 2

For a given ρ that satisfies $0 < \rho < 1 - \frac{\mu}{dL}$, we choose $\epsilon, \delta > 0$ such that they satisfy

$$\tfrac{1}{\mu}\left(\tfrac{\tilde{L}\epsilon}{2} + L\delta(1 + M\sqrt{L}\epsilon)^2 + L^{\frac{3}{2}}M\epsilon(2 + M\sqrt{L}\epsilon)\right) \leq \frac{\rho}{1+\rho} < 1. \tag{52}$$

Remark 6 *Indeed, there exists positive constants $\epsilon, \delta > 0$ that satisfy 52. Recall from the premise that δ is a function of ϵ, σ_0, and we can define the left-hand-side of 52 as a function $h(\epsilon, \sigma_0)$ as*

$$h(\epsilon, \sigma_0) \coloneqq \tfrac{1}{\mu}\left(\tfrac{\tilde{L}\epsilon}{2} + Le^{\frac{4M\sqrt{L}\epsilon}{1-\rho}}\big(\sigma_0 + \epsilon\tfrac{4Md\sqrt{L}}{1-\rho(1-\frac{\mu}{dL})^{-1}}\big)(1 + M\sqrt{L}\epsilon)^2 + L^{\frac{3}{2}}M\epsilon(2 + M\sqrt{L}\epsilon)\right).$$

Fix $\sigma_0 = \frac{\mu}{2L}(\frac{\rho}{1+\rho}) > 0$. It is easy to see that $h(\epsilon, \sigma_0)$ is continuous and monotonically increasing in ϵ. Also, note that $h(0, \frac{\mu}{2L}(\frac{\rho}{1+\rho})) = \frac{\rho}{2(1+\rho)}$ and $\lim_{\epsilon\to\infty} h(\epsilon, \frac{\mu}{2L}(\frac{\rho}{1+\rho})) = \infty$. We can therefore apply the Intermediate Value Theorem (IVT) to guarantee that there exists $\epsilon > 0$ such that $h(\epsilon, \frac{\mu}{2L}(\frac{\rho}{1+\rho})) \leq \frac{\rho}{1+\rho}$.

Similar to Lemma E.3, we use Induction on t to prove the result.

Base case: At $t = 1$, from Lemma 1, we have

$$\left\|\boldsymbol{x}^1 - \boldsymbol{x}^\star\right\| \leq \frac{\tilde{L}\Gamma^0}{2}\sum_{i=1}^n \left\|\boldsymbol{z}_i^0 - \boldsymbol{x}^\star\right\|^2 + \Gamma^0\sum_{i=1}^n \left\|\boldsymbol{D}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\| \left\|\boldsymbol{z}_i^0 - \boldsymbol{x}^\star\right\|.$$

From the initialization, we have that $\boldsymbol{D}_i^0 = (1+\alpha_0)^2\boldsymbol{I}_i^0$ and $\boldsymbol{z}_i^0 = \boldsymbol{x}^0$, for all $i \in [n]$, and $\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| \leq \epsilon$. Substituting these in the above expression, we obtain

$$\begin{aligned}
\left\|\boldsymbol{x}^1 - \boldsymbol{x}^\star\right\| &\leq \Gamma^0\left(n\tfrac{\tilde{L}\epsilon}{2} + \sum_{i=1}^n \left\|(1+\alpha_0)^2\boldsymbol{I}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\|\right)\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|, \qquad (53)\\
&\overset{(a)}{\leq} \Gamma^0\left(n\tfrac{\tilde{L}\epsilon}{2} + (1+\alpha_0)^2\sum_{i=1}^n \left\|\boldsymbol{I}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\| + \alpha_0(\alpha_0+2)\sum_{i=1}^n \underbrace{\left\|\nabla^2 f_i(\boldsymbol{z}_i^0)\right\|}_{\leq L}\right)\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|,\\
&\overset{(b)}{\leq} \Gamma^0\left(n\tfrac{\tilde{L}\epsilon}{2} + nL(1+\alpha_0)^2\sigma_0 + nL\alpha_0(\alpha_0+2)\right)\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|,\\
&\leq \Gamma^0\left(n\tfrac{\tilde{L}\epsilon}{2} + nL(1+M\sqrt{L}\epsilon)^2\sigma_0 + nL^{\frac{3}{2}}M\epsilon(M\sqrt{L}\epsilon+2)\right)\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|,\\
&\leq \Gamma^0\left(n\tfrac{\tilde{L}\epsilon}{2} + nL(1+M\sqrt{L}\epsilon)^2\delta + nL^{\frac{3}{2}}M\epsilon(M\sqrt{L}\epsilon+2)\right)\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|, \qquad (54)
\end{aligned}$$

where (a) follows by adding and subtracting $(1+\alpha_0)^2\nabla^2 f_i(\boldsymbol{z}_i^0)$ to $(1+\alpha_0)^2\boldsymbol{I}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)$ and applying the Triangle inequality. To see why inequality (b) is true, first recall from the initialization that $\sigma(\boldsymbol{I}_i^0, \nabla^2 f_i(\boldsymbol{z}_i^0)) \leq \sigma_0$ and $\boldsymbol{I}_i^0 \succeq \nabla^2 f_i(\boldsymbol{z}_i^0)$. Applying Lemma B.1, we have $\left\|\boldsymbol{I}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\| \leq L\sigma(\boldsymbol{I}_i^0, \nabla^2 f_i(\boldsymbol{z}_i^0)) \leq L\sigma_0$.

We now upper bound Γ^0. Define $\boldsymbol{X}^0 \coloneqq \frac{1}{n}\sum_{i=1}^n \boldsymbol{D}_i^0$ and $\boldsymbol{Y}^0 \coloneqq \frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(\boldsymbol{z}_i^0)$. Then, we have

$$\frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{D}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\| \overset{\Delta}{\geq} \left\|\boldsymbol{X}^0 - \boldsymbol{Y}^0\right\| = \left\|(\boldsymbol{Y}^0)((\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I})\right\| \overset{(a)}{\geq} \mu\left\|(\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I}\right\|,$$

where the inequality (a) follows from Assumption **A1** which implies $\mu\boldsymbol{I} \preceq \frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(\boldsymbol{z}_i^0) = \boldsymbol{Y}^0$. By tracking the parts of steps (53)-(54) which bound $\left\|\boldsymbol{D}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\|$, we get

$$\sum_{i=1}^n \left\|\boldsymbol{D}_i^0 - \nabla^2 f_i(\boldsymbol{z}_i^0)\right\| \leq nL(1+M\sqrt{L}\epsilon)^2\delta + nL^{\frac{3}{2}}M\epsilon(M\sqrt{L}\epsilon+2).$$

From the above two chain of inequalities, we obtain

$$\left\|(\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I}\right\| \leq \tfrac{1}{\mu}\left(L(1+M\sqrt{L}\epsilon)^2\delta + L^{\frac{3}{2}}M\epsilon(M\sqrt{L}\epsilon+2)\right) \overset{(52)}{<} \frac{\rho}{1+\rho}.$$

We can now upper bound Γ^0 using Banach's Lemma A.1. Consider the matrix $(\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I}$ and note that it satisfies the requirement of A.1 from the above result. Therefore, we have

$$\left\|(\boldsymbol{X}^0)^{-1}\boldsymbol{Y}^0\right\| = \left\|\left(\boldsymbol{I} + ((\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0) - \boldsymbol{I})\right)^{-1}\right\| \overset{Lem.A.1}{\leq} \frac{1}{1 - \|(\boldsymbol{Y}^0)^{-1}\boldsymbol{X}^0 - \boldsymbol{I}\|} \leq 1 + \rho.$$

Recall that $\mu\boldsymbol{I} \preceq \boldsymbol{Y}^0$. Using this and the previous result, we can upper bound $\left\|(\boldsymbol{X}^0)^{-1}\right\|$ as $\mu\left\|(\boldsymbol{X}^0)^{-1}\right\| \leq \left\|(\boldsymbol{X}^0)^{-1}\boldsymbol{Y}^0\right\| \leq 1 + \rho$. This gives us, $\Gamma^0 = \frac{1}{n}\left\|(\boldsymbol{X}^0)^{-1}\right\| \leq \frac{1+\rho}{n\mu}$.

Substituting this bound on Γ^0 in 54, we obtain

$$\left\|\boldsymbol{x}^1 - \boldsymbol{x}^\star\right\| \leq \tfrac{1+\rho}{\mu}\big(\tfrac{\tilde{L}\epsilon}{2} + L(1 + M\sqrt{L}\epsilon)^2\delta + L^{\frac{3}{2}}M\epsilon(M\sqrt{L}\epsilon + 2)\big)\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| \overset{(52)}{\leq} \rho\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|.$$

To complete the base step, we now upper bound $\sigma(\omega_1^{-1}\boldsymbol{D}_1^1, \nabla^2 f_1(\boldsymbol{z}_1^1))$, where $\omega_1 = 1$. Applying Lemma B.3 with parameters as $T = 1, \tilde{\boldsymbol{x}} = \boldsymbol{x}^\star, \boldsymbol{P}^0 = (1+\alpha_0)^2\boldsymbol{I}_1^0 = \boldsymbol{D}_1^0$ (refer to Remark 3 we made for Lemma B.3, where we stated that the results of Lemma B.3 remain unchanged on redefining $r_k := 2\sqrt{L}\rho^{k-1}\epsilon \geq 2\sqrt{L}\rho^{k-1}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\|$), we get

$$\sigma(\boldsymbol{D}_1^1, \nabla^2 f_1(\boldsymbol{z}_1^1)) \leq \big(1-c\big)e^{\frac{4M\sqrt{L}\epsilon}{1-\rho}}\big(\underbrace{\sigma(\boldsymbol{I}_1^0, \nabla^2 f_1(\boldsymbol{z}_1^0))}_{\leq\sigma_0} + \epsilon\frac{4Md\sqrt{L}}{1-\frac{\rho}{1-c}}\big) \leq (1-c)\delta.$$

Finally, as a technical remark, the proof of Lemma B.3 already shows that $\boldsymbol{D}_1^1 \succeq \nabla^2 f_1(\boldsymbol{z}_1^1)$, and therefore $\sigma(\boldsymbol{D}_1^1, \nabla^2 f_1(\boldsymbol{z}_1^1))$ is well defined. This completes the proof for for the base case.

We now prove that (12) and (13) hold for any $t > 1$ by induction.

Induction hypothesis (IH): Let (12) and (13) hold for all $t \in [jn+m]$ for some $j \geq 0$ and $0 \leq m < n$.

Induction step: We then prove that (12) and (13) also hold for $t = jn + m + 1$. Recall that the tuples are updated in a deterministic cyclic order, and at the current time t, we are in the j^{th} cycle and have updated the m^{th} tuple. Therefore, it is easy to note that $\boldsymbol{z}_i^{jn+m} = \boldsymbol{x}^{jn+i}$, for all $i \in [m]$, which refer to the tuples updated in this cycle, and $\boldsymbol{z}_i^{jn+m} = \boldsymbol{x}^{jn-n+i}$ for all $i \in [n]\backslash[m]$, which refer to the tuples updated in the previous cycle. From the induction hypothesis, we have

$$\left\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\right\| \leq \begin{cases}\rho^{\lceil\frac{jn+i}{n}\rceil}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| & i \in [m],\\ \rho^{\lceil\frac{(j-1)n+i)}{n}\rceil}\left\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\right\| & i \in [n]\backslash[m].\end{cases} \tag{55}$$

We will execute the induction step in three distinct stages. In the first stage we will establish an upper bound on $\sum_{i=1}^n \|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\|$. In the second stage, we will use the previous result and Lemma 1 to bound $\|\boldsymbol{z}_{m+1}^{jn+m+1} - \boldsymbol{x}^\star\|$. In the final stage, we will prove the linear convergence of the updated Hessian approximation, i.e., $\sigma(\omega_{jn+m+1}^{-1}\boldsymbol{D}_{m+1}^{jn+m+1}, \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m+1})) \leq (1-c)^{j+1}\delta$.

Since $\boldsymbol{D}_{i_t}^t$ is updated in a different manner for $t \bmod n \neq 0$ and $t \bmod n = 0$, we split the first stage into two cases corresponding to $t \bmod n \neq 0$ and $t \bmod n = 0$.

Stage 1, Case 1: $t \bmod n \neq 0$

Since $t = jn + m + 1$, this case is equivalent to considering $0 \leq m < n-1$. From the structure of the cyclic updates and the pre-multiplication of the scaling factor, it is easy to note that $\boldsymbol{D}_i^{jn+m} = \boldsymbol{D}_i^{jn+i}$, for all $i \in [m]$, $\boldsymbol{D}_i^{jn+m} = (1 + M\sqrt{L}\epsilon\rho^j)^2\boldsymbol{D}_i^{jn-n+i}$, for all $i \in [n-1]\backslash[m]$, and $\boldsymbol{D}_i^{jn+m} = \boldsymbol{D}_i^{jn}$, for $i = n$.

We want to bound $\sum_{i=1}^n \|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\|$. For all $i \in [m]$, from the induction hypothesis, we have

$$\begin{aligned}&\sigma(\boldsymbol{D}_i^{jn+m}, \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})) \leq (1-c)^{\lceil\frac{jn+i}{n}\rceil}\delta,\\ \overset{Lem.B.1}{\Longrightarrow}&\left\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| = \left\|\boldsymbol{D}_i^{jn+i} - \nabla^2 f_i(\boldsymbol{x}^{jn+i})\right\| \leq L(1-c)^{\lceil\frac{jn+i}{n}\rceil}\delta.\end{aligned} \tag{56}$$

For all $i \in [n-1]\backslash[m]$, we follow in the footsteps of 53-54 from the base case to get

$$\left\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| = \left\|(1+M\sqrt{L}\epsilon\rho^j)^2\boldsymbol{D}_i^{jn-n+i} - \nabla^2 f_i(\boldsymbol{x}^{jn-n+i})\right\|,$$

$$\leq (1+M\sqrt{L}\epsilon\rho^j)^2\left\|\boldsymbol{D}_i^{jn-n+i} - \nabla^2 f_i(\boldsymbol{x}^{jn-n+i})\right\| + M\sqrt{L}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j)\underbrace{\left\|\nabla^2 f_i(\boldsymbol{x}^{jn-n+i})\right\|}_{\leq L},$$

$$\leq (1+M\sqrt{L}\epsilon\rho^j)^2\left\|\boldsymbol{D}_i^{jn-n+i} - \nabla^2 f_i(\boldsymbol{x}^{jn-n+i})\right\| + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j),$$

$$\overset{(a)}{\leq} (1+M\sqrt{L}\epsilon\rho^j)^2 L\delta(1-c)^j + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j). \tag{57}$$

The inequality (a) follows from $\|\boldsymbol{D}_i^{jn-n+i} - \nabla^2 f_i(\boldsymbol{x}^{jn-n+i})\| \leq (1-c)^j\delta$, which can be established from the induction hypothesis in a similar way as we did for the case with $i \in [m]$. Next, for $i = n$, we have

$$\begin{aligned}\left\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| &= \left\|\boldsymbol{D}_i^{jn} - \nabla^2 f_i(\boldsymbol{z}_i^{jn})\right\|,\\ &= \left\|\omega_{jn}(\omega_{jn}^{-1}\boldsymbol{D}_i^{jn} - \nabla^2 f_i(\boldsymbol{z}_i^{jn})) + (\omega_{jn}-1)\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|,\\ &\overset{\Delta}{\leq} \omega_{jn}\|\omega_{jn}^{-1}\boldsymbol{D}_i^{jn} - \nabla^2 f_i(\boldsymbol{z}_i^{jn})\| + |\omega_{jn}-1|\|\nabla^2 f_i(\boldsymbol{z}_i^{jn})\|,\\ &\overset{(a)}{\leq} (1+M\sqrt{L}\epsilon\rho^j)L\delta(1-c)^j + |\omega_{jn}-1|\|\nabla^2 f_i(\boldsymbol{z}_i^{jn})\|,\\ &\overset{(b)}{\leq} (1+M\sqrt{L}\epsilon\rho^j)L\delta(1-c)^j + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j).\end{aligned} \tag{58}$$

To see the deduction (a), we follow in the footsteps of the case with $i \in [m]$. Concretely, from induction and Lemma B.1, $\|\omega_{jn}^{-1}\boldsymbol{D}_i^{jn} - \nabla^2 f_i(\boldsymbol{z}_i^{jn})\| \leq L\sigma(\omega_{jn}^{-1}\boldsymbol{D}_i^{jn}, \nabla^2 f_i(\boldsymbol{z}_i^{jn})) \leq L\delta(1-c)^j$. Inequality (b) uses the fact that $\omega_{jn} = 1 + M\sqrt{L}\epsilon\rho^j$ and $\|\nabla^2 f_i(\boldsymbol{z}_i^{jn})\| \leq L$ ($\because$ Assumption **A1**).

We can now bound the quantity $\sum_{i=1}^n \|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\|$ using(56), (57) and (58), as follows:

$$\sum_{i=1}^n \left\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \leq mL\delta(1-c)^{j+1} + (n-m)\big((1+M\sqrt{L}\epsilon\rho^j)^2 L\delta(1-c)^j + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j)\big),$$

$$\begin{aligned}&\leq mL\delta + (n-m)\big((1+M\sqrt{L}\epsilon)^2 L\delta + ML^{\frac{3}{2}}\epsilon(2+M\sqrt{L}\epsilon)\big),\\ &\overset{(a)}{\leq} nL\delta(1+M\sqrt{L}\epsilon)^2 + nML^{\frac{3}{2}}\epsilon(2+M\sqrt{L}\epsilon),\end{aligned} \tag{59}$$

where (a) follows since $mL\delta + (n-m)(1+M\sqrt{L}\epsilon)^2 L\delta < mL(1+M\sqrt{L}\epsilon)^2 L\delta + (n-m)(1+M\sqrt{L}\epsilon)^2 L\delta = nL(1+M\sqrt{L}\epsilon)^2 L\delta$.

Stage 1, Case 2: $t \bmod n = 0$

Since $t = jn+m+1$, this case is equivalent to considering $m = n-1$. This is a simpler case, as compared to the previous case. Here, we have $\boldsymbol{D}_i^{jn} = (1+M\sqrt{L}\epsilon\rho^j)^2\boldsymbol{D}_i^{jn-n+i}$, for all $i \in [n-1]$, and $\boldsymbol{D}_n^{jn}$ would be used as it is.

We follow exactly the steps leading up to (57) and (58). For, $i \in [n-1]$, using the reasoning in the derivation of (57), we get

$$\left\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \leq (1+M\sqrt{L}\epsilon\rho^j)^2 L\delta(1-c)^j + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j),$$

For $i = n$, from equation (58), we get

$$\left\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \leq (1+M\sqrt{L}\epsilon\rho^j)L\delta(1-c)^j + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j).$$

We can now bound the quantity $\sum_{i=1}^{n}\left\|\boldsymbol{D}_i^{jn+m}-\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|$ in the following manner:

$$\sum_{i=1}^{n}\left\|\boldsymbol{D}_i^{jn+m}-\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \leq n(1+M\sqrt{L}\epsilon\rho^j)L\delta(1-c)^j + nML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j),$$
$$\overset{(a)}{\leq} nL\delta(1+M\sqrt{L}\epsilon)^2 + nML^{\frac{3}{2}}\epsilon(2+M\sqrt{L}\epsilon), \tag{60}$$

where (a) follows by bounding the terms < 1.

Stage 2

We now use the result from Stage 1 to bound $\|\boldsymbol{z}_{m+1}^{jn+m+1}-\boldsymbol{x}^\star\|$,

$$\|\boldsymbol{z}_{m+1}^{jn+m+1}-\boldsymbol{x}^\star\| \overset{Lem.1}{\leq} \frac{\tilde{L}\Gamma^{jn+m}}{2}\sum_{i=1}^{n}\|\boldsymbol{z}_i^{jn+m}-\boldsymbol{x}^\star\|^2 + \Gamma^{jn+m}\sum_{i=1}^{n}\left\|\boldsymbol{D}_i^{jn+m}-\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|\|\boldsymbol{z}_i^{jn+m}-\boldsymbol{x}^\star\|,$$

$$\overset{(a)}{\leq} \Gamma^{jn+m}\big(n\frac{\tilde{L}\epsilon}{2}\rho^j + \sum_{i=1}^{n}\left\|\boldsymbol{D}_i^{jn+m}-\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\|\rho^j\big)\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|,$$
$$\overset{(59),(60)}{\leq} \Gamma^{jn+m}\big(n\frac{\tilde{L}\epsilon}{2} + nL\delta(1+M\sqrt{L}\epsilon)^2 + nML^{\frac{3}{2}}\epsilon(2+M\sqrt{L}\epsilon)\big)\rho^j\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|.$$

The inequality (a) follows from the induction hypothesis that $\|\boldsymbol{z}_i^{jn+m}-\boldsymbol{x}^\star\| \leq \rho^j\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|$, for all $i \in [n]\backslash[m]$ and $\|\boldsymbol{z}_i^{jn+m}-\boldsymbol{x}^\star\| \leq \rho^{j+1}\|\boldsymbol{x}^0-\boldsymbol{x}^\star\|$, for all $i \in [m]$. Since $\rho < 1$, we can have a common upper bound, $\|\boldsymbol{z}_i^{jn+m}-\boldsymbol{x}^\star\| \leq \rho^j\|\boldsymbol{x}^0-\boldsymbol{x}^\star\|$, for all $i \in [n]$.

We now bound Γ^{jn+m}. Define $\boldsymbol{X}^{jn+m} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{D}_i^{jn+m}$ and $\boldsymbol{Y}^{jn+m} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})$. We follow the same recipe when we bound Γ^0 in the base case. Observe that

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{D}_i^{jn+m}-\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \overset{\Delta}{\geq} \left\|\boldsymbol{X}^{jn+m}-\boldsymbol{Y}^{jn+m}\right\| \overset{(a)}{\geq} \mu\left\|(\boldsymbol{Y}^{jn+m})^{-1}\boldsymbol{X}^{jn+m}-\boldsymbol{I}\right\|.$$

The inequality (a) follows from Assumption **A1** that $\mu\boldsymbol{I} \preceq \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(\boldsymbol{z}_i^{jn+m}) = \boldsymbol{Y}^{jn+m}$. Also, restating the result from Stage 1, we have

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{D}_i^{jn+m}-\nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\right\| \leq L\delta(1+M\sqrt{L}\epsilon)^2 + ML^{\frac{3}{2}}\epsilon(2+M\sqrt{L}\epsilon)$$

From the above two chain of inequalities, we deduce

$$\|(\boldsymbol{Y}^{jn+m})^{-1}\boldsymbol{X}^{jn+m}-\boldsymbol{I}\| \leq \frac{1}{\mu}\bigg(L\delta(1+M\sqrt{L}\epsilon)^2 + ML^{\frac{3}{2}}\epsilon(2+M\sqrt{L}\epsilon)\bigg) \overset{(52)}{<} \frac{\rho}{1+\rho}.$$

We can now upper bound Γ^{jn+m} using Banach's Lemma by exactly following the procedure laid out in the base case. We get that $\Gamma^{jn+m} = \|(\sum_{i=1}^{n}\boldsymbol{D}_i^{jn+m})^{-1}\| \leq \frac{1+\rho}{n\mu}$. Substituting this above, we get

$$\|\boldsymbol{z}_{m+1}^{jn+m+1}-\boldsymbol{x}^\star\| \leq \frac{1+\rho}{n\mu}\big(n\frac{\tilde{L}\epsilon}{2} + nL\delta(1+M\sqrt{L}\epsilon)^2 + nML^{\frac{3}{2}}\epsilon(2+M\sqrt{L}\epsilon)\big)\rho^j\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|,$$
$$\overset{(52)}{\leq} \rho^{j+1}\left\|\boldsymbol{x}^0-\boldsymbol{x}^\star\right\|. \tag{61}$$

This completes the induction step proof for (12) at $t = jn+m+1$.

Stage 3

In this stage, we prove the linear convergence of the updated Hessian approximation, that is, we show that $\sigma(\omega_{jn+m+1}^{-1}\boldsymbol{D}_{m+1}^{jn+m+1}, \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m+1})) \leq (1-c)^{j+1}\delta$. But first, we establish that the $\sigma(\cdot)$ metric is well defined, by showing that

$$\omega_{jn+m+1}^{-1}\boldsymbol{D}_{m+1}^{jn+m+1} \succeq \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m+1}). \tag{62}$$

We make two observations to establish 62. The first observation is that

$$\omega_{jn+m}^{-1}\boldsymbol{D}_{m+1}^{jn+m} \overset{(a)}{\succeq} \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m}) \overset{(b)}{\succeq} (1+\tfrac{1}{2}Mr_{jn+m+1})^{-1}\boldsymbol{K}^{jn+m},$$

where (a) follows from the induction hypothesis and (b) follows from Lemma A.3. For convenience, we restate that $r_t \coloneqq \|\boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}\|_{\boldsymbol{z}_{i_t}^{t-1}}$.

For the next observation, we first bound r_{jn+m+1} as we did in Corollary E.1, in the following manner:

$$\begin{aligned} r_{jn+m+1} = \|\boldsymbol{z}_{m+1}^{jn+m+1} - \boldsymbol{z}_{m+1}^{jn+m}\|_{\boldsymbol{z}_{m+1}^{jn+m}} &\overset{(a)}{\leq} \sqrt{L}\|\boldsymbol{z}_{m+1}^{jn+m+1} - \boldsymbol{z}_{m+1}^{jn+m}\|, \\ &= \sqrt{L}\|\boldsymbol{z}_{m+1}^{jn+m+1} - \boldsymbol{x}^\star - \boldsymbol{z}_{m+1}^{jn+m} + \boldsymbol{x}^\star\|, \\ &\overset{\Delta}{\leq} \sqrt{L}\left(\|\boldsymbol{z}_{m+1}^{jn+m+1} - \boldsymbol{x}^\star\| + \|\boldsymbol{z}_{m+1}^{jn+m} - \boldsymbol{x}^\star\|\right), \\ &\overset{(b)}{\leq} \sqrt{L}\big(\rho^{j+1} + \rho^j\big)\epsilon \leq 2\sqrt{L}\rho^j\epsilon = \frac{2\alpha_j}{M}, \end{aligned}$$

where (a) follows from Assumption **A1** which implies $\nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m}) \preceq L\boldsymbol{I}$, (b) follows from the induction hypothesis and (61). Therefore, our second observation is that $\frac{Mr_{jn+m+1}}{2} \leq \alpha_j$.

We consider two cases depending on m (or equivalently t), similar to Stage 1.

Case 1: $0 \leq m < n$

Since all the $\boldsymbol{D}_i$'s were scaled by a factor $(1+\alpha_j)^2$ at the end of the cycle $j-1$, i.e., at $t = jn$, we have $\boldsymbol{D}_{m+1}^{jn+m} = (1+\alpha_j)^2\boldsymbol{D}_{m+1}^{(j-1)n+m+1}$. Also, from the induction hypothesis, we have

$$\omega_{(j-1)n+m+1}^{-1}\boldsymbol{D}_{m+1}^{(j-1)n+m+1} \succeq \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{(j-1)n+m+1}).$$

Note that if $(j-1)n+m+1 < 0$, we can simply assume those quantities to be super scripted/sub scripted (as appropriate) with 0. For example, $\boldsymbol{D}_{m+1}^{-n+m+1} = \boldsymbol{D}_{m+1}^0, \omega_{-n+m+1} = \omega_0$, etc.

Since $0 \leq m < n$, we have $\omega_{(j-1)n+m+1} = 1$. Further, $\boldsymbol{z}_{m+1}^{jn+m} = \boldsymbol{z}_{m+1}^{(j-1)n+m+1}$. Therefore,

$$\begin{aligned} \boldsymbol{D}_{m+1}^{jn+m} = (1+\alpha_j)^2\boldsymbol{D}_{m+1}^{(j-1)n+m+1} &\succeq (1+\alpha_j)^2\nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m}), \\ &\overset{Lem.A.2}{\succeq} \big(1+\alpha_j\big)^2\big(1+\frac{Mr_{jn+m+1}}{2}\big)^{-1}\boldsymbol{K}^{jn+m+1}, \\ &\succeq \big(1+\alpha_j\big)\boldsymbol{K}^{jn+m+1}. \end{aligned}$$

Case 2: $m = n-1$

Since the current index $m+1$ was last updated at time $t = jn$, we have $\boldsymbol{D}_{m+1}^{jn+m} = \boldsymbol{D}_{m+1}^{jn}$ and $\boldsymbol{z}_{m+1}^{jn+m} = \boldsymbol{z}_{m+1}^{jn}$. Further, the induction hypothesis yields $\omega_{jn}^{-1}\boldsymbol{D}_{m+1}^{jn} \succeq \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn})$. Also, $\omega_{jn} = (1+\alpha_j)^2$ by definition. Therefore,

$$\begin{aligned} \boldsymbol{D}_{m+1}^{jn+m} = \omega_{jn}(\omega_{jn}^{-1}\boldsymbol{D}_{m+1}^{jn}) \succeq (1+\alpha_j)^2\nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m}). \\ \overset{Lem.A.3}{\succeq} \big(1+\alpha_j\big)^2\big(1+\frac{Mr_{jn+m+1}}{2}\big)^{-1}\boldsymbol{K}^{jn+m+1}, \\ \succeq \big(1+\alpha_j\big)\boldsymbol{K}^{jn+m+1}. \end{aligned}$$

Summarizing, for both cases $0 \leq m < n$ and $m = n-1$, we have established that $\boldsymbol{D}_{m+1}^{jn+m} \succeq (1+\alpha_j)\boldsymbol{K}^{jn+m+1}$. The next steps are common for both the cases.

Since $\boldsymbol{D}_{m+1}^{jn+m} \succeq \big(1+\alpha_j\big)\boldsymbol{K}^{jn+m+1}$, applying Lemma A.2, we obtain

$$\begin{aligned}\boldsymbol{Q}^{jn+m+1} &= \text{BFGS}(\boldsymbol{D}^{jn+m}, \big(1+\alpha_j\big)\boldsymbol{K}^{jn+m+1}, \boldsymbol{z}_{m+1}^{jn+m+1} - \boldsymbol{z}_{m+1}^{jn+m+1}),\\ &\overset{Lem.A.2}{\succeq} \big(1+\alpha_j\big)\boldsymbol{K}^{jn+m+1}.\end{aligned}$$

Applying Lemma A.3 to relate $\boldsymbol{K}^{jn+m+1}$ and $\nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m})$, we obtain

$$\boldsymbol{Q}^{jn+m+1} \succeq \big(1+\alpha_j\big)\boldsymbol{K}^{jn+m+1} \succeq \big(1+\frac{Mr_{jn+m+1}}{2}\big)\boldsymbol{K}^{jn+m+1} \overset{Lem.A.2}{\succeq} \nabla^2 f_{i_m}(\boldsymbol{z}_{i_m}^m).$$

Since $\boldsymbol{Q}^{jn+m+1} \succeq \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m+1})$, applying Lemma A.2, we obtain

$$\begin{aligned}\omega_{jn+m+1}^{-1}\boldsymbol{D}_{m+1}^{jn+m+1} &= \text{BFGS}(\boldsymbol{Q}^{jn+m+1}, \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m+1}), \bar{\boldsymbol{u}}(\boldsymbol{Q}^{jn+m+1}, \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m+1})),\\ &\overset{Lem.A.2}{\succeq} \nabla^2 f_{m+1}(\boldsymbol{z}_{m+1}^{jn+m+1}).\end{aligned}$$

We can now prove the linear convergence of the Hessian approximation.

We define the sequence $\{\boldsymbol{y}_k\}$, for $k = 0, \ldots, j+1$, such that $\{\boldsymbol{y}_k\} = \{\boldsymbol{x}^0, \boldsymbol{z}_{m+1}^{m+1}, \ldots, \boldsymbol{z}_{m+1}^{jn+m+1}\}$. From the induction hypothesis and Stage 2, we have that $\|\boldsymbol{y}_k - \boldsymbol{x}^\star\| \leq \rho^k \|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|$, for all k. Since $\{\boldsymbol{y}_k\}$ comes about from the application of BFGS updates with, $r_k := 2\sqrt{L}\rho^{k-1}\epsilon$, as described in the statement of Lemma B.3, therefore $\{\boldsymbol{y}_k\}$ satisfies the conditions of Lemma B.3. This implies that,

$$\sigma(\omega_{jn+m+1}^{-1}\boldsymbol{D}_{m+1}^{jn+m+1}, \nabla^2 f_{m+1}(\boldsymbol{y}_{j+1})) \leq (1-c)^{j+1}\delta.$$

Since $\boldsymbol{y}_{j+1} = \boldsymbol{z}_{m+1}^{jn+m+1}$, the proof is complete via induction.

G.3 Proof of Lemma 3

We prove the Lemma for a generic iteration $t = jn + m + 1$, for some $j \geq 0$ and $0 \leq m < n$. We restate a few observations derived in the proof of Lemma 2. First, we proved an upper bound on $\Gamma^{jn+m} = \|\big(\sum_{i=1}^n \boldsymbol{D}_i^{jn+m}\big)^{-1}\|$, given by

$$\Gamma^{jn+m} \leq \frac{1+\rho}{n\mu}. \tag{63}$$

We also derived upper bounds (57), (58) on $\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\|$:

$$\Big\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\Big\| \leq \begin{cases}(1+M\sqrt{L}\epsilon\rho^j)^2 L\delta(1-c)^j + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j) & i \in [n]\backslash[m],\\ L\delta(1-c)^{j+1} & i \in [m].\end{cases}$$

A common larger upper bound for both the cases is given by

$$\Big\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\Big\| \leq (1+M\sqrt{L}\epsilon\rho^j)^2 L\delta(1-c)^j + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j),\ \ i \in [n]. \tag{64}$$

Finally, we also established that,

$$\begin{aligned}&\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\| \leq \rho^{j+1}\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\| \leq \rho^j\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|,\ \ i \in [m]\\ &\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\| \leq \rho^j\|\boldsymbol{x}^0 - \boldsymbol{x}^\star\|,\ \ i \in [n]\backslash[m].\end{aligned} \tag{65}$$

We are now ready to prove the mean-superlinear convergence. From Lemma 1, we have

$$\begin{aligned}\|\boldsymbol{z}_{m+1}^{jn+m+1} - \boldsymbol{x}^\star\| &\leq \Gamma^{jn+m}\tfrac{\tilde{L}}{2}\sum_{i=1}^n \|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\|^2 + \Gamma^{jn+m}\sum_{i=1}^n \Big\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\Big\|\,\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\|,\\ &\leq \Gamma^{jn+m}\sum_{i=1}^n \Big(\tfrac{\tilde{L}}{2}\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\| + \Big\|\boldsymbol{D}_i^{jn+m} - \nabla^2 f_i(\boldsymbol{z}_i^{jn+m})\Big\|\Big)\|\boldsymbol{z}_i^{jn+m} - \boldsymbol{x}^\star\|,\end{aligned}$$

We now substitute the upper bounds from 63, 64, and 65 to get

$$\leq \tfrac{1+\rho}{n\mu}\big(\tfrac{\tilde{L}}{2}\rho^j\epsilon + (1+M\sqrt{L}\epsilon\rho^j)^2 L\delta(1-c)^j + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon\rho^j)\big)\sum_{i=1}^{n}\|\boldsymbol{z}_i^{jn+m}-\boldsymbol{x}^\star\|,$$

$$\leq \tfrac{1+\rho}{n\mu}\big(\tfrac{\tilde{L}}{2}\rho^j\epsilon + (1+M\sqrt{L}\epsilon)^2 L\delta(1-c)^j + ML^{\frac{3}{2}}\epsilon\rho^j(2+M\sqrt{L}\epsilon)\big)\sum_{i=1}^{n}\|\boldsymbol{z}_i^{jn+m}-\boldsymbol{x}^\star\|,$$

$$\overset{(a)}{\leq} (1-c)^j\tfrac{1+\rho}{\mu}\big(\tfrac{\tilde{L}}{2}\epsilon + (1+M\sqrt{L}\epsilon)^2 L\delta + ML^{\frac{3}{2}}\epsilon(2+M\sqrt{L}\epsilon)\big)\big(\tfrac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{z}_i^{jn+m}-\boldsymbol{x}^\star\|\big),$$

$$\overset{(52)}{\leq} \rho(1-c)^j\big(\tfrac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{z}_i^{jn+m}-\boldsymbol{x}^\star\|\big) \overset{(b)}{\leq} (1-c)^{j+1}\big(\tfrac{1}{n}\sum_{i=1}^{n}\big\|\boldsymbol{x}^{t-i}-\boldsymbol{x}^\star\big\|\big),$$

where (a) and (b) follow since $\rho < 1-c$. This completes the proof.

G.4 Proof of Theorem 1

Define the sequence $\{\zeta^t\}$, for all $t \geq 0$, as $\zeta^t := \max_{j\in[n]}\|\boldsymbol{x}^{nt+j}-\boldsymbol{x}^\star\|$. Let the constant c be defined as $c := \frac{\mu}{dL}$. Then, from Lemma 3, we have

$$\big\|\boldsymbol{x}^{nt+i}-\boldsymbol{x}^\star\big\| \leq (1-c)^{t+1}\tfrac{1}{n}\sum_{j=1}^{n}\|\boldsymbol{x}^{nt+i-j}-\boldsymbol{x}^\star\| \tag{66}$$

$$\leq (1-c)^{t+1}\max_{j\in[n]}\|\boldsymbol{x}^{nt+i-j}-\boldsymbol{x}^\star\|. \tag{67}$$

By induction on i, we prove the following first:

$$\big\|\boldsymbol{x}^{nt+i}-\boldsymbol{x}^\star\big\| \leq (1-c)^{t+1}\max_{j=1,\dots,n}\big\|\boldsymbol{x}^{nt+1-j}-\boldsymbol{x}^\star\big\| = (1-c)^{t+1}\zeta^{t-1}, \tag{68}$$

for $i = 1,\dots,n$. Substituting $i = 1$ in (67) we get

$$\big\|\boldsymbol{x}^{nt+1}-\boldsymbol{x}^\star\big\| \leq (1-c)^{t+1}\max_{j=1,\dots,n}\big\|\boldsymbol{x}^{nt+1-j}-\boldsymbol{x}^\star\big\| = (1-c)^{t+1}\zeta^{t-1}.$$

This proves the base step ($i = 1$). Assume, (68) holds for $i = k$. We prove that (68) holds for $i = k+1$. Substituting $i = k+1$ in (67) we get

$$\begin{aligned}\big\|\boldsymbol{x}^{nt+k+1}-\boldsymbol{x}^\star\big\| &\leq (1-c)^{t+1}\max_{j=1,\dots,n}\big\|\boldsymbol{x}^{nt+k+1-j}-\boldsymbol{x}^\star\big\|,\\ &\leq (1-c)^{t+1}\max_{j=1,\dots,n+k}\big\|\boldsymbol{x}^{nt+k+1-j}-\boldsymbol{x}^\star\big\|,\\ &= (1-c)^{t+1}\max\big(\big\|\boldsymbol{x}^{nt+k}-\boldsymbol{x}^\star\big\|,\dots,\big\|\boldsymbol{x}^{nt+1}-\boldsymbol{x}^\star\big\|,\max_{j=1,\dots,n}\big\|\boldsymbol{x}^{nt+1-j}-\boldsymbol{x}^\star\big\|\big),\\ &\overset{(a)}{\leq} (1-c)^{t+1}\max_{j=1,\dots,n}\big\|\boldsymbol{x}^{nt+1-j}-\boldsymbol{x}^\star\big\|,\end{aligned}$$

where (a) follows since $\big\|\boldsymbol{x}^{nt+1}-\boldsymbol{x}^\star\big\| \leq (1-c)^{t+1}\max_{j=1,\dots,n}\big\|\boldsymbol{x}^{nt+1-j}-\boldsymbol{x}^\star\big\|,\dots,\big\|\boldsymbol{x}^{nt+k}-\boldsymbol{x}^\star\big\| \leq (1-c)^{t+1}\max_{j=1,\dots,n}\big\|\boldsymbol{x}^{nt+1-j}-\boldsymbol{x}^\star\big\|$ by the induction hypothesis. Therefore (68) holds for $i = k+1$. This proves (68) for $i = 1,\dots,n$.

Since (67) holds for $i = 1,\dots,n$, we have

$$\zeta^t \leq (1-c)^{t+1}\zeta^{t-1}, \tag{69}$$

for all $t \geq 1$. Unrolling this recursion, we get

$$\zeta^t \leq (1-c)^{\sum_{j=2}^{t+1}j}\zeta^0 \overset{(a)}{\leq} (1-c)^{\frac{t(t+3)}{2}}\rho\big\|\boldsymbol{x}^0-\boldsymbol{x}^\star\big\| \overset{(b)}{\leq} \epsilon(1-c)^{\frac{(t+1)(t+2)}{2}},$$

where (a) follows from Lemma 2 and (b) follows since $\rho < 1-c$. This completes the proof.

H GENERALIZED SHARPENED INCREMENTAL QUASI-NEWTON METHOD (G-SLIQN)

In this section, we extend the SLIQN algorithm, whose Hessian approximation updates 3, 3 are built on the BFGS operator, to the class of *restricted Broyden* operators. We refer to this class of algorithms as G-SLIQN. First, we define the DFP operator

$$\mathrm{DFP}(\boldsymbol{B},\boldsymbol{K},\boldsymbol{z}) \coloneqq \boldsymbol{B} - \frac{\boldsymbol{K}\boldsymbol{z}\boldsymbol{z}^\mathsf{T}\boldsymbol{B} + \boldsymbol{B}\boldsymbol{z}\boldsymbol{z}^\mathsf{T}\boldsymbol{K}}{\langle \boldsymbol{z}, \boldsymbol{K}\boldsymbol{z}\rangle} + \left(1 + \frac{\langle \boldsymbol{z}, \boldsymbol{B}\boldsymbol{z}\rangle}{\langle \boldsymbol{z}, \boldsymbol{K}\boldsymbol{z}\rangle}\right)\frac{\boldsymbol{K}\boldsymbol{z}\boldsymbol{z}^\mathsf{T}\boldsymbol{K}}{\langle \boldsymbol{z}, \boldsymbol{K}\boldsymbol{z}\rangle},$$

for $\boldsymbol{B}, \boldsymbol{K} \succ \boldsymbol{0}$ and $\boldsymbol{z} \in \mathbb{R}^d \backslash \{\boldsymbol{0}\}$. The restricted Broyden operator with parameter $0 \leq \tau \leq 1$ is defined as a convex combination of DFP and BFGS as follows:

$$\mathrm{Broyd}_\tau^{res}(\boldsymbol{B},\boldsymbol{K},\boldsymbol{z}) \coloneqq \tau \mathrm{DFP}\left(\boldsymbol{B},\boldsymbol{K},\boldsymbol{z}\right) + (1-\tau)\mathrm{BFGS}\left(\boldsymbol{B},\boldsymbol{K},\boldsymbol{z}\right),$$

where $\mathrm{BFGS}(\boldsymbol{B},\boldsymbol{K},\boldsymbol{z})$ is given by (3).

H.1 The generalized algorithm G-SLIQN

As with SLIQN, we first present G-SLIQN with a maximum per-iteration cost of $\mathcal{O}(nd^2)$ and an average $\mathcal{O}(d^2)$ cost per epoch. Similar to SLIQN, G-SLIQN can be implemented with a maximum per-iteration cost of $\mathcal{O}(d^2)$ as per Appendix F.1.

Hyperparameters: Choose $\tau_1, \tau_2 \in [0,1]$ as the restricted Broyd operator paramater for the classic and the greedy updates respectively. Note that setting $\tau_1 = \tau_2 = 0$ reduces G-SLIQN to SLIQN.

We denote the Hessian approximation matrices at time t as $\{\boldsymbol{G}_i^t\}_{i=1}^n$.

Initialize: At $t=0$, we initialize $\{\boldsymbol{z}_i^0\}_{i=1}^n$ as $\boldsymbol{z}_i^0 = \boldsymbol{x}^0$, $\forall i \in [n]$, for a suitably chosen $\boldsymbol{x}^0$. We initialize $\{\boldsymbol{G}_i^0\}_{i=1}^n$ as $\boldsymbol{G}_i^0 = (1+\alpha_0)^2\boldsymbol{I}_i^0$, where $\{\boldsymbol{I}_i^0\}_{i=1}^n$ are chosen such that $\boldsymbol{I}_i^0 \succeq \nabla^2 f_i(\boldsymbol{z}_i^0)$, $\forall i \in [n]$. Here $\{\alpha_k\}, k \in \mathbb{N}$ is as defined in Lemma 2.

Algorithm: For any iteration $t \geq 1$, just like the update in SLIQN 10, we update $\boldsymbol{x}^t$ as

$$\boldsymbol{x}^t = \left(\sum_{i=1}^n \boldsymbol{G}_i^{t-1}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{G}_i^{t-1}\boldsymbol{z}_i^{t-1} - \sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^{t-1})\right). \tag{70}$$

Next, we update $\boldsymbol{z}_{i_t}^t$ as $\boldsymbol{z}_{i_t}^t = \boldsymbol{x}^t$. To update $\boldsymbol{Q}^t$ and $\boldsymbol{G}_{i_t}^t$, we use the chosen restricted Broyd operators in place of the BFGS operators 3, 3:

$$\boldsymbol{Q}^t = \mathrm{Broyd}_{\tau_1}^{res}(\boldsymbol{G}_{i_t}^{t-1}, (1+\alpha_{\lceil t/n-1\rceil})\boldsymbol{K}^t, \boldsymbol{z}_{i_t}^t - \boldsymbol{z}_{i_t}^{t-1}), \tag{71}$$

$$\boldsymbol{G}_{i_t}^t = \omega_t \mathrm{Broyd}_{\tau_2}^{res}(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t), \bar{\boldsymbol{u}}^t(\boldsymbol{Q}^t, \nabla^2 f_{i_t}(\boldsymbol{z}_{i_t}^t))), \tag{72}$$

where $\omega_t \coloneqq (1+\alpha_{\lceil t/n\rceil})^2$ if t is a multiple of n and 1 otherwise. For the indices $i \neq i_t$, we update $\boldsymbol{z}_i^t$ and $\boldsymbol{G}_i^t$ in the following manner:

$$\boldsymbol{z}_i^t = \boldsymbol{z}_i^{t-1}, \boldsymbol{G}_i^t = \omega_t \boldsymbol{G}_i^{t-1}, \forall i \in [n]; i \neq i_t. \tag{73}$$

Finally, we update $\left(\sum_{i=1}^n \boldsymbol{G}_i^t\right)^{-1}$, $\sum_{i=1}^n \boldsymbol{G}_i^t\boldsymbol{z}_i^t$ and $\sum_{i=1}^n \nabla f_i(\boldsymbol{z}_i^t)$. Observe that the restricted Broyd operator induces a correction of at-most 5 rank-1 matrices. We can therefore, carry out this update in $\mathcal{O}(d^2)$ cost by repeatedly applying Sherman-Morrison formula. This is similar to what was done for SLIQN in Appendix F.1. All other updates are a constant number of matrix-vector multiplications which can be done in $\mathcal{O}(d^2)$ cost.

H.2 Overview of the Convergence Analysis of G-SLIQN

The analysis of SLIQN can be readily extended to G-SLIQN. Since most of the results established for SLIQN would continue to hold for G-SLIQN, we do not explictly state them here for the sake of brevity. However, we discuss the mappings that allow us to conclude that similar results hold for SLIQN in this section.

Algorithm 4 Generalized Sharpened Lazy Incremental Quasi-Newton (G-SLIQN)

Function {Sherman-Morrison} $\{\boldsymbol{A}^{-1}, \boldsymbol{u}, \boldsymbol{v}\}$
  return $\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{A}^{-1}}{1+\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}$
EndFunction

Initialize: Initialize $\{\boldsymbol{z}_i, \boldsymbol{G}_i\}_{i=1}^n$ as described in Section H.1;
Evaluate $\bar{\boldsymbol{G}} \coloneqq \left(\sum_i \boldsymbol{G}_i\right)^{-1}, \boldsymbol{\phi} \coloneqq \sum_i \boldsymbol{G}_i\boldsymbol{z}_i$, and $\boldsymbol{g} \coloneqq \sum_i \nabla f_i(\boldsymbol{z}_i)$;
while *not converged:*
  Current index to be updated is $i_t \leftarrow (t-1) \mod n + 1$;
  Update ω_t as $\omega_t \leftarrow 1 + \alpha_{\lceil t/n-1 \rceil}$;
if $i_t = 1$ **then**
  Update $\bar{\boldsymbol{G}}$ as $\bar{\boldsymbol{G}} \leftarrow \bar{\boldsymbol{G}}/\omega_t^2$;
  Update $\boldsymbol{\phi}$ as $\boldsymbol{\phi} \leftarrow \omega_t^2\boldsymbol{\phi}$;
end if
  Update $\boldsymbol{x}^t$ as $\boldsymbol{x}^t \leftarrow \left(\bar{\boldsymbol{G}}\right)\left(\boldsymbol{\phi} - \boldsymbol{g}\right)$ as per (70);
  Update $\boldsymbol{G}_{i_t}$ as $\boldsymbol{G}_{i_t} \leftarrow \omega_t^2\boldsymbol{G}_{i_t}$;
  Update $\boldsymbol{v}_1$ as $\boldsymbol{v}_1 \leftarrow \boldsymbol{x}^t - \boldsymbol{z}_{i_t}$;
  Update $\boldsymbol{Q}_{i_t}$ as $\boldsymbol{Q}_{i_t} \leftarrow \text{Broyd}_{\tau_1}^{res}(\boldsymbol{G}_{i_t}, \omega_t\boldsymbol{K}^t, \boldsymbol{v}_1)$ as per (71);
  Update $\boldsymbol{v}_2$ as $\boldsymbol{v}_2 \leftarrow \bar{\boldsymbol{u}}^t(\boldsymbol{Q}_{i_t}, \nabla^2 f_{i_t}(\boldsymbol{x}^t))$;
  Update $\tilde{\boldsymbol{G}}_{i_t}$ as $\tilde{\boldsymbol{G}}_{i_t} \leftarrow \text{Broyd}_{\tau_2}^{res}(\boldsymbol{Q}_{i_t}, \nabla^2 f_{i_t}(\boldsymbol{x}^t), \boldsymbol{v}_2)$ as per (72);
  Update $\boldsymbol{\phi}$ as $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} - \boldsymbol{G}_{i_t}\boldsymbol{z}_{i_t} + \tilde{\boldsymbol{G}}_{i_t}\boldsymbol{x}^t$;
  Update $\boldsymbol{g}$ as $\boldsymbol{g} \leftarrow \boldsymbol{g} - \nabla f_{i_t}(\boldsymbol{z}_{i_t}) + \nabla f_{i_t}(\boldsymbol{x}^t)$;
  Update $\bar{\boldsymbol{G}}$ as $\bar{\boldsymbol{G}} \leftarrow \text{Sherman-Morrison}(\bar{\boldsymbol{G}}, -\frac{\boldsymbol{K}\boldsymbol{v}_1}{\boldsymbol{v}_1^T\boldsymbol{K}\boldsymbol{v}_1}, \boldsymbol{G}_{i_t}\boldsymbol{v}_1)$;
  Update $\bar{\boldsymbol{G}}$ as $\bar{\boldsymbol{G}} \leftarrow \text{Sherman-Morrison}(\bar{\boldsymbol{G}}, -\boldsymbol{G}_{i_t}\boldsymbol{v}_1, \frac{\boldsymbol{K}\boldsymbol{v}_1}{\boldsymbol{v}_1^T\boldsymbol{K}\boldsymbol{v}_1})$;
  Update $\bar{\boldsymbol{G}}$ as $\bar{\boldsymbol{G}} \leftarrow \text{Sherman-Morrison}(\bar{\boldsymbol{G}}, \omega_t\left(1 + \frac{\boldsymbol{v}_1^T\boldsymbol{G}_{i_t}\boldsymbol{v}_1}{\omega_t^2\boldsymbol{v}_1^T\boldsymbol{K}\boldsymbol{v}_1}\right)\frac{\boldsymbol{K}\boldsymbol{v}_1}{\boldsymbol{v}_1^T\boldsymbol{K}\boldsymbol{v}_1}, \frac{\boldsymbol{K}\boldsymbol{v}_1}{\boldsymbol{v}_1^T\boldsymbol{K}\boldsymbol{v}_1})$;
  Update $\bar{\boldsymbol{G}}$ as $\bar{\boldsymbol{G}} \leftarrow \text{Sherman-Morrison}(\bar{\boldsymbol{G}}, -\frac{\boldsymbol{Q}_{i_t}\boldsymbol{v}_2}{\boldsymbol{v}_2^T\boldsymbol{Q}_{i_t}\boldsymbol{v}_2}, \boldsymbol{Q}_{i_t}\boldsymbol{v}_2)$;
  Update $\bar{\boldsymbol{G}}$ as $\bar{\boldsymbol{G}} \leftarrow \text{Sherman-Morrison}(\bar{\boldsymbol{G}}, \frac{\nabla^2 f_{i_t}(\boldsymbol{x}^t)\boldsymbol{v}_2}{\boldsymbol{v}_2^T\nabla^2 f_{i_t}(\boldsymbol{x}^t)\boldsymbol{v}_2}, \nabla^2 f_{i_t}(\boldsymbol{x}^t)\boldsymbol{v}_2)$;
  Update $\boldsymbol{z}_{i_t}$ as $\boldsymbol{z}_{i_t} \leftarrow \boldsymbol{x}^t$;
Increment the iteration counter t;
end while

Firstly, the result in Lemma A.2 holds even for the restricted Broyden operator as per (Rodomanov and Nesterov, 2021c, Lemma 2.1) and (Rodomanov and Nesterov, 2021c, Lemma 2.2), which are restated here for completeness. Note that the results in Rodomanov and Nesterov (2021c) are for the Broyden operator, but as per Nocedal and Wright (1999) the restricted Broyden operator is a subset of the Broyden operator, hence the results in Rodomanov and Nesterov (2021c) are applicable for the restricted Broyden operator as well.

Lemma H.1 *(Rodomanov and Nesterov, 2021c, Lemma 2.1) Let, $\boldsymbol{G}, \boldsymbol{A}$ be positive definite matrices such that $\frac{1}{\xi}\boldsymbol{A} \preceq \boldsymbol{G} \preceq \boldsymbol{A}$, where $\xi, \eta \geq 1$. Then, for any $\boldsymbol{u} \neq 0$, and any $\tau \in [0,1]$, we have*

$$\frac{1}{\xi}\boldsymbol{A} \preceq \boldsymbol{G}_+ := \text{Broyd}_\tau^{res}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) \preceq \eta \boldsymbol{A}.$$

Lemma H.2 *(Rodomanov and Nesterov, 2021c, Lemma 2.2) Let $\boldsymbol{G}, \boldsymbol{A}$ be positive definite matrices such that $\boldsymbol{A} \preceq \boldsymbol{G} \preceq \eta \boldsymbol{A}$, for some $\eta \geq 1$. Then, for any $\tau \in [0,1]$ and any $\boldsymbol{u} \neq \boldsymbol{0}$, we have*

$$\sigma(\boldsymbol{G}, \boldsymbol{A}) - \sigma(\text{Broyd}_\tau^{res}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}), \boldsymbol{A}) \geq \left(\frac{\tau}{\eta} + 1 - \tau\right)\theta^2(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}),$$

where

$$\theta(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}) := \left(\frac{\langle (\boldsymbol{G} - \boldsymbol{A})\boldsymbol{u}, \boldsymbol{A}^{-1}(\boldsymbol{G} - \boldsymbol{A})\boldsymbol{u}\rangle}{\langle \boldsymbol{G}\boldsymbol{u}, \boldsymbol{A}^{-1}\boldsymbol{G}\boldsymbol{u}\rangle}\right)^{\frac{1}{2}}.$$

Since $\tau \in [0,1]$ and $\eta \geq 1$, we have $\frac{\tau}{\eta} + 1 - \tau \geq 0 \iff \tau \leq \frac{\eta}{\eta - 1}$ holds vacuously. Therefore, on taking a restricted Broyden update, we get $\sigma(\boldsymbol{G}, \boldsymbol{A}) \geq \sigma(\text{Broyd}_\tau^{res}(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{u}), \boldsymbol{A})$ under the assumptions of Lemma H.2.

Futher, the result in Lemma A.4 holds for the restricted Broyd operator as per (Rodomanov and Nesterov, 2021a, Theorem 2.5) which is restated here for completeness.

Lemma H.3 *(Rodomanov and Nesterov, 2021a, Theorem 2.5) Let $\boldsymbol{G}, \boldsymbol{A}$ be positive definite matrices such that $\boldsymbol{A} \preceq \boldsymbol{G}$. Further, let $\mu, L > 0$ be such that $\mu \boldsymbol{I} \preceq \boldsymbol{A} \preceq L\boldsymbol{I}$. Then, for any $\tau \in [0,1]$, we have*

$$\sigma(\text{Broyd}_\tau^{res}(\boldsymbol{G}, \boldsymbol{A}, \bar{\boldsymbol{u}}(\boldsymbol{G}, \boldsymbol{A})), \boldsymbol{A}) \leq \left(1 - \frac{\mu}{dL}\right)\sigma(\boldsymbol{G}, \boldsymbol{A}),$$

where $\bar{\boldsymbol{u}}(\boldsymbol{G}, \boldsymbol{A})$ the greedy vector (4)*.*

Using Lemma H.1, H.2, H.3, we can establish that the Lemma B.2 (with BFGS replaced by restricted Broyd), Corollary B.1, and Lemma B.3 (with BFGS replaced by restricted Broyd) hold. Therefore, the supporting lemmas in Appendix B hold even for the Broyden update.

Next, we discuss about the main results in Section 5. Firstly, observe that Lemma 1 remains the same for G-SLIQN. This is because the proof of Lemma 1 hinges on the structure of update 10 and and abstracts out the specific updates made to $\boldsymbol{D}_{i_t}^t$. Since the update 70 for G-SLIQN is the same as 10, the guarantees of the Lemma and its proof carries through for G-SLIQN. Further, since the supporting lemmas in Appendix B hold for G-SLIQN, using Lemma 1 and the supporting lemmas, we can establish that Lemma 2 holds even for G-SLIQN. Since the proof of Lemma 3 leverages the result of Lemma 2, we can establish that the mean superlinear convergence result given by Lemma 3 holds for G-SLIQN as well. Finally, using Lemma 3 we can show that Theorem 1 holds for G-SLIQN.

I NUMERICAL SIMULATIONS

As can be clearly observed, the proposed algorithm SLIQN requires the knowledge about ϵ, σ_0 in order to tune the correction factor α_t. However, we observed that, empirically SLIQN outperforms a number of incremental and stochastic QN methods without the correction factor, i.e., $\alpha_t = 0$. For IGS however, the performance is quite sensitive to the correction factor, β_t, and $\beta_t = 0$ was not the best performing correction factor for all the simulations. Therefore, SLIQN does not require hyper-parameter tuning, unlike IGS.

Initialization: For all our simulations, all algorithms start at the same initial $\boldsymbol{x}_0 = \alpha \boldsymbol{v}$, where $\boldsymbol{v} \in \mathbb{R}^d$ is such that each coordinate $v_{i,j} \sim \text{Unif}[0, 1]$. Since, all the algorithms considered for performance comparison are only locally convergent, the parameter α affects the convergence of the algorithms.

Stopping Criterion: We stop the execution of each algorithm when the gradient norm of f is less than a threshold. Formally, letting the threshold be *gstop*, the stopping condition can be expressed as

$$\frac{1}{N} \left\| \sum_{i=1}^{N} \nabla f_i(\boldsymbol{x}^t) \right\| < gstop.$$

Typical values of *gstop* used in our simulations range from 10^{-7} to 10^{-8}.

I.1 Generating Scheme for Quadratic minimization

We follow the scheme proposed in Mokhtari et al. (2018) to generate $\{\boldsymbol{A}_i, \boldsymbol{b}_i\}_{i=1}^n$. We set each matrix $\boldsymbol{A}_i \coloneqq \text{diag}(\{\boldsymbol{a}_i\}_{i=1}^d)$, by sampling the diagonal elements as $\{\boldsymbol{a}_i\}_{i=1}^{d/2} \overset{\text{i.i.d.}}{\sim} \text{Unif}[1, 10^{\frac{\xi}{2}}]$ and $\{\boldsymbol{a}_i\}_{i=d/2+1}^{d} \overset{\text{i.i.d.}}{\sim} \text{Unif}[10^{-\frac{\xi}{2}}, 1]$. The parameter ξ controls the condition number of the matrix $\boldsymbol{A}_i$. Under the limit $d \to \infty$, the condition number of $\boldsymbol{A}_i$ is given by 10^{ξ}. Each coordinate $b_{i,j}$ of the vector $\boldsymbol{b}_i$ is sampled as $b_{i,j} \sim \text{Unif}[0, 1000]$.