Journal of Machine Learning Research 26 (2025) 1-49 Submitted 3/25; Revised 8/25; Published 11/25

Second-order Information Promotes Mini-Batch Robustness in Variance-Reduced Gradients

Sachin Garg SACHG@UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109, USA

Albert S. Berahas ABERAHAS@UMICH.EDU
Department of Industrial and Operations Engineering
University of Michigan
Ann Arbor, MI 48109, USA

Michał Dereziński DEREZIN@UMICH.EDU
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Lam Nguyen

Abstract

We show that, for finite-sum minimization problems, incorporating partial second-order information of the objective function can dramatically improve the robustness to mini-batch size of variance-reduced stochastic gradient methods, making them more scalable while retaining their benefits over traditional Newton-type approaches. We demonstrate this phenomenon on a prototypical stochastic second-order algorithm, called Mini-Batch Stochastic Variance-Reduced Newton (`Mb-SVRN`), which combines variance-reduced gradient estimates with access to an approximate Hessian oracle. In particular, we show that when the data size n is sufficiently large, i.e., $n \gg \alpha^2\kappa$, where κ is the condition number and α is the Hessian approximation factor, then `Mb-SVRN` achieves a fast linear convergence rate that is independent of the gradient mini-batch size b, as long b is in the range between 1 and $b_{\max} = O(n/(\alpha \log n))$. Only after increasing the mini-batch size past this critical point $b_{\max}$, the method begins to transition into a standard Newton-type algorithm which is much more sensitive to the Hessian approximation quality. We verify this empirically on benchmark tasks showing that, after tuning the step size, the convergence rate of `Mb-SVRN` remains fast for a wide range of mini-batch sizes, and the dependence of the phase transition point $b_{\max}$ on the Hessian approximation factor α agrees with our theory.

Keywords: Finite-sum minimization, Stochastic gradient, Newton-type methods, Variance reduction, Robustness

1. Introduction

Consider the following finite-sum convex minimization problem:

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \psi_i(\mathbf{x}), \qquad \text{with} \qquad \mathbf{x}^* \in \operatorname{argmin} f(\mathbf{x}). \tag{1}$$

The finite-sum formulation given in (1) can be used to express many empirical minimization tasks, where for a choice of the underlying parameter vector $\mathbf{x} \in \mathbb{R}^d$, each $\psi_i : \mathbb{R}^d \to \mathbb{R}$ models the loss on a particular observation (Sharpe, 1989; Rigollet and Tong, 2011; Xiao and Zhang, 2014; Necoara and Singh, 2022). In modern-day settings, it is common to encounter problems with a very large number of observations n, making deterministic optimization methods prohibitively expensive (Gasnikov et al., 2019; Bottou et al., 2018). In particular, we are interested in and investigate problems where n is much larger than the condition number of the problem, κ (see Assumption 1). In this regime ($n \gg \kappa$), our aim is to develop a stochastic algorithm with guarantees of returning an ϵ-approximate solution i.e., $\tilde{\mathbf{x}}$ such that $f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*) \leq \epsilon$ with high probability. As is commonly assumed in the convex optimization literature, we consider the problem setting where f is μ-strongly convex and each ψ_i is λ-smooth.

A prototypical optimization method for (1) is stochastic gradient descent (SGD) (Robbins and Monro, 1951; Toulis and Airoldi, 2017), which relies on computing a gradient estimate based on randomly sampling a single component ψ_i, or more generally, a subset of components (i.e., a mini-batch). Unfortunately, SGD with constant step size does not converge to the optimal solution $\mathbf{x}^*$, but only to a neighborhood around $\mathbf{x}^*$ that depends on the step size and variance in the stochastic gradient approximations. A slower sub-linear convergence rate to the optimal solution can be guaranteed with diminishing step sizes. To mitigate this deficiency of SGD, several works have explored variance reduction techniques for first-order stochastic methods, e.g., `SDCA` (Shalev-Shwartz and Zhang, 2013), `SAG`, (Roux et al., 2012), `SAGA` (Defazio et al., 2014), `SVRG` (Johnson and Zhang, 2013), `Katyusha` (Allen-Zhu, 2017) and `S2GD` (Konecnỳ and Richtárik, 2013), and have derived faster linear convergence rates to the optimal solution.

One of the popular first-order stochastic optimization methods that employs variance reduction is Stochastic Variance Reduced Gradient (`SVRG`) (Johnson and Zhang, 2013). The method has two stages and operates with inner and outer loops. At every inner iteration, the method randomly selects a component gradient, $\nabla\psi_i(\mathbf{x})$, and couples this with an estimate of the true gradient computed at every outer iteration to compute a step. This combination, and the periodic computation of the true gradient at outer iterations, leads to a reduction in the variance of the stochastic gradients employed, and allows for global linear convergence guarantees. The baseline `SVRG` method (and its associated analysis) assumes only one observation is sampled at every inner iteration, and as a result, requires $\mathcal{O}(n)$ inner iterations in order to achieve the best convergence rate per data pass. This makes baseline `SVRG` inherently and highly sequential, and unable to take advantage of modern-day massively parallel computing environments. A natural remedy is to use larger mini-batches (gradient samples sizes; denoted as b) at every inner iteration. Unfortunately, this natural idea does not have the desired advantages as increasing the mini-batch size b in `SVRG` leads to deterioration in the convergence rate per data pass. In particular, one can show that regardless of the chosen mini-batch size b, `SVRG` requires as many as $\mathcal{O}(\kappa)$ inner iterations to ensure fast convergence rate, again rendering the method unable to take advantage of parallelization. On the other hand, one can use stochastic second-order methods for solving (1), including Subsampled Newton (Roosta-Khorasani and Mahoney, 2019; Bollapragada et al., 2019; Erdogdu and Montanari, 2015), Newton Sketch (Pilanci and Wainwright, 2017; Berahas et al., 2020), and others (Moritz et al., 2016; Mokhtari et al., 2018; Derezinski et al., 2018; Berahas et al., 2016). These methods use either full gradients or require extremely large gradient mini-batches $b \gg \kappa$ at every iteration, as in Dereziński (2025), and as a result, their convergence rate (per data pass) is sensitive to the quality of the Hessian estimate and to the mini-batch size b. Several works (Gonen et al., 2016; Gower et al., 2016; Liu et al., 2019;

Moritz et al., 2016) have also explored the benefits of including second-order information to accelerate first-order stochastic variance-reduced methods. However, these are primarily first-order stochastic methods (and analyzed as such) resulting again in highly sequential algorithms that are unable to exploit modern parallel computing frameworks.

The shortcomings of the aforementioned stochastic methods raise a natural question, which is central to our work:

Can second-order information plus variance-reduction achieve an accelerated and robust convergence rate for a wide range of gradient mini-batch sizes?

Here, robustness of a stochastic method with regards to the gradient mini-batch size refers to the range of gradient mini-batch sizes for which the stochastic method can provide an ϵ-approximation solution in $\mathcal{O}(\log(1/\epsilon))$ data passes (see more below). In this work, we provide an affirmative answer to the above question. Recently, Dereziński (2025) proposed `SVRN`, combining variance-reduced gradient estimates, with second-order information to accelerate stochastic Newton-type methods. However, as we discuss in more detail in Section 3, `SVRN` (and its analysis) is not robust to deviations in the gradient mini-batch size. To this end, following up on Dereziński (2025), we analyze the convergence behavior of a similar prototypical stochastic optimization algorithm called Mini-batch Stochastic Variance-Reduced Newton (`Mb-SVRN`), and prove its robustness to the gradient mini-batch size. Similar to `SVRN`, we combine variance reduction in the stochastic gradient, via a scheme based on `SVRG` (Johnson and Zhang, 2013) with second-order information from a Hessian oracle. Our Hessian oracle is general and returns an estimate that satisfies a relatively mild Hessian approximation condition. This condition can be, for instance, satisfied by computing the Hessian of the subsampled objective (as in Subsampled Newton, Roosta-Khorasani and Mahoney, 2019, which is what we use in our experiments), but can also be satisfied by other Hessian approximation techniques such as sketching.

In our main result (Theorem 4) we show that even a relatively weak Hessian estimate leads to a dramatic improvement in robustness to the gradient mini-batch size for `Mb-SVRN`: the method is endowed with *the same* fast local linear convergence guarantee *for any* gradient mini-batch size up to a critical point $b_{\max}$, which we characterize in terms of the number of function components n and the Hessian oracle approximation factor α. As b increases beyond $b_{\max}$, the method has to inevitably transition into a standard Newton-type algorithm with a weaker convergence rate that depends much more heavily on the Hessian approximation factor α. Remarkably, unlike results for most stochastic gradient methods, our result holds with high probability rather than merely in expectation, both for large and small mini-batch sizes. It is also a direct improvement over the prior result of Dereziński (2025) both in convergence rate and the range of robustness to mini-batch sizes (see Section 3 for a comparison), thanks to relying on an entirely new submartingale analysis framework.

We verify these results empirically on the Logistic Regression task on the `EMNIST` and `CIFAR10` datasets (see Figure 1 and Section 5); we show that the convergence rate of `Mb-SVRN` indeed exhibits robustness to mini-batch size b, while at the same time having a substantially better convergence rate per data pass than the Subsampled Newton method (`SN`) that uses full gradients. Furthermore, `Mb-SVRN` proves to be remarkably robust to the Hessian approximation quality in our experiments, showing that even low-quality Hessian estimates can significantly improve the scalability of the method.

Outline. In Section 2, we provide an informal version of our main result along with a comparison to the popular `SVRG`, `Katyusha`, and `SN` methods. In Section 3, we give a detailed overview of the

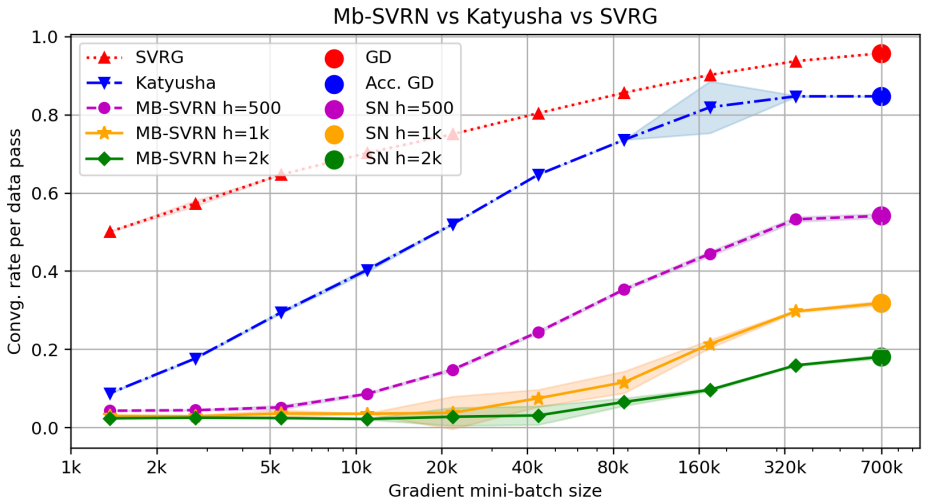


Figure 1: Convergence rate of `Mb-SVRN` (smaller is better, see Section 5), as we vary gradient mini-batch size b and Hessian sample size h, including the extreme cases of `SN` ($b=n$) and `SVRG` ($h=0$). The plot shows that after adding some second-order information (increasing h), the convergence rate of `Mb-SVRN` quickly becomes robust to gradient mini-batch size. On the other hand, the performance of `SVRG` and `Katyusha` rapidly degrades as we increase the gradient mini-batch size b, which ultimately turns them into simple gradient methods.

related work in stochastic variance reduction and second-order methods. In Section 4 we present our technical analysis for the local convergence of `Mb-SVRN`, as well as a global convergence guarantee. We provide experimental evidence in agreement with our theoretical results in Section 5.

2. Main Convergence Result

In this section, we provide an informal version of our main result, Theorem 25. We start by formalizing the assumptions and algorithmic setup.

Assumption 1 *Consider a twice continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ defined as in (1), where $\psi_i : \mathbb{R}^d \to \mathbb{R}$ and $i \in \{1, 2, ..., n\}$. We make standard λ-smoothness, μ-strong convexity, and L-Lipschitz Hessian assumptions, with $\kappa = \lambda/\mu$ as the condition number.*

1. *For each $i \in \{1, 2, ..., n\}$, ψ_i is a λ-smooth convex function:*

$$\psi_i(\mathbf{y}) \leq \psi_i(\mathbf{x}) + \nabla\psi_i(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\lambda}{2}\|\mathbf{y} - \mathbf{x}\|^2, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d,$$

   and f is a μ-strongly convex function:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|^2, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

2. *The Hessian $\nabla f^2(\mathbf{x})$ of the objective function is L-Lipschitz continuous:*

$$\left\|\nabla f^2(\mathbf{x}) - \nabla f^2(\mathbf{y})\right\| \leq L\,\|\mathbf{x} - \mathbf{y}\|, \quad \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

In our computational setup, we assume that the algorithm has access to a stochastic gradient at any point via the following gradient oracle.

Definition 1 (Gradient oracle) *Given mini-batch size $1 \leq b \leq n$ and indices $i_1, ..., i_b \in \{1, ..., n\}$, the gradient oracle $\mathcal{G}$ returns $\hat{\mathbf{g}} \sim \mathcal{G}(\{i_1, ..., i_b\})$ such that:*

$$\hat{\mathbf{g}}(\mathbf{x}) = \frac{1}{b} \sum_{j=1}^{b} \nabla \psi_{i_j}(\mathbf{x}), \qquad \text{for any } \mathbf{x} \in \mathbb{R}^d.$$

We formalize this notion of a gradient oracle to highlight that the algorithm only accesses component gradients in batches, enabling hardware acceleration for the gradient computations, which is one of the main motivations of this work. We also use the gradient oracle to measure the per-data-pass convergence rate of the algorithms. Here, a data pass refers to a sequence of gradient oracle queries that together access up to n component gradients.

We next define what we refer to as the robustness of a stochastic method to the gradient mini-batch size. Here and throughout, we focus on the $n \gg \kappa$ regime, so that a standard variance-reduced stochastic method (such as `SVRG`) with mini-batch size $b = 1$ can converge ϵ-close within $\mathcal{O}(\log(1/\epsilon))$ data passes.

Definition 2 *A stochastic method $\mathcal{S}$ is said to be r-robust to gradient mini-batch size, if for any $b \leq r$, $\mathcal{S}$ guarantees to provide an ϵ-approximate solution $f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*) \leq \epsilon$ to (1) within $\mathcal{O}(\log(1/\epsilon))$ data passes, i.e., within $\mathcal{O}(n \log(1/\epsilon))$ component gradient evaluations.*

In other words, we say that a stochastic method is r-robust to the gradient mini-batch size if, for any $b \leq r$, the method reduces the excess function value $f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)$ after each data pass by an absolute constant factor less than 1. Figure 2 provides a visual depiction of the increased robustness of `Mb-SVRN` in comparison to first-order methods such as `SVRG` and `Katyusha`[1], as well as second-order methods like `SVRN`[2].

In order to attain improved robustness in `Mb-SVRN`, we utilize second-order information about the objective, which is accessed through the following approximate Hessian oracle.

Definition 3 (Hessian oracle) *Given fixed $0 < \beta_1 \leq \beta_2$, we say that $\mathcal{H} = \mathcal{H}_{\beta_1,\beta_2}$ is a β_2/β_1-approximate Hessian oracle, and use $\alpha := \beta_2/\beta_1$ to denote the Hessian approximation factor, if for any $\mathbf{x} \in \mathbb{R}^d$, $\mathcal{H}$ returns a possibly random $\bar{\mathbf{H}} \sim \mathcal{H}(\mathbf{x})$ such that:*

$$\beta_1 \cdot \nabla^2 f(\mathbf{x}) \preceq \bar{\mathbf{H}} \preceq \beta_2 \cdot \nabla^2 f(\mathbf{x}).$$

We treat the cost of Hessian approximation separate from the gradient data passes, as the complexity of Hessian estimation is largely problem and method-dependent. For instance, the above guarantee can be obtained with high probability by using a sub-sampled Hessian estimate of the form $\frac{1}{h} \sum_{j=1}^{h} \nabla^2 \psi_{i_j}(\mathbf{x})$, with appropriately chosen sample size h (which is what we use in all our numerical experiments). In particular, as guaranteed in Lemma 30, a uniformly drawn sample

1. Allen-Zhu (2017) claims mini-batch robustness of b up to $\mathcal{O}(\sqrt{n})$ for `Katyusha`, however in our regime of $n \gg \kappa$, it is easy to see that this can be further increased to $\mathcal{O}(n/\sqrt{\kappa})$.
2. We note that full-batch ($b = n$) second-order methods such as Subsampled Newton (Roosta-Khorasani and Mahoney, 2019) can converge ϵ-close after $\mathcal{O}(\log(1/\epsilon))$ only when using a sufficiently accurate Hessian estimate, which corresponds to $\alpha = \mathcal{O}(1)$.

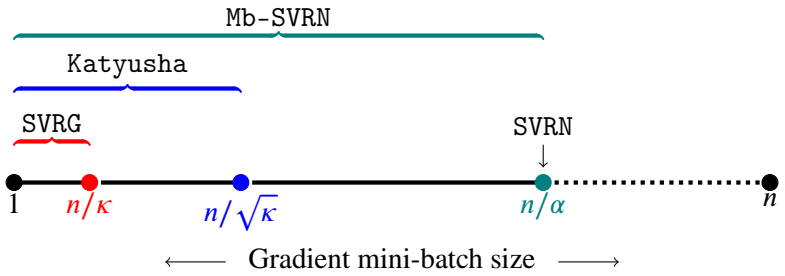


Figure 2: A visual illustration of mini-batch regimes under which different methods provide an ϵ-approximate solution to (1) within $\mathcal{O}(\log(1/\epsilon))$ data passes. We assume $n \gg \kappa$. While SVRG (Johnson and Zhang, 2013) and Katyusha (Allen-Zhu, 2017) can achieve this guarantee only until $b < n/\kappa$, and $b < n/\sqrt{\kappa}$ respectively, Mb-SVRN achieves the guarantee until $b \lesssim n/\alpha$, where α is the Hessian approximation factor (we assume $\kappa \gg \alpha^2$). We also show SVRN (Dereziński, 2025), which achieves the same guarantee only for $b \approx n/\alpha$.

$\{\nabla^2 \psi_{i_j}\}_{j=1}^h$ achieves the above Hessian approximation guarantee with probability $1-\delta$ for $\alpha = 1 + O(\kappa \log(d/\delta)/h + \sqrt{\kappa \log(d/\delta)/h})$, where $\beta_1 = 1/\sqrt{\alpha}$ and $\beta_2 = \sqrt{\alpha}$. In Mb-SVRN, we query the oracle once at the start of every outer iteration. So, in order to ensure that our Hessian oracle model is satisfied across the entire run of, say s outer iterations, each sub-sampled Hessian estimate should be an α-approximation with probability $1-\delta/s$ (leading to a $\log(sd/\delta)$ factor in the Hessian sample size h). Then, taking a union bound over the s outer iterations ensures that the Hessian oracle property is satisfied with probability $1-\delta$. Due to the above argument, from now on for the sake of simplicity we assume that the Hessian oracle returns an α-approximation with probability 1.

2.1 Main algorithm and result

In this section, we present an algorithm that follows the above computational model (Algorithm 1, Mb-SVRN), and provide our main technical result, the local convergence analysis of this algorithm across different values of the Hessian approximation factor α and gradient mini-batch sizes b. We note that the algorithm was deliberately chosen as a natural extension of both a standard stochastic variance-reduced first-order method (SVRG) and a standard Stochastic Newton-type method (SN), so that we can explore the effect of variance reduction in conjunction with second-order information on the convergence rate and robustness. Furthermore, Algorithm 1 is similar to SVRN (Dereziński, 2025), except that Mb-SVRN does not require large gradient mini-batch sizes and provides convergence guarantee for b as small as 1 and as large as $\mathcal{O}(\frac{n}{\alpha \log n})$. Moreover, due to potentially smaller mini-batches, the step size in Mb-SVRN is picked differently as compared to SVRN, which picks the step size as $O(1/\sqrt{\alpha})$ due to large b.

We now informally state our main result, which is a local convergence guarantee for Algorithm 1 (for completeness, we also provide a global convergence guarantee in Section 4.6). In this result, we show a high-probability convergence bound, that holds for any gradient mini-batch size b between 1 and $\mathcal{O}(\frac{n}{\alpha \log n})$, where the fast linear rate of convergence is independent of the mini-batch size.

Theorem 4 (Main result; informal Theorem 25) *Suppose that Assumption 1 holds and $n \gtrsim \alpha^2 \kappa$. Then, in a local neighborhood around $\mathbf{x}^*$, Algorithm 1 using α-approximate Hessian oracle and any*

Algorithm 1 Mini-batch Stochastic Variance-Reduced Newton (`Mb-SVRN`)

Require: $\tilde{\mathbf{x}}_0$, gradient mini-batch size b, gradient/Hessian oracles $\mathcal{G}, \mathcal{H}$, inner iterations $t_{\max}$
  for $s = 0, 1, 2, \cdots,$ **do**
    Compute the Hessian estimate $\bar{\mathbf{H}}_s \sim \mathcal{H}(\tilde{\mathbf{x}}_s)$
    Compute the full gradient $\mathbf{g}_s = \mathbf{g}(\tilde{\mathbf{x}}_s)$ for $\mathbf{g} \sim \mathcal{G}(\{1, ..., n\})$
    Set $\mathbf{x}_{0,s} = \tilde{\mathbf{x}}_s$
    for $t = 0, 1, 2, \cdots, t_{\max} - 1$ **do**
      Compute $\hat{\mathbf{g}} \sim \mathcal{G}(\{i_1, ..., i_b\})$ for $i_1, ..., i_b \sim \{1, ..., n\}$ uniformly random
      Compute $\hat{\mathbf{g}}_{t,s} = \hat{\mathbf{g}}(\mathbf{x}_{t,s})$ and $\hat{\mathbf{g}}_{0,s} = \hat{\mathbf{g}}(\tilde{\mathbf{x}}_s)$
      Compute variance-reduced gradient $\bar{\mathbf{g}}_{t,s} = \hat{\mathbf{g}}_{t,s} - \hat{\mathbf{g}}_{0,s} + \mathbf{g}_s$
      Update $\mathbf{x}_{t+1,s} = \mathbf{x}_{t,s} - \eta \bar{\mathbf{H}}_s^{-1} \bar{\mathbf{g}}_{t,s}$
    end for
    $\tilde{\mathbf{x}}_{s+1} = \mathbf{x}_{t_{max},s}$
  end for

gradient mini-batch size $b \lesssim \frac{n}{\alpha \log n}$*, with* $t_{\max} = n/b$ *and optimally chosen* η*, attains the following high-probability convergence bound:*

$$f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*) \leq \rho^s \cdot \left(f(\tilde{\mathbf{x}}_0) - f(\mathbf{x}^*) \right), \quad \textit{where} \quad \rho \lesssim \alpha^2 \cdot \frac{\kappa}{n}.$$

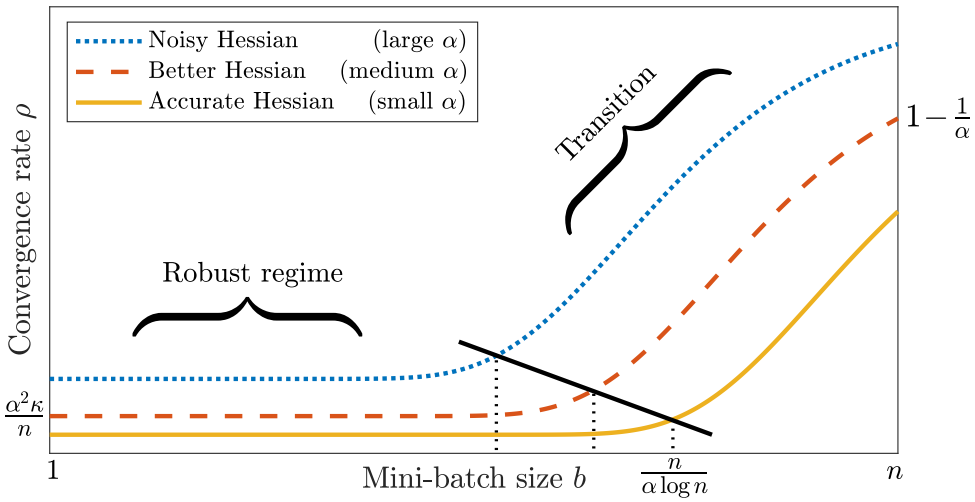
Remark 5 *Under our assumption that* $n \gtrsim \alpha^2 \kappa$*, the convergence rate satisfies* $\rho \ll 1/2$*, i.e., it is a fast condition-number-free linear rate of convergence that gets better for larger data sizes n. Since we used* $t_{\max} = n/b$*, one outer iteration of the algorithm corresponds to roughly two passes over the data (as measured by the gradient oracle calls).*

The proof of Theorem 4 (given in Section 4) requires a careful blend of the techniques from stochastic optimization with local convergence analysis of approximate Newton methods. We achieve this by developing a custom submartingale framework with random starting and stopping times that ensures the iterates stay in the local neighborhood, and using modified Freedman's inequality to convert in-expectation convergence of the inner iterates into a high-probability guarantee over the outer iterates.

2.2 Discussion

Our convergence analysis in Theorem 4 shows that incorporating second-order information into a stochastic variance-reduced gradient method makes it robust to increasing the gradient mini-batch size b up to the point where $b = \mathcal{O}\big(\frac{n}{\alpha \log n}\big)$. This is illustrated in Figure 3 (compare to the empirical Figure 1), where the robust regime corresponds to the convergence rate ρ staying flat as we vary b. Improving the Hessian oracle quality (smaller α) expands the robust regime, allowing for even larger mini-batch sizes with a flat convergence rate profile.

Note that, unless α is very close to 1 (extremely accurate Hessian oracle), the convergence rate per gradient data pass must eventually degrade, reaching $\rho \approx 1 - 1/\alpha$, which is the rate achieved by the corresponding Newton-type method with full gradients (e.g., see Lemma 8). The transition point of $b \approx \frac{n}{\alpha \log n}$ arises naturally in this setting, because the convergence rate after one outer iteration of `Mb-SVRN` could not be better than $\rho \approx (1 - 1/\alpha)^{t_{\max}}$ where $t_{\max} = n/b$ (obtained by treating the stochastic gradient noise as negligible).



plot/robustness.pdf

Figure 3: Illustration of the Mb-SVRN convergence analysis from Theorem 4 (for $n \gtrsim \alpha^2\kappa$), showing how the regime of robustness to gradient mini-batch size b depends on the quality of the Hessian oracle (smaller α means better Hessian estimate). As we increase b past $\frac{n}{\alpha \log n}$, the algorithm gradually transitions to a full-gradient Newton-type method.

Implications for stochastic gradient methods. In the case where the Hessian oracle always returns $\bar{\mathbf{H}}$ as $\mathbf{I}$, Mb-SVRN reduces to a first-order stochastic gradient method which is a variant of SVRG. So, it is natural to ask what our convergence analysis implies about the effect of second-order preconitioning on an SVRG-type algorithm. In our setting, SVRG with mini-batch size b achieves an expected convergence rate of $\rho = \mathcal{O}\left(\sqrt{\frac{b\kappa}{n}}\right)$, compared to a high-probability convergence rate of $\rho = \mathcal{O}\left(\frac{\kappa}{n}\right)$ for Mb-SVRN. This means that, while for small mini-batches $b = \mathcal{O}(n/\kappa)$ the advantage of preconditioning is not significant, the robust regime of Mb-SVRN (which is not present in SVRG) leads to a gap in convergence rates between the two methods that becomes larger as we increase b (see Figure 1). On the other hand, methods such as Catalyst (Lin et al., 2015) and Katyusha (Allen-Zhu, 2017, 2018), among others (Driggs et al., 2022), incorporate momentum acceleration techniques into SVRG and achieve improved convergence guarantees where the condition number dependence is improved from κ to $\sqrt{\kappa}$. However, similar to SVRG, all these are first-order methods and the convergence rate per data pass of these methods is still not nearly as robust to the gradient mini-batch size as Mb-SVRN. In particular, the convergence rate per data pass of Katyusha is comparable to that of Mb-SVRN up to mini-batch size $O(n/\sqrt{\kappa})$, but begins to deteriorate after that point (see Figure 2). This suggests the following general rule-of-thumb:

To improve the robustness of an SVRG-type method to the gradient mini-batch size, one can precondition the method with second-order information based on a Hessian estimate.

Implications for Newton-type methods. A different perspective on our results arises if our starting point is a full-gradient Newton-type method, such as Subsampled Newton (Roosta-Khorasani and Mahoney, 2019) or Newton Sketch (Pilanci and Wainwright, 2017), which also arises as a simple corner case of Mb-SVRN by choosing mini-batch size $b = n$ and inner iterations $t_{\max} = 1$. In

this setting, as mentioned above, the convergence rate of the method behaves as $\rho \approx 1 - 1/\alpha$, which means that it is highly dependent on the quality of the Hessian oracle via the Hessian approximation factor α. This sensitivity to the quality of Hessian approximation directly affects the scalability of these second-order methods, as generally it is more difficult to construct good Hessian approximations when n or κ become larger. In such scenarios, `Mb-SVRN` can provide a strategy to overcome this problem with scalability, as we explain below. As we decrease b in `Mb-SVRN`, the method gradually transitions into an `SVRG`-type method that is more robust to Hessian quality. We note that while our theoretical convergence result in Theorem 4 still exhibits a dependence on α, empirical results suggest that for large datasets this dependence is much less significant in the robust regime of `Mb-SVRN` than it is in the full-gradient regime. This suggests the following general prescription:

To improve the robustness of a Newton-type method to the Hessian estimation quality, one can replace the full gradients with variance-reduced sub-sampled gradient estimates.

Implications for parallel computing architectures. In recent times, there has been a considerable effort in developing parallelizable algorithms for problems arising in optimization and machine learning (Liu et al., 2014; Feng et al., 2024; Smith et al., 2018; Lin et al., 2025). In modern computing architectures, it is increasingly the case that computing the gradient of multiple component functions at a single iterate is much faster compared to computing the gradient of a single component function at multiple iterates due to advancements in parallel computing infrastructure and vectorized implementations. Under these developments, it is natural to look at the parallel complexity of an algorithm (Dereziński, 2025), which in case of problem (1) translates to the number of batch gradient queries. For instance, the optimal parallel complexity of `SVRG` is $O(\kappa \log(1/\epsilon))$, obtained by picking respective optimal gradient mini-batch size of $b = O(n/\kappa)$. Similarly, for methods like `Catalyst` and `Katyusha`, the parallel complexity is optimized at mini-batch size $b = O(n/\sqrt{\kappa})$, giving $O(\sqrt{\kappa}\log(1/\epsilon))$. However, for `Mb-SVRN`, the convergence rate per data pass is robust to b, and the parallel complexity is given as $O(\frac{n}{b}\log(1/\epsilon))$ for b up to roughly n/α. This means that as b increases beyond $O(n/\sqrt{\kappa})$, `Mb-SVRN` can leverage parallel computing hardware by offering the flexibility of choosing the mini-batch size b that best suits the communication-computation trade-off in a given architecture. This suggests the following general principle:

Preconditioning an SVRG-type method with a Hessian estimate can optimize its parallel complexity with respect to the gradient mini-batch size.

3. Related Work

Robustness improvement in optimization has been studied through various approaches across different problem settings, particularly with respect to gradient scaling, ill-conditioning, parameter sensitivity, and stochastic noise. For instance, popular diagonally scaled methods such as ADAGRAD (Duchi et al., 2011) and ADAM (Kingma and Ba, 2014) were introduced to make iterates more resilient to gradient magnitude disparities. Below, we outline the approaches most closely related to our work, namely those involving variance reduction, which ensures robustness to the stochastic gradient noise, and second-order methods, which address stability issues arising from ill-conditioning.

Among first-order variance-reduction methods, `SDCA` (Shalev-Shwartz and Zhang, 2013) and `SAG` (Schmidt et al., 2017) achieve convergence guarantees comparable to those of `SVRG`. Also,

Prox-SVRG (Xiao and Zhang, 2014) incorporates a weighted sampling strategy to improve the complexity so that it depends on the so-called average condition number, $\hat{\kappa}$, instead of κ. A similar weighted sampling strategy can be incorporated into Mb-SVRN with little effort. In addition to Catalyst and Katyusha, (Driggs et al., 2022) proposes a universal acceleration framework for first-order methods by including a simpler momentum and exploiting a bias-variance decomposition of the stochastic gradient. Also, loopless variants of first-order variance reduced methods using biased coin flips have been proposed by Hofmann et al. (2015); Kovalev et al. (2020).

Methods that incorporate stochastic second-order information have been proposed. In Gonen et al. (2016), using similar approaches to those in Xiao and Zhang (2014), a sketched preconditioned SVRG method is proposed, for solving ridge regression problems, that reduces the average condition number of the problem. The algorithms in Gower et al. (2016); Lucchi et al. (2015); Moritz et al. (2016), employ L-BFGS-type updates to approximate the Hessian inverse which is then used as a preconditioner to variance-reduced gradients. In Liu et al. (2019), the authors proposed an inexact preconditioned second-order method. All these methods use preconditioning to improve the dependence of SVRG-type methods on the condition number, and for practical considerations. However, their convergence analyses do not show scalability and robustness to gradient mini-batch sizes.

Popular stochastic second-order methods, e.g., Subsampled Newton (Roosta-Khorasani and Mahoney, 2019; Bollapragada et al., 2019; Erdogdu and Montanari, 2015) and Newton Sketch (Pilanci and Wainwright, 2017; Berahas et al., 2020), either use the full gradient or require extremely large gradient mini-batch sizes $b \gg \kappa$ at every iteration. Several other works have used strategies to reduce the variance of the stochastic second-order methods. One such work is Incremental Quasi-Newton (Mokhtari et al., 2018) which uses aggregated gradient and Hessian information coupled with solving a second-order Taylor approximation of f in a local neighborhood around $\mathbf{x}^*$ in order to reduce the variance of stochastic estimates employed. In Bollapragada et al. (2018) the authors propose methods that combine adaptive sampling strategies to reduce the variance and quasi-Newton updating to robustify the method. Finally, in Na et al. (2023) the authors show that averaging certain types of stochastic Hessian estimates (such as those based on subsampling) results in variance reduction and improved local convergence guarantees when used with full gradients.

The most closely related prior work is Dereziński (2025), where the authors investigated the effect of variance-reduced gradient estimates in conjunction with a Newton-type method, and introduced SVRN which is similar to Mb-SVRN as discussed earlier. They show that using gradient mini-batch size b of the order $\Theta(\frac{n}{\alpha \log n})$, one can achieve a local linear convergence rate of $\rho = \tilde{\mathcal{O}}\big(\frac{\alpha^3 \kappa}{n}\big)$. However, deviating from this prescribed mini-batch size (in either direction) causes the local convergence guarantee to rapidly degrade, and in particular, the results are entirely vacuous unless $b \gg \alpha^2 \kappa$. Our work can be viewed as a direct improvement over this prior work: first, through a better convergence rate dependence on α (with α^2 versus α^3); and second, in showing that this faster convergence can be achieved for any mini-batch size $b \lesssim \frac{n}{\alpha \log n}$. We achieve this by using a fundamentally different analysis via a chain of martingale concentration arguments combined with using much smaller step sizes, which allows us to compensate for the additional stochasticity in the gradients in the small and moderate mini-batch size regimes.

Another line of work has proposed combining variance-reduction with cubic regularized second-order methods (Wang et al., 2019; Zhang et al., 2022; Zhou et al., 2019). However, these methods primarily consider non-convex settings and at times use variance-reduction on Hessian estimates in addition to gradient estimates, rendering them largely incomparable to our problem setup.

4. Convergence Analysis

In this section, we present our technical analysis that concludes with the main convergence result, Theorem 25, informally stated earlier as Theorem 4. We start by introducing some auxiliary lemmas in Section 4.1. Then, in Theorem 11 (Section 4.2), we establish a one inner iteration convergence result in expectation, forming the key building block of our analysis. Then, in Section 4.3, we present our main technical contribution; we construct a submartingale framework with random start and stopping times, which is needed to establish fast high-probability linear convergence of outer iterates in Theorem 25, our main convergence result (Section 4.5). At the end of this section, we supplement our main local convergence results with a global convergence guarantee for `Mb-SVRN` in Theorem 26 (Section 4.6).

Notation. Let $\tilde{\mathbf{x}}_0$ denote the starting outer iterate and $\tilde{\mathbf{x}}_s$ denote the outer iterate after s outer iterations. Within the $(s+1)^{th}$ outer iteration, the inner iterates are indexed as $\mathbf{x}_{t,s}$. Since our results require working only within one outer iteration, we drop the subscript s throughout the analysis, and denote $\tilde{\mathbf{x}}_0$ as $\tilde{\mathbf{x}}$ and corresponding t^{th} inner iterate as $\mathbf{x}_t$. Similarly let $\tilde{\mathbf{H}}$ and $\mathbf{H}_t$ denote the Hessian at $\tilde{\mathbf{x}}$ and $\mathbf{x}_t$, respectively. For gradient mini-batch size b, let $\hat{\mathbf{g}}$ denote $\frac{1}{b}\sum_{j=1}^{b}\nabla\psi_{i_j}(\tilde{\mathbf{x}})$, let $\hat{\mathbf{g}}_t$ denote $\frac{1}{b}\sum_{j=1}^{b}\nabla\psi_{i_j}(\mathbf{x}_t)$ and let $\bar{\mathbf{g}}_t = \hat{\mathbf{g}}_t - \hat{\mathbf{g}} + \tilde{\mathbf{g}}$ denote the variance-reduced gradient at $\mathbf{x}_t$, where $\tilde{\mathbf{g}}$ represents exact gradient at $\tilde{\mathbf{x}}$. Moreover, $\mathbf{g}_t$ represents exact gradient at $\mathbf{x}_t$, $\mathbf{H}$ represents Hessian at $\mathbf{x}^*$, and let $\tilde{\Delta} = \tilde{\mathbf{x}} - \mathbf{x}^*$, $\Delta_t = \mathbf{x}_t - \mathbf{x}^*$. Also, we use η to denote the fixed step size, and $\alpha = \beta_2/\beta_1$ denotes the Hessian approximation factor. Finally, we refer to $\mathbb{E}_t$ as the conditional expectation given the event that the algorithm has reached iterate $\mathbf{x}_t$ starting with the outer iterate $\tilde{\mathbf{x}}$.

Next, we define the local convergence neighborhood, by using the notion of a Mahalanobis norm: $\|\mathbf{x}\|_{\mathbf{M}} = \sqrt{\mathbf{x}^\top\mathbf{M}\mathbf{x}}$, for any positive semi-definite matrix $\mathbf{M}$ and vector $\mathbf{x}$.

Definition 6 *Given $\epsilon_0 > 0$, we define a local neighborhood $\mathcal{U}_f(\epsilon_0)$ around $\mathbf{x}^*$ as follows:*

$$\mathcal{U}_f(\epsilon_0) := \left\{\mathbf{x} : \|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{H}}^2 < \frac{\mu^{3/2}}{L}\epsilon_0\right\},$$

where $L > 0$ is the Lipschitz constant for Hessians of f, and $\mu > 0$ is the strong convexity parameter for f.

4.1 Auxiliary lemmas

In this section we present auxiliary lemmas that will be used in the analysis. The first result upper bounds the variance of the `SVRG`-type stochastic gradient in terms of the smoothness parameter λ and gradient mini-batch size b. Similar variance bounds have also been derived in prior works (Johnson and Zhang, 2013; Dereziński, 2025; Berahas et al., 2023).

Lemma 7 (Upper bound on variance of stochastic gradient) *Let $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$ and $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$ for some $c \geq 1$. Then:*

$$\mathbb{E}_t[\|\bar{\mathbf{g}}_t - \mathbf{g}_t\|^2] \leq (1 + c\epsilon_0\eta)\frac{\lambda}{b}\|\tilde{\mathbf{x}} - \mathbf{x}_t\|_{\mathbf{H}}^2.$$

Proof Refer to Appendix A.1. ■

Assuming the access to Hessian oracle returning $\bar{\mathbf{H}}$ according to Definition 3, we define $\hat{\mathbf{H}} := \bar{\mathbf{H}}/\sqrt{\beta_1\beta_2}$. It immediately follows that,

$$\frac{1}{\sqrt{\alpha}}\nabla^2 f(\tilde{\mathbf{x}}) \preceq \hat{\mathbf{H}} \preceq \sqrt{\alpha}\nabla^2 f(\tilde{\mathbf{x}}),$$

for $\alpha = \beta_2/\beta_1$. We refer to the above property as α-Hessian approximation and say $\hat{\mathbf{H}}$ is an α-approximate Hessian at $\tilde{\mathbf{x}}$, denoted as $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}} \tilde{\mathbf{H}}$. Thus, without loss of generality, in the rest of the proof we will use $\hat{\mathbf{H}}$ instead of $\bar{\mathbf{H}}$. In the following result, we use standard approximate Newton analysis to establish that, using an α-approximate Hessian at $\tilde{\mathbf{x}}$, a Newton step with a sufficiently small step size η reduces the distance to $\mathbf{x}^*$ (in $\mathbf{H}$ norm), at least by a factor of $\left(1 - \frac{\eta}{8\sqrt{\alpha}}\right)$. We call this an approximate Newton step since the gradient is exact and only the second-order information is stochastic.

Lemma 8 (Guaranteed error reduction for approximate Newton) *Consider an approximate Newton step where* $\mathbf{x}_{\text{AN}} = \mathbf{x}_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t$, $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}} \tilde{\mathbf{H}}$, $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$, *and* $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$, *for some* $c \geq 1$, $\epsilon_0 < \frac{1}{8c\sqrt{\alpha}}$ *and* $\eta < \frac{1}{4\sqrt{\alpha}}$. *Then:*

$$\left\|\mathbf{x}_{\text{AN}} - \mathbf{x}^*\right\|_{\mathbf{H}} \leq \left(1 - \frac{\eta}{8\sqrt{\alpha}}\right)\left\|\mathbf{x}_t - \mathbf{x}^*\right\|_{\mathbf{H}},$$

where $\mathbf{H}$ *denotes the Hessian matrix at* $\mathbf{x}^*$.

Proof Refer to Appendix A.2. ■

Note that the approximate Newton step in Lemma 8 uses an exact gradient at every iteration, requiring the use of all n data samples. In our work, in addition to approximate Hessian, we incorporate variance-reduced gradients as well, calculated using b samples out of n. For a high probability analysis, we require an upper bound on the noise introduced due to the stochasticity in the gradients. The next result provides an upper bound on the noise (with high probability) in a single iteration introduced due to noise in the stochastic gradient.

Lemma 9 (High-probability bound on stochastic gradient noise) *Let* $\epsilon_0 < \frac{1}{8c\sqrt{\alpha}}$, $\eta < \frac{1}{4\sqrt{\alpha}}$, $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$, $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$ *for some* $c \geq 1$. *Then, for any* $\delta > 0$, *with probability at least* $1 - \delta\frac{b^2}{n^2}$ *(depending on the gradient mini-batch size b):*

$$\left\|\mathbf{g}_t - \bar{\mathbf{g}}_t\right\| \leq \begin{cases} (16\sqrt{\kappa\lambda}/3b)\ln(n/b\delta)\left\|\mathbf{x}_t - \tilde{\mathbf{x}}\right\|_{\mathbf{H}} & \text{for } b < \frac{8}{9}\kappa, \\ 4\sqrt{\lambda/b}\,\ln(n/b\delta)\left\|\mathbf{x}_t - \tilde{\mathbf{x}}\right\|_{\mathbf{H}} & \text{for } b \geq \frac{8}{9}\kappa. \end{cases}$$

Proof Refer to Appendix A.3. ■

The next result provides an upper bound on the gradient $\mathbf{g}_t = \mathbf{g}(\mathbf{x}_t)$ in terms of the distance of $\mathbf{x}_t$ to the minimizer $\mathbf{x}^*$. The result is useful in establishing a lower bound on $\left\|\Delta_{t+1}\right\|_{\mathbf{H}}$ in terms of $\left\|\Delta_t\right\|_{\mathbf{H}}$, essential to control the randomness in our submartingale framework, described in later sections.

Lemma 10 (Mean Value Theorem) *Let* $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$, $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$ *for* $\epsilon_0 < \frac{1}{8c\sqrt{\alpha}}$ *and for some* $c \geq 1$. *If the Hessian estimate satisfies* $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}} \tilde{\mathbf{H}}$, *then*

$$\left\|\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}} \leq 2\sqrt{\alpha}\left\|\Delta_t\right\|_{\mathbf{H}}.$$

Proof Refer to Appendix A.4. ■

4.2 One step expectation result

In the following theorem, we use the auxiliary lemmas (Lemma 7 and Lemma 8) to show that for a sufficiently small step size η, in expectation, `Mb-SVRN` is similar to the approximate Newton step (Lemma 8), with an additional small error term that can be controlled by the step size η. We note that while this result illustrates the convergence behavior of `Mb-SVRN`, it is by itself not sufficient to guarantee overall convergence (either in expectation or with high probability), because it does not ensure that the iterates will remain in the local neighborhood $\mathcal{U}_f$ throughout the algorithm (this is addressed with our submartingale framework, Section 4.3).

Theorem 11 (One inner iteration conditional expectation result) *Let* $\epsilon_0 < \frac{1}{8c\sqrt{\alpha}}$, $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$, $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$ *for some* $c \geq 1$. *Consider* `Mb-SVRN` *with step size* $0 < \eta < \min\{\frac{1}{4\sqrt{\alpha}}, \frac{b}{48\alpha^{3/2}\kappa}\}$ *and gradient mini-batch size b. Then:*

$$\mathbb{E}_t[\left\|\Delta_{t+1}\right\|_{\mathbf{H}}^2] \leq \left(1 - \frac{\eta}{8\sqrt{\alpha}}\right)\left\|\Delta_t\right\|_{\mathbf{H}}^2 + 3\eta^2\frac{\alpha\kappa}{b}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2.$$

Proof Taking the conditional expectation $\mathbb{E}_t$ of $\left\|\Delta_{t+1}\right\|_{\mathbf{H}}^2$,

$$\begin{aligned}\mathbb{E}_t[\left\|\Delta_{t+1}\right\|_{\mathbf{H}}^2] &= \mathbb{E}_t\left[\left\|\mathbf{x}_t - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t - \mathbf{x}^*\right\|_{\mathbf{H}}^2\right] \\ &\leq \mathbb{E}_t\left[\left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}}^2\right] + \eta^2\cdot\mathbb{E}_t\left[\left\|\hat{\mathbf{H}}^{-1}(\mathbf{g}_t - \bar{\mathbf{g}}_t)\right\|_{\mathbf{H}}^2\right].\end{aligned}$$

This is because $\mathbb{E}_t[\bar{\mathbf{g}}_t - \mathbf{g}_t] = 0$ and therefore the cross term vanishes. Since we know that $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$ and $\epsilon_0 < \frac{1}{8c\sqrt{\alpha}}$, by Lemma 8 we have $\left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}} \leq (1 - \frac{\eta}{8\sqrt{\alpha}})\left\|\Delta_t\right\|_{\mathbf{H}}$. Substituting this in the previous inequality, it follows that

$$\begin{aligned}\mathbb{E}_t[\left\|\Delta_{t+1}\right\|_{\mathbf{H}}^2] &\leq \left(1 - \frac{\eta}{8\sqrt{\alpha}}\right)^2\mathbb{E}_t\left[\left\|\Delta_t\right\|_{\mathbf{H}}^2\right] + \eta^2\cdot\mathbb{E}_t\left[\left\|\hat{\mathbf{H}}^{-1}(\mathbf{g}_t - \bar{\mathbf{g}}_t)\right\|_{\mathbf{H}}^2\right] \\ &\leq \left(1 - \frac{\eta}{8\sqrt{\alpha}}\right)^2\left\|\Delta_t\right\|_{\mathbf{H}}^2 + \eta^2\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1/2}\|^2\cdot\mathbb{E}_t\left[\|\hat{\mathbf{H}}^{-1/2}(\mathbf{g}_t - \bar{\mathbf{g}}_t)\|^2\right].\end{aligned}$$

Using that $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}} \tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}} \approx_{(1+\epsilon_0\eta)} \mathbf{H}$, we have $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}(1+\epsilon_0\eta)} \mathbf{H}$. Therefore, $\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1/2}\|^2 = \left\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\mathbf{H}^{1/2}\right\| \leq \sqrt{\alpha}(1+\epsilon_0\eta)$. So we get,

$$\mathbb{E}_t[\|\Delta_{t+1}\|_{\mathbf{H}}^2] \leq \left(1 - \frac{\eta}{8\sqrt{\alpha}}\right)^2 \|\Delta_t\|_{\mathbf{H}}^2 + \eta^2\sqrt{\alpha}(1+\epsilon_0\eta) \cdot \mathbb{E}_t\left[\|\hat{\mathbf{H}}^{-1/2}(\mathbf{g}_t - \bar{\mathbf{g}}_t)\|^2\right].$$

Upper bounding $\|\hat{\mathbf{H}}^{-1/2}\|^2 = \left\|\hat{\mathbf{H}}^{-1}\right\| \leq \frac{\sqrt{\alpha}}{\mu}$,

$$\mathbb{E}_t[\|\Delta_{t+1}\|_{\mathbf{H}}^2] \leq \left(1 - \frac{\eta}{8\sqrt{\alpha}}\right)^2 \|\Delta_t\|_{\mathbf{H}}^2 + \eta^2\frac{\alpha}{\mu}(1+\epsilon_0\eta) \cdot \mathbb{E}_t[\|(\mathbf{g}_t - \bar{\mathbf{g}}_t)\|^2].$$

By Lemma 7, we bound the last term of the previous inequality,

$$\begin{aligned}\mathbb{E}_t\left[\|\mathbf{g}_t - \bar{\mathbf{g}}_t\|^2\right] &\leq \frac{(1+c\epsilon_0\eta)\lambda}{b}\|\Delta_t - \tilde{\Delta}\|_{\mathbf{H}}^2 \\ &\leq \frac{2(1+c\epsilon_0\eta)\lambda}{b}\left(\|\Delta_t\|_{\mathbf{H}}^2 + \|\tilde{\Delta}\|_{\mathbf{H}}^2\right).\end{aligned}$$

Putting it all together,

$$\begin{aligned}\mathbb{E}_t[\|\Delta_{t+1}\|_{\mathbf{H}}^2] &\leq \left[\left(1 - \frac{\eta}{8\sqrt{\alpha}}\right)^2 + 3\eta^2\alpha\frac{\kappa}{b}\right]\|\Delta_t\|_{\mathbf{H}}^2 + 3\eta^2\frac{\alpha\kappa}{b} \cdot \|\tilde{\Delta}\|_{\mathbf{H}}^2 \\ &= \left[1 - \frac{\eta}{4\sqrt{\alpha}} + \frac{\eta^2}{64\alpha} + 3\eta^2\frac{\alpha\kappa}{b}\right]\|\Delta_t\|_{\mathbf{H}}^2 + 3\eta^2\frac{\alpha\kappa}{b} \cdot \|\tilde{\Delta}\|_{\mathbf{H}}^2 \\ &< \left(1 - \frac{\eta}{8\sqrt{\alpha}}\right)\|\Delta_t\|_{\mathbf{H}}^2 + 3\eta^2\frac{\alpha\kappa}{b} \cdot \|\tilde{\Delta}\|_{\mathbf{H}}^2, \end{aligned} \tag{2}$$

where we used that $(1+\epsilon_0\eta) < 1 + \frac{1}{8}$ and $\eta < \frac{b}{48\alpha^{3/2}\kappa}$ imply $3\eta^2\frac{\alpha\kappa}{b} < \frac{\eta}{16\sqrt{\alpha}}$. ■

In the remainder of the analysis, we set $c = e^2$, where $e \approx 2.718$ represents Euler's number, and consider $\eta = \frac{b\sqrt{\alpha}\beta}{n}$ for some $\beta > 1$. These specific values for c and η are set in hindsight based on the optimal step size and the neighborhood scaling factor (c) derived later as artifacts of our analysis. On the other hand, one can do the analysis with general c and η and later derive these assignments for η and c. The value for β also gets specified as the analysis proceeds.

4.3 Building blocks for the submartingale framework

We begin by building an intuitive understanding of our submartingale framework, where we define a random process Y_t to describe the convergence behavior of the algorithm. Informally, one can think of Y_t as representing the error $\|\Delta_t\|_{\mathbf{H}}^2$, and aiming to establish the submartingale property $\mathbb{E}_t[Y_{t+1}] < Y_t$. We next highlight two problematic scenarios with achieving the submartingale property, which are illustrated in Figure 4.

First, the assumption $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$ in Theorem 11 ensures that the approximate Newton step reduces the error at least by a factor of $(1 - \frac{\eta}{8\sqrt{\alpha}})$. However, if $\mathbf{x}_t \notin \mathcal{U}_f(c\epsilon_0\eta)$, this claim is no longer valid, leading to the first scenario that disrupts the submartingale behavior. This implies the necessity of the property $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$, for some constant c, to establish $\mathbb{E}_t[Y_{t+1}] < Y_t$. In the course of our martingale concentration analysis, we show the existence of an absolute constant $c > 0$, such that $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$, for all t, with high probability.

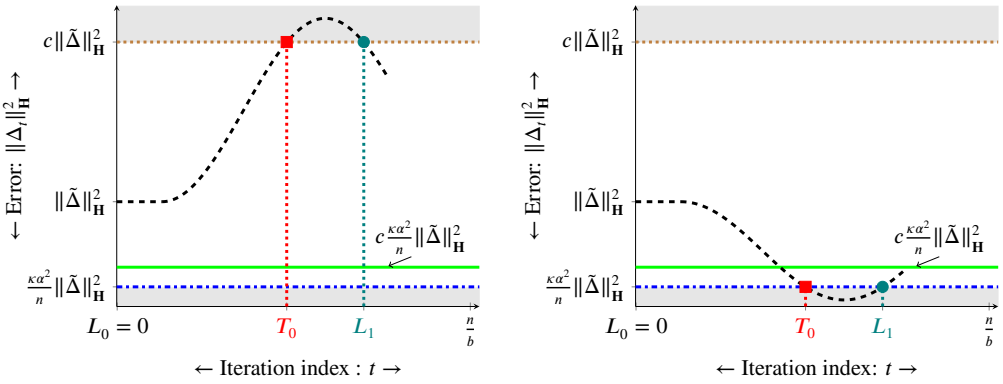
We now explain the second scenario that disrupts the submartingale property. Note that Y_{t+1} involves noise arising due to the stochasticity of the variance-reduced gradient $\bar{\mathbf{g}}_t$. This stochasticity leads to the term $3\eta^2\frac{\alpha\kappa}{b}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ in the statement of Theorem 11. Now, in the event that this noise dominates the error reduction due to the approximate Newton step, one cannot establish that $\mathbb{E}_t[Y_{t+1}] \leq Y_t$. However, we show later in Lemma 14 that this problematic scenario occurs only when $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \lesssim \frac{\kappa\alpha^2}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$, i.e., when the error becomes small enough to recover our main convergence guarantee. As the algorithm cannot automatically detect this scenario and restart a new outer iteration, it still performs the remaining inner iterations. This results in subsequent inner iterates being affected by high stochastic gradient noise, disrupting the submartingale property.

To address these scenarios, we introduce random stopping times T_i and random resume times L_i, where stopping times capture the disruptive property scenarios, and resume times capture property restoration after encountering stopping times. Refer to Figure 4 for a visual representation, where the dashed curve denotes the error $\left\|\Delta_t\right\|_{\mathbf{H}}^2$, the x-axis denotes the iteration index t, ■ denotes the stopping time, and • denotes the resume time. The submartingale property holds as long as $\left\|\Delta_t\right\|_{\mathbf{H}}^2$ stays between the brown (dotted) and blue (dashdotted) lines (i.e., in the white strip between the grey). The brown (dashed) line at the top denotes the scenario where $\mathbf{x}_t \notin \mathcal{U}_f(c\epsilon_0\eta)$, and the blue (dashdotted) line at the bottom denotes the scenario where $\left\|\Delta_t\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$. The green (solid) line denotes the convergence guarantee provided in our work, which remarkably is a constant multiple of the blue (dashdotted) line. Our martingale concentration argument shows that after n/b iterations, the error will, with high probability, fall below the green (solid) line.

We proceed to formally define the random stopping and resume times. Consider a fixed $\gamma > 0$. Define random stopping times T_i and random resume times L_i, with $L_0 = 0$ and for $i \geq 0$, as

$$
\begin{aligned}
T_i &= \min\left(\left\{t > L_i \mid \left\|\Delta_t\right\|_{\mathbf{H}}^2 > e^2\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \text{ or } \left\|\Delta_t\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2\right\} \cup \left\{\frac{n}{b}\right\}\right), \\
L_{i+1} &= \min\left(\left\{t > T_i \mid \left\|\Delta_t\right\|_{\mathbf{H}}^2 \leq e^2\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \text{ and } \left\|\Delta_t\right\|_{\mathbf{H}}^2 \geq \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2\right\} \cup \left\{\frac{n}{b}\right\}\right). \qquad (3)
\end{aligned}
$$

The random stopping time T_i denotes the $(i+1)^{th}$ iteration index when the random sequence $(\mathbf{x}_t)$ leaves the local neighborhood $\mathcal{U}_f(e^2\epsilon_0\eta)$ or satisfies $\left\|\Delta_t\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$. The random resume time L_i denotes the $(i+1)^{th}$ iteration index when the random sequence $(\mathbf{x}_t)$ returns back to the local neighborhood and also satisfies $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \geq \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$. The condition $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \leq e^2\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$, ensures that $\mathbf{x}_t$ lies in a small local neighborhood around $\mathbf{x}^*$ such that the previous results (Lemma 7, 8, 9) hold with $c = e^2$, whereas the condition $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \geq \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ intuitively means that $\mathbf{x}_t$ hasn't reached too close to $\mathbf{x}^*$(necessarily far from $\tilde{\mathbf{x}}$) such that behaviour of $\mathbb{E}\left\|\Delta_{t+1}\right\|_{\mathbf{H}}$ is determined by noise in the stochastic gradient. The latter condition also captures the fact that the optimal convergence rate has not been achieved at $\mathbf{x}_t$. These two different conditions are required to design a submartingale from random time L_i to T_i. Note that, in hindsight, we have defined stopping time in a way similar



(a) Scenario 1: Leaving the neighborhood

(b) Scenario 2: Too close to the optimum

Figure 4: Visual illustration of two problematic scenarios disrupting the submartingale behavior of $\|\Delta_t\|_{\mathbf{H}}^2$. The green (solid) line denotes the convergence guarantee we prove in our work (Theorem 25). In the left plot, the red square (■) represents the first stopping time obtained due to failure of the local neighborhood condition, plotted via the brown (dotted) line near the top. In the right plot, the red square (■) represents the first stopping time obtained due the iterate $\mathbf{x}_t$ lying too close to $\mathbf{x}^*$, plotted via the blue (dashdotted) line near the bottom. The teal colored dots (•) denote the resume time representing the restoration of submartingale property.

to our convergence guarantee, $\left\|\Delta_{\frac{n}{b}}\right\|_{\mathbf{H}}^2 \lesssim \frac{\kappa\alpha^2}{n}\|\tilde{\Delta}\|_{\mathbf{H}}^2$. We add a factor γ to capture constants and log factors that can appear in the convergence guarantee. One can show that for $T_i \leq t < L_{i+1}$, the behavior of $\|\Delta_t\|_{\mathbf{H}}^2$ is dictated by the variance in the stochastic gradient and not the progress made by the approximate Newton step.

Remark 12 *The stopping and resume times satisfy the following properties:*

(a) For $L_0 \leq t < T_0$, we have $\|\Delta_t\|_{\mathbf{H}}^2 \leq e^2\|\tilde{\Delta}\|_{\mathbf{H}}^2$. This is due to the definitions of L_0 and T_0 in (3). Equivalently, we can write,

$$\Pr\left(\|\Delta_t\|_{\mathbf{H}}^2 \leq e^2\|\tilde{\Delta}\|_{\mathbf{H}}^2 \mid L_0 \leq t < T_0\right) = 1. \tag{4}$$

(b) For $T_i \leq t < L_{i+1}$ and given that $\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\|\tilde{\Delta}\|_{\mathbf{H}}^2$, we have $\|\Delta_t\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\|\tilde{\Delta}\|_{\mathbf{H}}^2$. This is because L_{i+1} is the first instance after T_i such that $\frac{\kappa\alpha^2\gamma}{n}\|\tilde{\Delta}\|_{\mathbf{H}}^2 \leq \left\|\Delta_{L_i}\right\|_{\mathbf{H}}^2 \leq e^2\|\tilde{\Delta}\|_{\mathbf{H}}^2$. Equivalently, we can write,

$$\Pr\left(\|\Delta_t\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\|\tilde{\Delta}\|_{\mathbf{H}}^2 \mid T_i \leq t < L_{i+1}, \left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\|\tilde{\Delta}\|_{\mathbf{H}}^2\right) = 1. \tag{5}$$

(c) For any gradient mini-batch size $1 \leq b \leq n$, `Mb-SVRN` performs n/b inner iterations, and therefore by construction random times T_i and L_i would not be realized more than n/b times.

Note that due to (5), Mb-SVRN provides a very strong guarantee on the error of the iterates $\mathbf{x}_t$ when $T_i \leq t < L_{i+1}$ and $\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$. However, the stopping time T_i can be realized due to $\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 > e^2\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ or $T_i = n/b$, and in that case the conditioning event in (5) may never hold.

Now we proceed to construct a submartingale framework from the random resume time L_i to the random stopping time T_i. Our aim is to apply Freedman's inequality on a carefully constructed martingale and prove strong concentration guarantees on $\left\|\Delta_{T_i}\right\|_{\mathbf{H}}$, resulting in our main result. We state a version of Freedman's inequality for submartingales, which is a minor modification of standard Freedman for martingales (Tropp, 2011), see Appendix A.5.

Theorem 13 (Freedman's inequality for submartingales) *For a random process Y_t satisfying $\mathbb{E}_t\left[Y_{t+1}\right] \leq Y_t$ and $|Y_{t+1} - Y_t| \leq R$, for any $\zeta, \sigma > 0$ it follows that:*

$$\Pr\left(\exists t \mid Y_t > Y_0 + \zeta \text{ and } \sum_{j=0}^{t-1}\mathbb{E}_j\left(Y_{j+1} - Y_j\right)^2 \leq \sigma^2\right) \leq \exp\left(-\frac{1}{4}\min\left\{\frac{\zeta^2}{\sigma^2}, \frac{\zeta}{R}\right\}\right).$$

As required for the application of Freedman's inequality on any random process, we need to establish the following three properties:

(a) the submartingale property, i.e., $\mathbb{E}_t[Y_{t+1}] \leq Y_t$;

(b) the predictable quadratic variation bound, i.e., a bound on $\mathbb{E}_t\left(Y_{t+1} - Y_t\right)^2$;

(c) and the almost sure upper bound, i.e., a bound on $|Y_{t+1} - Y_t|$.

Now, having defined stopping and resume times, we provide a few lemmas that provide fundamental results for analyzing the formal submartingale framework developed in Section 4.4. However, we note that Lemmas 14, 16, 17, 18 are interesting on their own and can be understood outside of the submartingale viewpoint. The next lemma shows that if $\mathbf{x}_t$ lies in the local neighborhood $\mathcal{U}_f(e^2\epsilon_0\eta)$, and does not satisfy $\left\|\Delta_t\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$, then we have a submartingale property ensuring that in expectation $\left\|\Delta_{t+1}\right\|_{\mathbf{H}}$ is smaller than $\left\|\Delta_t\right\|_{\mathbf{H}}$.

Lemma 14 (Submartingale property till stopping time) *Let the gradient mini-batch size be $1 \leq b \leq n$ and $\epsilon_0 < \frac{1}{8e^2\sqrt{\alpha}}$. Let $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$, $\mathbf{x}_t \in \mathcal{U}_f(e^2\epsilon_0\eta)$, $\eta = \frac{b\sqrt{\alpha\beta}}{n} < \min\{\frac{1}{4\sqrt{\alpha}}, \frac{b}{48\alpha^{3/2}\kappa}\}$, $\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \leq \frac{n}{\kappa\alpha^2\gamma}\left\|\Delta_t\right\|_{\mathbf{H}}^2$, and $\frac{3\beta}{\gamma} < \frac{1}{16}$. Then:*

$$\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}} \leq \left(1 - \frac{\eta}{32\sqrt{\alpha}}\right)\left\|\Delta_t\right\|_{\mathbf{H}}.$$

Proof By Theorem 11 and using $\frac{3\beta}{\gamma} < \frac{1}{16}$, it follows that,

$$\begin{aligned}
\mathbb{E}_t \left\|\Delta_{t+1}\right\|_{\mathbf{H}}^2 &\leq \left(1 - \frac{\eta}{8\sqrt{\alpha}}\right) \left\|\Delta_t\right\|_{\mathbf{H}}^2 + 3\eta^2 \frac{\alpha\kappa}{b} \left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \\
&\leq \left(1 - \frac{\eta}{8\sqrt{\alpha}}\right) \left\|\Delta_t\right\|_{\mathbf{H}}^2 + 3\eta \cdot \frac{b\sqrt{\alpha}\beta}{n} \cdot \frac{\alpha\kappa}{b} \cdot \frac{n}{\kappa\alpha^2\gamma} \left\|\Delta_t\right\|_{\mathbf{H}}^2 \\
&= \left(1 - \frac{\eta}{8\sqrt{\alpha}} + \frac{3\eta}{\sqrt{\alpha}} \cdot \frac{\beta}{\gamma}\right) \left\|\Delta_t\right\|_{\mathbf{H}}^2 \\
&\leq \left(1 - \frac{\eta}{16\sqrt{\alpha}}\right) \left\|\Delta_t\right\|_{\mathbf{H}}^2,
\end{aligned}$$

implying that

$$\mathbb{E}_t \left\|\Delta_{t+1}\right\|_{\mathbf{H}} \leq \left(1 - \frac{\eta}{32\sqrt{\alpha}}\right) \left\|\Delta_t\right\|_{\mathbf{H}},$$

which concludes the proof. ■

Remark 15 *We make a few remarks about the step size $\eta = \frac{b\sqrt{\alpha}\beta}{n}$ which comes out of our analysis as the optimal choice. Lemma 14 requires that $\eta < \min\left\{\frac{1}{4\sqrt{\alpha}}, \frac{b}{48\alpha^{3/2}\kappa}\right\}$, where the first term in the bound ensures convergence of the approximate Newton step, whereas the second term guarantees control over the gradient noise. In the regime where $n \gg \alpha^2\kappa\beta$ (which we assume in our main result), it is always true that our chosen step size satisfies $\eta < \frac{b}{48\alpha^{3/2}\kappa}$. However, to ensure that $\eta < \frac{1}{4\sqrt{\alpha}}$ we must restrict the mini-batch size to $b < \frac{n}{4\alpha\beta}$. In our main result, we use $\beta = \mathcal{O}(\log(n/\alpha^2\kappa))$. This suggests that, in the regime of $b \gtrsim \frac{n}{\alpha\log(n)}$, the choice of the step size is primarily restricted by the Hessian approximation factor α. Ultimately, this leads to deterioration of the convergence rate as b increases beyond $\frac{n}{\alpha\log(n)}$.*

Throughout our analysis, we assume $\frac{b\sqrt{\alpha}\beta}{n} < \frac{1}{4\sqrt{\alpha}}$, which corresponds to the condition $b \lesssim \frac{n}{\alpha\log n}$ stated informally in Theorem 4. This is the regime where the convergence rate of `Mb-SVRN` does not deteriorate with b. Having established the submartingale property, we now prove the predictable quadratic variation bound property. For proving a strong upper bound on the quadratic variation, we need a result upper bounding $\left\|\tilde{\Delta}\right\|_{\mathbf{H}}$ by $\mathbb{E}_t \left\|\Delta_{t+1}\right\|_{\mathbf{H}}$, as long as the stopping time criteria are not satisfied at $\mathbf{x}_t$.

Lemma 16 (Bounded variation) *Let $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$ with $\epsilon_0 < \frac{1}{8e^2\sqrt{\alpha}}$ and $\eta \leq \frac{1}{4\sqrt{\alpha}}$. Also, let $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \leq e^2 \left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ and $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \geq \frac{\kappa\alpha^2\gamma}{n} \left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ for some $\gamma > 0$. Then, with $\omega = \sqrt{\frac{n}{\kappa\alpha^2\gamma}}$:*

$$\left\|\tilde{\Delta}\right\|_{\mathbf{H}} \leq 2\omega\mathbb{E}_t \left\|\Delta_{t+1}\right\|_{\mathbf{H}}.$$

Proof We apply Jensen's inequality to $\mathbb{E}_t \left\|\Delta_{t+1}\right\|_{\mathbf{H}}$ to get,

$$\begin{aligned}\mathbb{E}_t \left\|\Delta_{t+1}\right\|_{\mathbf{H}} &\geq \left\|\mathbb{E}_t \Delta_{t+1}\right\|_{\mathbf{H}} = \left\|\Delta_t - \eta \hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}} \\ &\geq \left\|\Delta_t\right\|_{\mathbf{H}} - \eta \left\|\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}}.\end{aligned}$$

Since $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \leq e^2 \left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$, we have $\mathbf{x}_t \in \mathcal{U}_f(e^2\epsilon_0\eta)$. By using Lemma 10 on the term $\left\|\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}}$,

$$\mathbb{E}_t \left\|\Delta_{t+1}\right\|_{\mathbf{H}} \geq \left\|\Delta_t\right\|_{\mathbf{H}} - 2\eta\sqrt{\alpha}\left\|\Delta_t\right\|_{\mathbf{H}} \geq \frac{1}{2}\left\|\Delta_t\right\|_{\mathbf{H}}.$$

The last inequality holds because $\eta \leq \frac{1}{4\sqrt{\alpha}}$. Finally, we use the condition that $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \geq \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$, concluding the proof. ■

We proceed with the predictable quadratic variation bound for $\mathbb{E}_t \left\|\Delta_{t+1}\right\|_{\mathbf{H}}$, assuming that $\mathbf{x}_t$ does not satisfy the stopping time criteria.

Lemma 17 (Predictable quadratic variation bound) *Let $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$ with $\epsilon_0 < \frac{1}{8e^2\sqrt{\alpha}}$ and $\eta = \frac{b\sqrt{\alpha}\beta}{n} \leq \min\left\{\frac{1}{4\sqrt{\alpha}}, \frac{b}{48\alpha^{3/2}\kappa}\right\}$ for some $\beta > 0$. Also, let $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \leq e^2\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ and $\left\|\Delta_t\right\|_{\mathbf{H}}^2 \geq \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ for some $\gamma > 0$. Then, with $\omega = \sqrt{\frac{n}{\kappa\alpha^2\gamma}}$:*

$$\mathbb{E}_t\left(\left\|\Delta_{t+1}\right\|_{\mathbf{H}} - \mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}\right)^2 < 80\eta\frac{\beta}{\gamma\sqrt{\alpha}}\left(\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}\right)^2.$$

Proof Consider $\mathbb{E}_t\left(\left\|\Delta_{t+1}\right\|_{\mathbf{H}} - \mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}\right)^2$. We have,

$$\begin{aligned}\mathbb{E}_t\left(\left\|\Delta_{t+1}\right\|_{\mathbf{H}} - \mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}\right)^2 &= \mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}^2 - (\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}})^2 \\ &= \mathbb{E}_t\left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t\right\|_{\mathbf{H}}^2 - (\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}})^2 \\ &= \left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}}^2 + \eta^2\mathbb{E}_t\left\|\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|_{\mathbf{H}}^2 \\ &\quad + 2\eta\mathbb{E}_t\left(\Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right)^\top\left(\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right) - (\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}})^2.\end{aligned}$$

Using $\mathbb{E}_t\left(\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right) = \hat{\mathbf{H}}^{-1}\mathbb{E}_t\left(\bar{\mathbf{g}}_t - \mathbf{g}_t\right) = 0$ and $\Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t = \mathbb{E}_t\Delta_{t+1}$, we get,

$$\mathbb{E}_t\left(\left\|\Delta_{t+1}\right\|_{\mathbf{H}} - \mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}\right)^2 = \left\|\mathbb{E}_t\Delta_{t+1}\right\|_{\mathbf{H}}^2 + \eta^2\mathbb{E}_t\left\|\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|_{\mathbf{H}}^2 - (\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}})^2.$$

Using Jensen's inequality on the first term, we have $\left\|\mathbb{E}_t\Delta_{t+1}\right\|_{\mathbf{H}} \leq \mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}$, from which it follows that,

$$\begin{aligned}\mathbb{E}_t\left(\left\|\Delta_{t+1}\right\|_{\mathbf{H}} - \mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}\right)^2 &\leq (\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}})^2 + \eta^2\mathbb{E}_t\left\|\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|_{\mathbf{H}}^2 - (\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}})^2 \\ &= \eta^2\mathbb{E}_t\left\|\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|_{\mathbf{H}}^2 \\ &= \eta^2\mathbb{E}_t\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1/2}\hat{\mathbf{H}}^{-1/2}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\|^2 \\ &\leq \eta^2 \cdot \|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1/2}\|^2 \cdot \mathbb{E}_t\|\hat{\mathbf{H}}^{-1/2}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\|^2. \qquad (6)\end{aligned}$$

Using the fact that $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}} \tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}} \approx_{(1+\epsilon_0\eta)} \mathbf{H}$, we have $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}(1+\epsilon_0\eta)} \mathbf{H}$. Therefore, $\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1/2}\|^2 = \left\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\mathbf{H}^{1/2}\right\| \leq \sqrt{\alpha}(1+\epsilon_0\eta)$. Combining this with the inequality (6),

$$\mathbb{E}_t\left(\|\Delta_{t+1}\|_{\mathbf{H}} - \mathbb{E}_t\|\Delta_{t+1}\|_{\mathbf{H}}\right)^2 \leq \eta^2\sqrt{\alpha}(1+\epsilon_0\eta)\cdot\mathbb{E}_t\|\hat{\mathbf{H}}^{-1/2}(\mathbf{g}_t - \bar{\mathbf{g}}_t)\|^2.$$

Upper bounding $\|\hat{\mathbf{H}}^{-1/2}\|^2 = \left\|\hat{\mathbf{H}}^{-1}\right\| \leq \frac{\sqrt{\alpha}}{\mu}$, we get,

$$\mathbb{E}_t\left(\|\Delta_{t+1}\|_{\mathbf{H}} - \mathbb{E}_t\|\Delta_{t+1}\|_{\mathbf{H}}\right)^2 \leq \eta^2\frac{\alpha}{\mu}(1+\epsilon_0\eta)\cdot\mathbb{E}_t[\|(\mathbf{g}_t - \bar{\mathbf{g}}_t)\|^2]. \tag{7}$$

Since $\|\Delta_t\|_{\mathbf{H}}^2 < e^2\|\tilde{\Delta}\|_{\mathbf{H}}^2$, we have $\mathbf{x}_t \in \mathcal{U}_f(e^2\epsilon_0\eta)$. By using Lemma 7 on the last term of the inequality (7), it follows that,

$$\mathbb{E}_t\left[\|\mathbf{g}_t - \bar{\mathbf{g}}_t\|^2\right] \leq \frac{(1+c\epsilon_0\eta)\lambda}{b}\|\Delta_t - \tilde{\Delta}\|_{\mathbf{H}}^2 \leq \frac{2(1+c\epsilon_0\eta)\lambda}{b}\left(\|\Delta_t\|_{\mathbf{H}}^2 + \|\tilde{\Delta}\|_{\mathbf{H}}^2\right). \tag{8}$$

Substituting (8) in the inequality (7), we get,

$$\mathbb{E}_t\left(\|\Delta_{t+1}\|_{\mathbf{H}} - \mathbb{E}_t\|\Delta_{t+1}\|_{\mathbf{H}}\right)^2 \leq 2(1+\epsilon_0\eta)(1+c\epsilon_0\eta)\eta^2\frac{\kappa\alpha}{b}\left(\|\Delta_t\|_{\mathbf{H}}^2 + \|\tilde{\Delta}\|_{\mathbf{H}}^2\right).$$

Again, using $\|\Delta_t\|_{\mathbf{H}}^2 < e^2\|\tilde{\Delta}\|_{\mathbf{H}}^2$,

$$\mathbb{E}_t\left(\|\Delta_{t+1}\|_{\mathbf{H}} - \mathbb{E}_t\|\Delta_{t+1}\|_{\mathbf{H}}\right)^2 < 20\eta^2\frac{\kappa\alpha}{b}\|\tilde{\Delta}\|_{\mathbf{H}}^2.$$

Since $\eta \leq \frac{1}{4\sqrt{\alpha}}$, we use Lemma 16 on $\|\tilde{\Delta}\|_{\mathbf{H}}^2$ to obtain,

$$\mathbb{E}_t\left(\|\Delta_{t+1}\|_{\mathbf{H}} - \mathbb{E}_t\|\Delta_{t+1}\|_{\mathbf{H}}\right)^2 < 80\eta^2\frac{\kappa\alpha}{b}\omega^2\left(\mathbb{E}_t\|\Delta_{t+1}\|_{\mathbf{H}}\right)^2.$$

Substituting one of the η factors as $\frac{b\sqrt{\alpha}\beta}{n}$, and $\omega^2 = \frac{n}{\kappa\alpha^2\gamma}$, concludes the proof. ■

Finally, as the last building block of the submartingale framework, we establish a high probability upper and lower bound on $\|\Delta_{t+1}\|_{\mathbf{H}}$ in terms of $\|\Delta_t\|_{\mathbf{H}}$. Here, due to the noise in the stochastic gradient, we cannot prove almost sure bounds. However, by Lemmas 9 and 16, we can get the upper and lower bounds, holding with probability at least $1-\delta\frac{b^2}{n^2}$, for any $\delta > 0$. For notational convenience, we use M to denote the high probability upper bound on $\|\bar{\mathbf{g}}_t - \mathbf{g}_t\|$, guaranteed in Lemma 9,

$$M = \begin{cases} \frac{\kappa}{b} & \text{if } b < \frac{8}{9}\kappa \\ \sqrt{\frac{\kappa}{b}} & \text{if } b \geq \frac{8}{9}\kappa. \end{cases}$$

Lemma 18 (One step high probability upper bound) *Let* $n > \frac{(96)^2\beta^2\ln(n/b\delta)^2}{\gamma}\kappa$, $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$ *with* $\epsilon_0 < \frac{1}{8e^2\sqrt{\alpha}}$, $\eta = \frac{b\sqrt{\alpha}\beta}{n} \leq \min\left\{\frac{1}{4\sqrt{\alpha}}, \frac{b}{48\alpha^{3/2}\kappa}\right\}$ *and* $b < \min\{\frac{n}{4\alpha\beta}, \frac{\gamma n}{(96)^2\beta^2\ln(n/b\delta)^2}\}$, *for some* $\beta > 0$, $\gamma > 0$, *and any* $\delta > 0$. *Also, let* $\|\Delta_t\|_{\mathbf{H}}^2 \leq e^2\|\tilde{\Delta}\|_{\mathbf{H}}^2$ *and* $\|\Delta_t\|_{\mathbf{H}}^2 \geq \frac{\kappa\alpha^2\gamma}{n}\|\tilde{\Delta}\|_{\mathbf{H}}^2$. *Then, with* $\omega = \sqrt{\frac{n}{\kappa\alpha^2\gamma}}$, $s = 96\eta\omega\sqrt{\alpha}M\ln(n/\delta b)$ *and probability at least* $1 - \delta\frac{b^2}{n^2}$*:*

$$(1+s)^{-1}\,\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}} \leq \left\|\Delta_{t+1}\right\|_{\mathbf{H}} \leq (1+s)\,\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}.$$

Proof We first prove the right-hand side inequality (upper bound),

$$\begin{aligned}\left\|\Delta_{t+1}\right\|_{\mathbf{H}} &= \left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t\right\|_{\mathbf{H}} \\ &\leq \left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}} + \eta\left\|\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|_{\mathbf{H}} \\ &\leq \left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}} + \eta\left\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\mathbf{H}^{1/2}\right\| \cdot \left\|\mathbf{H}^{-1/2}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|.\end{aligned}$$

Since, $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}} \tilde{\mathbf{H}}$ and $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$, we have $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}(1+\epsilon_0\eta)} \mathbf{H}$, and therefore, $\left\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\mathbf{H}^{1/2}\right\| \leq \sqrt{\alpha}(1+\epsilon_0\eta)$. Using this we get,

$$\left\|\Delta_{t+1}\right\|_{\mathbf{H}} \leq \left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}} + \eta\frac{\sqrt{\alpha}(1+\epsilon_0\eta)}{\sqrt{\mu}}\left\|(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|.$$

Note that the first term is $\left\|\mathbb{E}_t\Delta_{t+1}\right\|_{\mathbf{H}}$. Also as $\|\Delta_t\|_{\mathbf{H}}^2 \leq e^2\|\tilde{\Delta}\|_{\mathbf{H}}^2$, we have $\mathbf{x}_t \in \mathcal{U}_f(e^2\epsilon_0\eta)$, and therefore, we can apply Lemma 9 on the second term. As M captures the effect of whether $b < \frac{8}{9}\kappa$ or $b \geq \frac{8}{9}\kappa$, we have with probability at least $1 - \delta\frac{b^2}{n^2}$,

$$\left\|\Delta_{t+1}\right\|_{\mathbf{H}} \leq \mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}} + 6\eta\sqrt{\alpha}\cdot M\ln(n/b\delta)\left(\left\|\Delta_t\right\|_{\mathbf{H}} + \left\|\tilde{\Delta}\right\|_{\mathbf{H}}\right).$$

By $\left\|\Delta_t\right\|_{\mathbf{H}} \leq e\cdot\left\|\tilde{\Delta}\right\|_{\mathbf{H}}$ it follows that,

$$\left\|\Delta_{t+1}\right\|_{\mathbf{H}} \leq \mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}} + 20\eta\sqrt{\alpha}\cdot M\ln(n/b\delta)\left\|\tilde{\Delta}\right\|_{\mathbf{H}},$$

and using Lemma 16, we upper bound $\left\|\tilde{\Delta}\right\|_{\mathbf{H}}$ by $2\omega\cdot\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}$ and get,

$$\left\|\Delta_{t+1}\right\|_{\mathbf{H}} \leq \left(1 + 40\omega\eta\sqrt{\alpha}\cdot M\ln(n/b\delta)\right)\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}}. \tag{9}$$

Next, we prove the left-hand side inequality (lower bound), observing that,

$$\begin{aligned}\mathbb{E}_t\left\|\Delta_{t+1}\right\|_{\mathbf{H}} &= \mathbb{E}_t\left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t\right\|_{\mathbf{H}} \\ &\leq \left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}} + \eta\mathbb{E}_t\left\|\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|_{\mathbf{H}} \\ &\leq \left\|\Delta_t - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t\right\|_{\mathbf{H}} + \eta\left\|\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|_{\mathbf{H}} + \eta\mathbb{E}_t\left\|\hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t)\right\|_{\mathbf{H}}.\end{aligned}$$

Note that,

$$\mathbb{E}_t \left\| \hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t) \right\|_{\mathbf{H}} \leq \left\| \mathbf{H}^{1/2} \hat{\mathbf{H}}^{-1} \mathbf{H}^{1/2} \right\| \cdot \left\| \mathbf{H}^{-1/2} \right\| \cdot \mathbb{E}_t \left\| (\bar{\mathbf{g}}_t - \mathbf{g}_t) \right\|$$
$$\leq (1 + \epsilon_0 \eta) \frac{\sqrt{\alpha}}{\sqrt{\mu}} \cdot \mathbb{E}_t \left\| (\bar{\mathbf{g}}_t - \mathbf{g}_t) \right\|.$$

Using Lemma 7 to upper bound $\mathbb{E}_t \left\| (\bar{\mathbf{g}}_t - \mathbf{g}_t) \right\|$ and substituting in the previous inequality for $\mathbb{E}_t \left\| \Delta_{t+1} \right\|_{\mathbf{H}}$, we get,

$$\mathbb{E}_t \left\| \Delta_{t+1} \right\|_{\mathbf{H}} \leq \left\| \Delta_{t+1} \right\|_{\mathbf{H}} + \eta \left\| \hat{\mathbf{H}}^{-1}(\bar{\mathbf{g}}_t - \mathbf{g}_t) \right\|_{\mathbf{H}}$$
$$+ \eta(1 + \epsilon_0 \eta) \sqrt{1 + c\epsilon_0 \eta} \cdot \frac{\sqrt{\kappa \alpha}}{\sqrt{b}} \left(\left\| \Delta_t \right\|_{\mathbf{H}} + \left\| \tilde{\Delta} \right\|_{\mathbf{H}} \right).$$

Now in the last term, we use $\left\| \Delta_t \right\|_{\mathbf{H}} < e \cdot \left\| \tilde{\Delta} \right\|_{\mathbf{H}}$ and upper bound the second term using Lemma 9. Following the same steps as we did for the right-hand side inequality, we get

$$\mathbb{E}_t \left\| \Delta_{t+1} \right\|_{\mathbf{H}} \leq \left\| \Delta_{t+1} \right\|_{\mathbf{H}} + 20\eta \sqrt{\alpha} \cdot M \ln(n/b\delta) \left\| \tilde{\Delta} \right\|_{\mathbf{H}} + 4\eta \sqrt{\alpha} \frac{\sqrt{\kappa}}{\sqrt{b}} \left\| \tilde{\Delta} \right\|_{\mathbf{H}}.$$

Using Lemma 16 on $\left\| \tilde{\Delta} \right\|_{\mathbf{H}}$,

$$\mathbb{E}_t \left\| \Delta_{t+1} \right\|_{\mathbf{H}} \leq \left\| \Delta_{t+1} \right\|_{\mathbf{H}} + 48\eta \sqrt{\alpha} \cdot M \ln(n/b\delta) \mathbb{E}_t \left\| \Delta_{t+1} \right\|_{\mathbf{H}},$$

implying,

$$\mathbb{E}_t \left\| \Delta_{t+1} \right\|_{\mathbf{H}} \leq \left(1 - 48\omega\eta \sqrt{\alpha} \cdot M \ln(n/b\delta) \right)^{-1} \left\| \Delta_{t+1} \right\|_{\mathbf{H}}.$$

By the condition on n, namely $n > \frac{(96)^2 \beta^2 \ln(n/b\delta)^2}{\gamma} \max\{\kappa, b\}$, it follows that,

$$\mathbb{E}_t \left\| \Delta_{t+1} \right\|_{\mathbf{H}} \leq \left(1 + 96\omega\eta \sqrt{\alpha} \cdot M \ln(n/b\delta) \right) \left\| \Delta_{t+1} \right\|_{\mathbf{H}}. \tag{10}$$

Combining (9) and (10) we conclude the proof. ■

Remark 19 *We make the following remark about the mini-batch size prescribed in Lemma 18, i.e.,* $b < \min\{\frac{n}{4\alpha\beta}, \frac{\gamma \cdot n}{(96)^2 \beta^2 \ln(n/b\delta)^2}\}$. *The first term,* $b < \frac{n}{4\alpha\beta}$, *ensures that optimal step size* η *can be picked for* $b \lesssim \frac{n}{\alpha \log(n)}$. *The second condition on* b *eventually reduces to* $b \lesssim \frac{n}{\log(n)}$, *as we set* $\gamma = 1280\beta^2 \ln(n/b\delta)$ *later in the analysis (see Theorem 21).*

4.4 Martingale setup

In what follows, we analyze the behavior of a carefully defined random process from random times L_i to T_i, for $i \geq 0$. In particular, we show that if $T_i = \frac{n}{b}$ then $\left\| \Delta_{T_i} \right\|_{\mathbf{H}}^2 < c_1 \frac{\kappa \alpha^2 \gamma}{n} \left\| \tilde{\Delta} \right\|_{\mathbf{H}}^2$ with very high

probability for some absolute constant c_1, and if $T_i < \frac{n}{b}$ then $\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ with very high probability. For any $t \geq 0$, such that $L_i + t < T_i$, we consider an event $\mathcal{A}_t^i$,

$$\mathcal{A}_t^i = \bigcap_{j=L_i}^{j=L_i+t} \left\{ \left\|\mathbf{g}_j - \bar{\mathbf{g}}_j\right\|_{\mathbf{H}} \leq 6M \ln(n/b\delta) \left\|\mathbf{x}_j - \tilde{\mathbf{x}}\right\|_{\mathbf{H}} \right\}.$$

The event $\mathcal{A}_t^i$ captures the occurrence of the high probability event mentioned in Lemma 9, for iterates ranging from random time L_i to $L_i + t$. Using Lemma 9, it follows that $\Pr\left(\mathcal{A}_{t+1}^i | \mathcal{A}_t^i\right) \geq 1 - \delta\frac{b^2}{n^2}$, and by the union bound $\Pr\left(\mathcal{A}_{t+1}^i | \mathcal{A}_0^i\right) \geq 1 - t\delta\frac{b^2}{n^2} > 1 - \delta\frac{b}{n}$. Consider a random process Y_t^i defined as:

$$Y_0^i = \ln(\left\|\Delta_{L_i}\right\|_{\mathbf{H}}),$$

and for $L_i + t < T_i$,

$$Y_{t+1}^i = \left(\ln(\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}) + \sum_{j=0}^{t} \ln\left(\frac{\left\|\Delta_{L_i+j}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+j}\left\|\Delta_{L_i+j+1}\right\|_{\mathbf{H}}} \right) \right) \cdot \mathbf{1}_{\mathcal{A}_t^i} + Y_t^i \cdot \mathbf{1}_{\neg\mathcal{A}_t^i},$$

where $\mathbb{E}_{L_i+t}$ means expectation conditioned on the past till iterate $\mathbf{x}_{L_i+t}$. At iteration $t+1$, the random process Y_{t+1}^i checks for the two halting conditions mentioned in the definition of T_i. If any of the halting condition is met then we get $L_i + t = T_i$, the random process halts, otherwise, the random process proceeds to iteration $t + 2$. We analyze the random process Y_{t+1}^i till $L_i + t = T_i$. The first observation is that Y_{t+1}^i is a sub-martingale, proven as follows. If $\mathbf{1}_{\mathcal{A}_t^i} = 0$ or $\mathbf{1}_{\mathcal{A}_{t-1}^i} = 0$ then $Y_{t+1}^i = Y_t^i$ and trivially $\mathbb{E}_t Y_i^{t+1} = Y_t^i$. So we consider $\mathbf{1}_{\mathcal{A}_t^i} = \mathbf{1}_{\mathcal{A}_{t-1}^i} = 1$ and get,

$$\begin{aligned}
\mathbb{E}_t[Y_{t+1}^i] - Y_t^i &= \mathbb{E}_{L_i+1} \ln(\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}) + \ln\left(\frac{\left\|\Delta_{L_i+t}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}} \right) - \ln(\left\|\Delta_{L_i+t}\right\|_{\mathbf{H}}) \\
&\leq \ln(\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}) + \ln\left(\frac{1}{\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}} \right) = 0. \qquad (11)
\end{aligned}$$

Due to the quadratic variation bound Lemma 17, for $L_i + t < T_i$ we have,

$$\mathbb{E}_{L_i+t}\left(\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}} - \mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}} \right)^2 \leq 80\eta\frac{\beta}{\gamma\sqrt{\alpha}}\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}^2. \qquad (12)$$

We use (12) to upper bound $\mathbb{E}_t\left(Y_{t+1}^i - Y_t^i\right)^2$, where $\mathbb{E}_t$ denotes the expectation conditioned on the past and assuming Y_t^i is known. First note that if $\mathbf{1}_{\mathcal{A}_t^i} = 0$ or $\mathbf{1}_{\mathcal{A}_{t-1}^i} = 0$, then $Y_{t+1}^i = Y_t^i$, and therefore

$\mathbb{E}_t\left(Y^i_{t+1} - Y^i_t\right)^2 = 0$. Hence, it remains to consider the case when $\mathbf{1}_{\mathcal{A}^i_t} = \mathbf{1}_{\mathcal{A}^i_{t-1}} = 1$,

$$\begin{aligned}\mathbb{E}_t\left(Y^i_{t+1} - Y^i_t\right)^2 &= \mathbb{E}_{L_i+t}\left(\ln\left\|\Delta_{L_i+t+1}\right\| - \ln\left\|\Delta_{L_i+t}\right\| + \ln\left(\frac{\left\|\Delta_{L_i+t}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|}\right)\right)^2 \\ &= \mathbb{E}_{L_i+t}\left(\ln\left(\frac{\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|}\right)\right)^2.\end{aligned}$$

Note that in the event of $\mathbf{1}_{\mathcal{A}^i_t} = 1$, by Lemma 18 we know that $\frac{1}{2} \leq \frac{\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|} \leq 2$, assuming that n and b satisfy the assumptions of Lemma 18. Consider the following inequality for $p, q > 0$ such that $\frac{1}{2} \leq \frac{p}{q} \leq 2$,

$$\left(\ln\left(\frac{p}{q}\right)\right)^2 \leq \max\left\{\ln\left(1 + \left(\frac{p}{q} - 1\right)^2\right), \ln\left(1 + \left(\frac{q}{p} - 1\right)^2\right)\right\}.$$

With $p = \left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}$ and $q = \mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}$, we use the above inequality to get,

$$\begin{aligned}\mathbb{E}_t\left(Y^i_{t+1} - Y^i_t\right)^2 &\leq \mathbb{E}_{L_i+t}\left[\ln\left(1 + 4\frac{\left(\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}} - \mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}\right)^2}{\left(\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}\right)^2}\right)\right] \\ &\leq \ln\left(1 + 320\eta\frac{\beta}{\gamma\sqrt{\alpha}}\right).\end{aligned}$$

where the last inequality is due to (12). Thus, we get the following predictable quadratic variation bound for random process Y^i_t,

$$\mathbb{E}_t\left(Y^i_{t+1} - Y^i_t\right)^2 \leq \ln\left(1 + 320\eta\frac{\beta}{\gamma\sqrt{\alpha}}\right). \tag{13}$$

Now we aim to upper bound $|Y^i_{t+1} - Y^i_t|$ for t such that $L_i + t < T_i$. Again, in the events $\mathbf{1}_{\mathcal{A}^i_t} = 0$ or $\mathbf{1}_{\mathcal{A}^i_{t-1}} = 0$, we have $Y^i_t = Y^i_{t-1}$ and we are done. Considering the case $\mathbf{1}_{\mathcal{A}^i_t} = \mathbf{1}_{\mathcal{A}^i_{t-1}} = 1$,

$$Y^i_{t+1} - Y^i_t = \ln\left(\frac{\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}}\right).$$

Again due to having $\mathbf{1}_{\mathcal{A}^i_t} = 1$, we invoke Lemma 18 to get,

$$\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}} \leq \left(1 + 96\eta\omega\sqrt{\alpha} \cdot M\ln(n/b\delta)\right)\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}},$$

where $\omega^2 = \frac{n}{\kappa\alpha^2\gamma}$, and

$$Y^i_{t+1} - Y^i_t \leq 2\ln\left(1 + 96\eta\omega\sqrt{\alpha}\cdot M\ln(n/b\delta)\right).$$

Similarly, we get,

$$Y^i_t - Y^i_{t+1} \leq 2\ln\left(1 + 96\eta\omega\sqrt{\alpha}\cdot M\ln(n/b\delta)\right).$$

Combining the above upper bound property with (11) and (13), we get the following submartingale framework.

Lemma 20 (Submartingale setup) *Let* $n > \frac{(96)^2\beta^2\ln(n/b\delta)^2}{\gamma}\kappa$, $\epsilon_0 < \frac{1}{8e^2\sqrt{\alpha}}$, $\eta = \frac{b\sqrt{\alpha}\beta}{n} \leq \min\left\{\frac{1}{4\sqrt{\alpha}}, \frac{b}{48\alpha^{3/2}\kappa}\right\}$ *and* $b < \min\{\frac{n}{4\alpha\beta}, \frac{\gamma\cdot n}{(96)^2\beta^2\ln(n/b\delta)^2}\}$ *for some* $\beta > 0$ *and* $\gamma > 0$. *Consider the random process defined as* $Y^i_0 = \ln(\left\|\Delta_{L_i}\right\|_{\mathbf{H}})$, *and for* $L_i + t < T_i$,

$$Y^i_{t+1} = \left(\ln(\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}) + \sum_{j=0}^{t}\ln\left(\frac{\left\|\Delta_{L_i+j}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+j+1}\right\|_{\mathbf{H}}}\right)\right)\cdot\mathbf{1}_{\mathcal{A}^i_t} + Y^i_t\cdot\mathbf{1}_{\neg\mathcal{A}^i_t}.$$

Then, letting $\omega^2 = \frac{n}{\kappa\alpha^2\gamma}$:

$$\begin{aligned}
\mathbb{E}_t[Y^i_{t+1}] &\leq Y^i_t,\\
|Y^i_{t+1} - Y^i_t| &\leq 2\ln\left(1 + 96\eta\omega\sqrt{\alpha}\cdot M\ln(n/b\delta)\right),\\
\mathbb{E}_t\left(Y^i_{t+1} - Y^i_t\right)^2 &\leq \ln\left(1 + 320\eta\frac{\beta}{\gamma\sqrt{\alpha}}\right).
\end{aligned}$$

4.5 High probability convergence via martingale framework

In the previous section, we constructed a submartingale framework satisfying a quadratic variation bound and high probability upper bound at every step. In this section, we invoke a well-known measure concentration result for martingales, Freedman's inequality stated in Theorem 13, on our framework. We apply Theorem 13 on the random process Y^i_t. For brevity, we provide the analysis only for the case $b < \frac{8}{9}\kappa$. This means replacing M by $\frac{\kappa}{b}$ in Lemma 20. The proof for the case $b \geq \frac{8}{9}\kappa$ follows similarly, by replacing M with $\sqrt{\frac{\kappa}{b}}$.

We consider $R = 2\ln\left(1 + \frac{96\eta\omega\sqrt{\alpha}\kappa\ln(n/b\delta)}{b}\right)$, $\sigma^2 = t\cdot\ln\left(1 + 320\eta\frac{\beta}{\gamma\sqrt{\alpha}}\right)$, fix $\lambda = 1$, and apply Freedman's inequality on the random process Y^i_t until $\left\|\Delta_t\right\|^2_{\mathbf{H}}$ does not satisfy any of stopping time criteria mentioned in the definition of T_i, (3). Consider two cases here,

Case 1: $\min\left\{\frac{\lambda^2}{\sigma^2}, \frac{\lambda}{R}\right\} = \frac{\lambda^2}{\sigma^2}$. Note that

$$\begin{aligned}\sigma^2 &\leq t \cdot \ln\left(1 + 320\eta\frac{\beta}{\gamma\sqrt{\alpha}}\right)\\ &\leq \frac{n}{b} \cdot \ln\left(1 + 320\eta\frac{\beta}{\gamma\sqrt{\alpha}}\right) \leq \frac{n}{b} \cdot 320\eta\frac{\beta}{\gamma\sqrt{\alpha}},\end{aligned}$$

where in the last inequality we use $t \leq \frac{n}{b}$ and $\ln(1+x) < x$. We get,

$$\exp\left(-\frac{1}{4} \cdot \frac{\lambda}{\sigma^2}\right) \leq \exp\left(-\frac{1}{4} \cdot \frac{1}{320\frac{n}{b}\frac{b\sqrt{\alpha\beta}}{n}\frac{\beta}{\gamma\sqrt{\alpha}}}\right) = \exp\left(-\frac{1}{4} \cdot \frac{\gamma}{320\beta^2}\right) < \delta\frac{b}{n},$$

where last inequality holds if $\gamma > 1280\beta^2 \ln(n/b\delta)$.

Case 2: $\min\left\{\frac{\lambda^2}{\sigma^2}, \frac{\lambda}{R}\right\} = \frac{\lambda}{R}$. Note that

$$\begin{aligned}R &= 2\ln\left(1 + \frac{96\eta\omega\sqrt{\alpha}\kappa\ln(n/b\delta)}{b}\right)\\ &< \frac{200\eta\omega\sqrt{\alpha}\kappa\ln(1/\delta)}{b},\end{aligned}$$

where in the last inequality we use $\ln(1+x) < x$. We get,

$$\exp\left(-\frac{1}{200} \cdot \frac{\lambda}{R}\right) \leq \exp\left(-\frac{1}{4} \cdot \frac{b}{200\eta\omega\sqrt{\alpha}\kappa\ln(n/b\delta)}\right).$$

Substitute $\eta = \frac{b\sqrt{\alpha}}{n} \cdot \beta$ and $\omega = \sqrt{\frac{n}{\kappa\alpha^2\gamma}}$, we get,

$$\begin{aligned}\exp\left(-\frac{1}{4} \cdot \frac{\lambda}{R}\right) &\leq \exp\left(-\frac{1}{4} \cdot \frac{b}{200\frac{b\sqrt{\alpha}\beta}{n}\sqrt{\frac{n}{\kappa\alpha^2\gamma}}\sqrt{\alpha}\kappa\ln(n/b\delta)}\right)\\ &= \exp\left(-\frac{1}{4} \cdot \frac{\sqrt{n}\sqrt{\gamma}}{200\beta\sqrt{\kappa}\ln(n/b\delta)}\right) \leq \delta\frac{b}{n}.\end{aligned}$$

Letting $\gamma > 1280\beta^2\ln(n/b\delta)$, $n > 400\kappa\alpha^2\gamma\ln(n/b\delta)^2$, we get the failure probability less than $\delta \cdot \frac{b}{n}$. Combining both cases, we get the following powerful concentration guarantee.

Theorem 21 (Freedman's concentration) *Let* $n > 400\kappa\alpha^2\gamma(\ln(n/b\delta))^2$, $\gamma > 1280\beta^2\ln(n/b\delta)$, $\epsilon_0 < \frac{1}{8e^2\sqrt{\alpha}}$, $\eta = \frac{b\sqrt{\alpha}\beta}{n} \leq \min\{\frac{1}{4\sqrt{\alpha}}, \frac{b}{48\alpha^{3/2}\kappa}\}$ *and* $b < \min\{\frac{n}{4\alpha\beta}, \frac{\gamma\cdot n}{(96)^2\beta^2\ln(n/b\delta)^2}\}$. *Then, for any* $t \geq 0$ *satisfying* $L_i + t \leq T_i$*:*

$$\ln\left(\frac{\left\|\Delta_{L_i+t}\right\|_{\mathbf{H}}}{\left\|\Delta_{L_i}\right\|_{\mathbf{H}}}\right) \leq -t\ln\left(\frac{1}{1-\rho}\right) + 1,$$

where $\rho = \frac{\eta}{32\sqrt{\alpha}}$, *with probability at least* $1 - \delta\frac{b}{n}$.

Proof By Theorem 13, we have $Y_t \leq Y_0 + 1$ with probability at least $1 - \delta\frac{b}{n}$. This implies,

$$\left(\ln(\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}) + \sum_{j=0}^{t} \ln\left(\frac{\left\|\Delta_{L_i+t}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+t}\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}}\right)\right) \cdot \mathbf{1}_{\mathcal{A}_t^i} + Y_t^i \cdot \mathbf{1}_{\neg\mathcal{A}_t^i} \leq \ln\left(\left\|\Delta_{L_i}\right\|_{\mathbf{H}}\right) + 1. \tag{14}$$

Since $L_i + t \leq T_i$, we have $\mathbf{x}_{L_i+t-1} \in \mathcal{U}_f(e^2\epsilon_0\eta)$ and also $\mathbf{x}_{L_i} \in \mathcal{U}_f(e^2\epsilon_0\eta)$. By Lemma 9, and applying the union bound for t inner iterations starting from $\mathbf{x}_{L_i}$ we have $\Pr\left(\mathcal{A}_t^i\right) \geq 1 - t\delta\frac{b^2}{n^2} \geq 1 - \delta\frac{b}{n}$. Combining this with (14) and rescaling δ by a factor of 2, we get with probability at least $1 - \delta\frac{b}{n}$,

$$\left(\ln(\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}) + \sum_{j=0}^{t} \ln\left(\frac{\left\|\Delta_{L_i+j}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+j}\left\|\Delta_{L_i+j+1}\right\|_{\mathbf{H}}}\right)\right) \leq \ln\left(\left\|\Delta_{L_i}\right\|_{\mathbf{H}}\right) + 1.$$

By Lemma 14, with $\rho = \frac{\eta}{32\sqrt{\alpha}}$, we have $\ln\left(\frac{1}{1-\rho}\right) < \ln\left(\frac{\left\|\Delta_{L_i+j}\right\|_{\mathbf{H}}}{\mathbb{E}_{L_i+j}\left\|\Delta_{L_i+j+1}\right\|_{\mathbf{H}}}\right)$, for any j such that $L_i + j \leq T_i$. So, with probability at least $1 - \delta\frac{b}{n}$,

$$\left(\ln(\left\|\Delta_{L_i+t+1}\right\|_{\mathbf{H}}) + (t+1) \cdot \ln\left(\frac{1}{1-\rho}\right)\right) \leq \ln\left(\left\|\Delta_{L_i}\right\|_{\mathbf{H}}\right) + 1,$$

which concludes the proof. ■

Remark 22 *We make the following remarks about the Freedman's concentration results (Theorem 21):*

(a) An upper bound on $\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2$,

$$\Pr\left(\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 \leq e^2\left\|\Delta_{L_i}\right\|_{\mathbf{H}}^2\right) \geq 1 - \delta\frac{b}{n}. \tag{15}$$

This holds since we run the submartingale Y_t^i *till* $L_i + t = T_i$. *Applying the Freedman concentration for* t *such that* $L_i + t = T_i$ *yields (15);*

(b) In particular for $i = 0$, *we have,*

$$\Pr\left(\left\|\Delta_{T_0}\right\|_{\mathbf{H}}^2 \leq e^2\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2\right) \geq 1 - \delta\frac{b}{n}; \tag{16}$$

(c) Suppose $T_0 < \frac{n}{b}$, *by combining the definition of* T_0 (3) *and* (16) *we get,*

$$\Pr\left(\left\|\Delta_{T_0}\right\|_{\mathbf{H}}^2 \leq \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \mid T_0 < \frac{n}{b}\right) \geq 1 - \delta\frac{b}{n}. \tag{17}$$

The next two results show that for any stopping time T_i, $\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ with high probability, and, moreover, if $T_i < \frac{n}{b}$ then $\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$. We start by proving the result for the first stopping time T_0.

Lemma 23 (High probability result for first stopping time) *Let the conditions of Theorem 21 hold. Then:*

$$\Pr\left(\left\|\Delta_{T_0}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2\right) \geq 1-\delta\frac{b}{n},$$

where $\beta \geq 32\ln\left(\frac{n}{2\alpha^2\kappa}\right)$, *and* $\gamma \geq 1280\beta^2\ln(n/b\delta)$.

Proof By conditioning on the first stopping time, whether $T_0 < \frac{n}{b}$ or $T_0 = \frac{n}{b}$, and using law of total probability, it follows that,

$$\begin{aligned}\Pr\left(\left\|\Delta_{T_0}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2\right) &= \Pr\left(\left\|\Delta_{T_0}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \;|T_0 = \frac{n}{b}\right)\Pr\left(T_0 = \frac{n}{b}\right)\\ &\quad + \Pr\left(\left\|\Delta_{T_0}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \;|T_0 < \frac{n}{b}\right)\Pr\left(T_0 < \frac{n}{b}\right).\end{aligned}$$

Note that we bound the second term above by (17). We proceed to bound the first term. Consider the submartingale Y_t^0 running till stopping time T_0. If $T_0 = \frac{n}{b}$, then this implies that the submartingale Y_t^0 continued for the full outer iteration. We apply the result of Theorem 21 on Y_t^0 to get, with probability at least $1-\delta\frac{b}{n}$,

$$\ln\left(\frac{\left\|\Delta_{n/b}\right\|_{\mathbf{H}}}{\left\|\tilde{\Delta}\right\|_{\mathbf{H}}}\right) \leq -\frac{n}{b}\ln\left(\frac{1}{1-\rho}\right)+1,$$

which implies,

$$\ln\left(\frac{\left\|\Delta_{n/b}\right\|_{\mathbf{H}}^2}{\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2}\right) \leq -\frac{2n}{b}\ln\left(\frac{1}{1-\rho}\right)+2 \quad\Rightarrow\quad \left\|\Delta_{n/b}\right\|_{\mathbf{H}}^2 \leq e^2(1-\rho)^{2n/b}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2.$$

For $\beta = 32\ln\left(\frac{n}{2\alpha^2\kappa}\right)$ and $\gamma = 1280\beta^2\ln(n/b\delta)$, it follows that $(1-\rho)^{2n/b} \leq 2\frac{\kappa\alpha^2\gamma}{n}$, which concludes the proof. ■

Since there can be more than one stopping time, we prove a result similar to Lemma 23 for stopping times occurring after the first stopping time. Note that this requires conditioning on the previous stopping time. As we already have an unconditional high probability result for T_0 in Lemma 23, having conditioning on previous stopping time allows us to establish a high probability result for any subsequent stopping time.

Lemma 24 (High probability result for non-first stopping time) *Let the conditions of Theorem 21 hold. Then, for any $i \geq 0$:*

$$\Pr\left(\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \mid \left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2\right) \geq 1-\delta\frac{b}{n},$$

where $\beta \geq 32\ln\left(\frac{n}{2\alpha^2\kappa}\right)$ and $\gamma \geq 1280\beta^2\ln(n/b\delta)$.

Proof First, note that as $\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$, we have $\left\|\Delta_{L_{i+1}-1}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$. Now due to one step high probability bound in Lemma 18 we have $\left\|\Delta_{L_{i+1}}\right\|_{\mathbf{H}}^2 \leq 2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$. Similar to Lemma 23 we show that,

$$\Pr\left(\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \mid \left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2, T_{i+1} = n/b\right) \geq 1-\delta\frac{b}{n}, \tag{18}$$

and,

$$\Pr\left(\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 \leq \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \mid \left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2, T_{i+1} < n/b\right) \geq 1-\delta\frac{b}{n}. \tag{19}$$

Now we invoke the Freedman concentration result Theorem 21 on the sub-martingale Y_t^{i+1}, to get with probability at least $1-\delta\frac{b}{n}$,

$$\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 \leq e^2\left\|\Delta_{L_{i+1}}\right\|_{\mathbf{H}}^2,$$

proving that

$$\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2.$$

Note that this condition does not depend on the value of T_{i+1}. In particular, for $T_{i+1} = \frac{n}{b}$ we get,

$$\Pr\left(\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \mid \left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 \leq \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2, T_{i+1} = \frac{n}{b}\right) \geq 1-\delta\frac{b}{n},$$

establishing the inequality (18). Next, to prove the inequality (19), observe that $T_{i+1} < \frac{n}{b}$ implies that either $\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 > 2e^2\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$ or $\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2$. We have already shown that with probability at least $1-\delta\frac{b}{n}$,

$$\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 < 2e^2\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2.$$

Combining this with $T_{i+1} < \frac{n}{b}$ we get,

$$\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2,$$

obtaining that

$$\Pr\left(\left\|\Delta_{T_{i+1}}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2 \mid \left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2, T_{i+1} < \frac{n}{b}\right) \geq 1 - \delta\frac{b}{n}.$$

which concludes the proof. ■

Note that for the β and γ values prescribed in Lemma 23 and the lower bound on n from Theorem 21, a sufficient condition on n can be stated as: $n > c_1\kappa\alpha^2\ln(2c_1\kappa\alpha^2/\delta)^7$ for $c_1 = 400\cdot 1280\cdot(32)^2$. We are now ready to prove our main convergence result.

Theorem 25 (High probability convergence over outer iterates) *Let Assumption 1 hold, and for fixed $\alpha \geq 1$, let Hessian oracle return an α-approximation to $\nabla^2 f(\mathbf{x})$ for any queried $\mathbf{x}$. Let $n > c_1\kappa\alpha^2\ln(2c_1\kappa\alpha^2/\delta)^7$, and $\epsilon_0 < \frac{1}{8e^2\sqrt{\alpha}}$. Let $\tilde{\mathbf{x}}_0 \in \mathcal{U}_f(\epsilon_0\eta)$, $\eta = \frac{b\sqrt{\alpha\beta}}{n}$ and $b < \frac{n}{c_2\alpha\ln(n/\kappa)}$. Then for any $0 < \delta < 1/2$, with probability at least $1 - 2\delta$:*

$$f(\tilde{\mathbf{x}}_s) - f(\mathbf{x}^*) \leq \left(\frac{c_3\kappa\alpha^2\ln(n/\delta)^2}{n}\right)^s \cdot \left(f(\tilde{\mathbf{x}}_0) - f(\mathbf{x}^*)\right),$$

for $\beta = O\big(\ln(\frac{n}{\alpha^2\kappa})\big)$, $\gamma = O\big(\beta^2\ln(\frac{n}{b\delta})\big)$ and absolute constants c_1, c_2 and c_3.

Proof Note that in our running notation, $\tilde{\mathbf{x}}_0$ is just $\tilde{\mathbf{x}}$. We prove one outer iteration result where $\tilde{\mathbf{x}}_1$ means $\mathbf{x}_{n/b}$. Consider $\beta = 32\ln\left(\frac{n}{2\alpha^2\kappa}\right)$ and $\gamma = 1280\beta^2\ln(n/b\delta)$ and note that due to our assumptions, the results of Theorem 21 and Lemmas 23 and 24 hold. Also for the given values of β and γ, the upper bound condition on b can be replaced with $b < \frac{n}{c_2\alpha\ln(n/2\alpha^2\kappa)}$, where $c_2 = 32\cdot 4$. Let $\mathcal{S}$ be the event denoting our convergence guarantee, i.e.,

$$\mathcal{S} := \left\{\left\|\Delta_{n/b}\right\|_{\mathbf{H}}^2 \leq 2e^2\frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2\right\}.$$

Also we define events $\mathcal{E}_i$ for all $i \geq 0$ as following,

$$\mathcal{E}_i := \left\{T_i < \frac{n}{b}\right\} \cap \left\{\left\|\Delta_{T_i}\right\|_{\mathbf{H}}^2 < \frac{\kappa\alpha^2\gamma}{n}\left\|\tilde{\Delta}\right\|_{\mathbf{H}}^2\right\}.$$

Due to Lemma 24 we know that $\Pr\left(\mathcal{E}_{i+1}|\mathcal{E}_i, T_{i+1} < \frac{n}{b}\right) \geq 1 - \delta\frac{b}{n}$. Moreover, by using the law of total probability in conjunction with conditioning on the first stopping time, we have,

$$\Pr(\mathcal{S}) = \Pr\left(\mathcal{S}\,|T_0 = \frac{n}{b}\right)\cdot\Pr\left(T_0 = \frac{n}{b}\right) + \Pr\left(\mathcal{S}\,|T_0 < \frac{n}{b}\right)\cdot\Pr\left(T_0 < \frac{n}{b}\right).$$

Using Lemma 23, we have $\Pr\left(\mathcal{S}\,|T_0 = \frac{n}{b}\right) \geq 1 - \delta\frac{b}{n}$, and

$$\Pr(\mathcal{S}) = \left(1 - \delta\frac{b}{n}\right)\cdot\Pr\left(T_0 = \frac{n}{b}\right) + \Pr\left(\mathcal{S}\,|T_0 < \frac{n}{b}\right)\cdot\Pr\left(T_0 < \frac{n}{b}\right). \tag{20}$$

Next, we consider the term $\Pr\left(\mathcal{S} \mid T_0 < \frac{n}{b}\right)$,

$$\Pr\left(\mathcal{S} \mid T_0 < \frac{n}{b}\right) \geq \Pr\left(\mathcal{S} \mid \mathcal{E}_0, T_0 < \frac{n}{b}\right) \cdot \Pr\left(\mathcal{E}_0 \mid T_0 < \frac{n}{b}\right).$$

By Lemma 23, $\Pr\left(\mathcal{E}_0 \mid T_0 < \frac{n}{b}\right) \geq 1 - \delta\frac{b}{n}$. Substituting this in (20),

$$\Pr(\mathcal{S}) \geq \left(1 - \delta\frac{b}{n}\right)\left[\Pr\left(T_0 = \frac{n}{b}\right) + \Pr\left(\mathcal{S} \mid \mathcal{E}_0\right) \cdot \Pr\left(T_0 < \frac{n}{b}\right)\right]. \tag{21}$$

We now show that $\Pr\left(\mathcal{S} \mid \mathcal{E}_i\right) \geq 1 - \delta$ for any i. Consider the following,

$$\begin{aligned}\Pr\left(\mathcal{S} \mid \mathcal{E}_i\right) = &\Pr\left(\mathcal{S} \mid T_{i+1} = \frac{n}{b}, \mathcal{E}_i\right) \cdot \Pr\left(T_{i+1} = \frac{n}{b} \mid \mathcal{E}_i\right) \\ &+ \Pr\left(\mathcal{S} \mid T_{i+1} < \frac{n}{b}, \mathcal{E}_i\right) \cdot \Pr\left(T_{i+1} < \frac{n}{b} \mid \mathcal{E}_i\right).\end{aligned}$$

By Lemma 24, we have $\Pr\left(\mathcal{S} \mid T_{i+1} = \frac{n}{b}, \mathcal{E}_i\right) \geq 1 - \delta\frac{b}{n}$,

$$\begin{aligned}\Pr\left(\mathcal{S} \mid \mathcal{E}_i\right) \geq &\left(1 - \delta\frac{b}{n}\right) \cdot \Pr\left(T_{i+1} = \frac{n}{b} \mid \mathcal{E}_i\right) \\ &+ \Pr\left(\mathcal{S} \mid T_{i+1} < \frac{n}{b}, \mathcal{E}_i\right) \cdot \Pr\left(T_{i+1} < \frac{n}{b} \mid \mathcal{E}_i\right).\end{aligned} \tag{22}$$

We write $\Pr\left(\mathcal{S} \mid T_{i+1} < \frac{n}{b}, \mathcal{E}_i\right)$ as follows,

$$\begin{aligned}\Pr\left(\mathcal{S} \mid T_{i+1} < \frac{n}{b}, \mathcal{E}_i\right) &\geq \Pr\left(\mathcal{S} \mid T_{i+1} < \frac{n}{b}, \mathcal{E}_i, \mathcal{E}_{i+1}\right) \cdot \Pr\left(\mathcal{E}_{i+1} | T_{i+1} < \frac{n}{b}, \mathcal{E}_i\right) \\ &= \Pr\left(\mathcal{S} \mid \mathcal{E}_{i+1}\right) \cdot \Pr\left(\mathcal{E}_{i+1} | T_{i+1} < \frac{n}{b}, \mathcal{E}_i\right),\end{aligned}$$

where in the last inequality we use $\Pr\left(\mathcal{S} \mid T_{i+1} < \frac{n}{b}, \mathcal{E}_i, \mathcal{E}_{i+1}\right) = \Pr\left(\mathcal{S} \mid \mathcal{E}_{i+1}\right)$. We also use Lemma 24 to write $\Pr\left(\mathcal{E}_{i+1} | T_{i+1} < \frac{n}{b}, \mathcal{E}_i\right) \geq 1 - \delta\frac{b}{n}$, implying,

$$\Pr\left(\mathcal{S} \mid T_{i+1} < \frac{n}{b}, \mathcal{E}_i\right) \geq \left(1 - \delta\frac{b}{n}\right)\Pr\left(\mathcal{S} \mid \mathcal{E}_{i+1}\right). \tag{23}$$

Substituting (23) in (22), we get,

$$\Pr\left(\mathcal{S} \mid \mathcal{E}_i\right) \geq \left(1 - \delta\frac{b}{n}\right)\left[\Pr\left(T_{i+1} = \frac{n}{b} \mid \mathcal{E}_i\right) + \Pr\left(\mathcal{S} \mid \mathcal{E}_{i+1}\right)\Pr\left(T_{i+1} < \frac{n}{b} \mid \mathcal{E}_i\right)\right].$$

Since we perform a finite number of iterations, $\Pr\left(T_{i+1} < \frac{n}{b} \mid \mathcal{E}_i\right) = 0$ for any $i \geq \frac{n}{b}$. This intuitively means that the number of stopping times is upper bounded by the number of inner iterations. In the worst case $i = \frac{n}{b} - 1$, which means that $\Pr\left(T_{i+1} < \frac{n}{b} \mid \mathcal{E}_i\right) = 0$ for $i = \frac{n}{b} - 1$. implying that $\Pr\left(\mathcal{S} | \mathcal{E}_{\frac{n}{b}-1}\right) \geq 1 - \delta\frac{b}{n}$. Also note that if $\Pr\left(\mathcal{S} \mid \mathcal{E}_{i+1}\right) \geq \left(1 - \delta\frac{b}{n}\right)^l$ for some integer $l \geq 0$, then

$\Pr\left(\mathcal{S} \mid \mathcal{E}_i\right) \geq \left(1 - \delta\frac{b}{n}\right)^{l+1}$. Combining this with the fact there cannot be more than $\frac{n}{b}$ stopping times, along with using (21) we get,

$$\Pr\left(\mathcal{S}\right) \geq 1 - \delta.$$

Incorporating the total failure probability of Lemma 9 for any of the n/b iterations, we get a failure probability of at most 2δ.

As a final step, we prove our result in terms of function values. Since f has continuous first- and second-order derivatives, by the quadratic Taylor expansion, for vectors $\mathbf{a}$ and $\mathbf{v}$, there exists a $\theta \in [0, 1]$ such that,

$$f(\mathbf{a} + \mathbf{v}) = f(\mathbf{a}) + \langle \nabla f(\mathbf{a}), \mathbf{v} \rangle + \frac{1}{2}\mathbf{v}^\top \nabla^2 f(\mathbf{a} + \theta\mathbf{v})\mathbf{v}.$$

Let $\mathbf{a} = \mathbf{x}^*$, $\mathbf{v} = \mathbf{x}_{n/b} - \mathbf{x}^*$,

$$f(\mathbf{x}_{n/b}) - f(\mathbf{x}^*) = \frac{1}{2}(\mathbf{x}_{n/b} - \mathbf{x}^*)^\top \nabla^2 f(\mathbf{x}^* + \theta(\mathbf{x}_{n/b} - \mathbf{x}^*))(\mathbf{x}_{n/b} - \mathbf{x}^*).$$

With high probability, we know that $\mathbf{x}_{n/b} \in \mathcal{U}_f(e^2\epsilon_0\eta)$ and have $\mathbf{x}^* + \theta(\mathbf{x}_{n/b} - \mathbf{x}^*) \in \mathcal{U}_f(e^2\epsilon_0\eta)$. Take $\mathbf{z} = \mathbf{x}^* + \theta(\mathbf{x}_{n/b} - \mathbf{x}^*)$, we have $\frac{1}{1+e^2\epsilon_0\eta} \cdot \mathbf{H} \preceq \nabla^2 f(\mathbf{z}) \preceq (1 + e^2\epsilon_0\eta) \cdot \mathbf{H}$. Then,

$$\begin{aligned}
\frac{1}{2}(\mathbf{x}_{n/b} - \mathbf{x}^*)^\top \nabla^2 f(\mathbf{x}^* + \theta(\mathbf{x}_{n/b} - \mathbf{x}^*))^\top (\mathbf{x}_{n/b} - \mathbf{x}^*) &= \frac{1}{2}\|\mathbf{x}_{n/b} - \mathbf{x}^*\|^2_{\nabla^2 f(\mathbf{z})} \\
&\leq \frac{1}{2}(1 + e^2\epsilon_0\eta)\left\|\Delta_{n/b}\right\|^2_{\mathbf{H}} \\
&< \left\|\Delta_{n/b}\right\|^2_{\mathbf{H}},
\end{aligned}$$

so it follows that,

$$f(\mathbf{x}_{n/b}) - f(\mathbf{x}^*) \leq \left\|\Delta_{n/b}\right\|^2_{\mathbf{H}}.$$

Using the reverse inequality with $\tilde{\mathbf{x}}$ in place of $\mathbf{x}_{n/b}$, we have,

$$f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*) \geq 2(1 + \epsilon_0)\left\|\tilde{\Delta}\right\|^2_{\mathbf{H}}.$$

Combining the above two relations with the definition of $\mathcal{S}$, we get with probability at least $1 - 2\delta$,

$$f(\mathbf{x}_{n/b}) - f(\mathbf{x}^*) \leq 6e^2\frac{\kappa\alpha^2\gamma}{n} \cdot \left(f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right).$$

Setting $c_3 = 6e^2 \cdot 1280 \cdot 32$, and writing the result in terms of outer iterates by noticing $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}_0$ and $\mathbf{x}_{n/b} = \tilde{\mathbf{x}}_1$, concludes the proof. ■

4.6 Global convergence analysis

In this section, we prove the global convergence of Mb-SVRN. Unlike Theorem 25 we have no local neighborhood condition and we do not put any condition on the gradient mini-batch size. Due to being a stochastic second-order method, the global rate of convergence of Mb-SVRN is provably much slower than the local convergence guarantee.

Theorem 26 (Global convergence of Mb-SVRN) *For any gradient mini-batch size b and step size $\eta = \min\left\{\frac{2}{\kappa\sqrt{\alpha}}, \frac{b}{8\kappa^3\alpha^{3/2}}\right\}$ there exists $\rho' < 1$ such that,*

$$\mathbb{E} f(\mathbf{x}_{n/b}) - f(\mathbf{x}^*) < \rho' \cdot \left(f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right).$$

Proof Refer to Appendix A.6. ■

5. Numerical Experiments

We now present more extensive empirical evidence to support our theoretical analysis. We considered a regularized logistic loss minimization task with two datasets, EMNIST dataset ($n \approx 700k$) and CIFAR10 dataset ($n = 60k$) transformed using a random feature map obtaining $d = 256$ features for both datasets.

5.1 Experimental setup

The regularized logistic loss function can be expressed as the following finite-sum objective:

$$f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \ln(1 + e^{-b_i \mathbf{a}_i^\top \mathbf{x}}) + \frac{\mu}{2}\|\mathbf{x}\|^2.$$

where $(\mathbf{a}_i, b_i)$ denote the training examples with $\mathbf{a}_i \in \mathbb{R}^d$ and $b_i \in \{+1, -1\}$ for $i \in \{1, 2, \ldots, n\}$. The function $f(\mathbf{x})$ is μ-strongly convex. We set $\mu = 10^{-6}$ in our experiments on EMNIST and CIFAR10. Before running Mb-SVRN, for both datasets, we find the respective $\mathbf{x}^*$ to high precision by using Newton's Method. This is done in order to calculate the error $f(\mathbf{x}_t) - f(\mathbf{x}^*)$ at every iteration t. For the Hessian approximation in Mb-SVRN, we use Hessian subsampling (Roosta-Khorasani and Mahoney, 2019; Bollapragada et al., 2019), sampling h component Hessians at the snapshot vector $\tilde{\mathbf{x}}$ and constructing $\hat{\mathbf{H}}_s$. The algorithmic implementation is discussed in Algorithm 1. For every combination of gradient mini-batch size (b), step size (η) and Hessian sample size (h) chosen for any particular dataset, we tune the number of inner iterations $t_{\max}$ to optimize the convergence rate per data pass and take the average over 5 runs.[3] Moreover, as Katyusha has two additional momentum parameters τ_1 and τ_2, we set $\tau_2 = 1/2b$ in our experiments. This is done to avoid high computational costs associated with tuning 3 hyperparameters for Katyusha. However, tuning η and τ_1 itself turns out to be a computationally heavy task, highlighting an additional drawback of Katyusha over Mb-SVRN, which in comparison requires tuning only η. All methods get a fixed budget equivalent to performing 4 data passes, *i.e.*, computing $4n$ component gradients.[4] The allowed computational

3. For SVRG and Katyusha, we have $h = 0$.
4. Setting $\tau_2 = 1/2b$ agrees with suggested parameter setting in Allen-Zhu (2017).

budget of 4 data passes can be used in 1 outer iteration by performing $3n$ inner iterations, or 2 outer iterations with n inner iterations each, or 3 outer iterations with $\lfloor n/3 \rfloor$ inner iterations each, or even 4 outer iterations with no inner iterations. In addition to these four regimes, we also consider a fifth scheme where we tune the number of inner iterations, say t_{tuned}, and run $\left\lfloor \frac{4}{1+bt_{tuned}/n} \right\rfloor$ outer iterations with t_{tuned} inner iterations each.[5] We tune the above mentioned 5 regimes over η, and τ_1 (only for `Katyusha`) and return $\hat{\mathbf{x}}$ as an estimate for $\mathbf{x}^*$, say after a total of w data passes. Here w is atmost 4 but could be less than 4; depending on t_{tuned}. The convergence rate per data pass is computed as $\hat{\rho} := \left(\frac{f(\hat{\mathbf{x}}) - f(\mathbf{x}^*)}{f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)} \right)^{1/w}$, where $\tilde{\mathbf{x}}$ is the initial iterate.

In the following sections, we discuss the robustness of `Mb-SVRN`, as well as `SVRG` and `Katyusha`, to 1) gradient mini-batch size and 2) step size.

5.2 Robustness to gradient mini-batch size

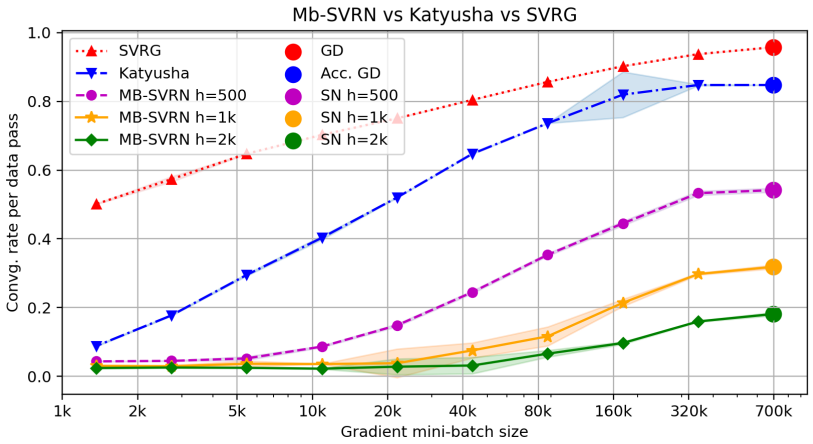
We first discuss Figure 5 showing the convergence rate of `Mb-SVRN` as we vary gradient mini-batch size and Hessian sample size for solving regularized logistic regression tasks on `EMNIST` and `CIFAR10` datasets. The convergence rates reported are obtained after tuning the convergence rate per data pass with respect to the step size η and number of inner iterations, for any (h, b) value pair. The theory suggests that the convergence rate of `Mb-SVRN` is independent of the gradient mini-batch size b for a wide range of mini-batch sizes, and the plot recovers this phenomenon remarkably accurately. The plot highlights that `Mb-SVRN` is robust to gradient mini-batch size, since the fast convergence rate of `Mb-SVRN` is preserved for a very large range of gradient mini-batch sizes (represented by the curves in Figure 5 staying flat). This phenomenon does not hold for first-order variance reduction methods, `SVRG` and `Katyusha`. The plots demonstrate that the performance of `SVRG` and `Katyusha` suffers with increasing b, which is consistent with the existing convergence analysis of `SVRG`-type first-order methods (Konecnỳ and Richtárik, 2013; Allen-Zhu, 2017). While `Katyusha` does provide an accelerated convergence in comparison to `SVRG`, the convergence rate per data pass still degrades rapidly as b grows.

We also note that as b increases and enters into a very large gradient mini-batch size regime, the convergence rate of `Mb-SVRN` starts to deteriorate and effectively turns into Subsampled Newton (`SN`) when $b = n$. The empirical evidence showing deterioration in convergence rate for very large b values agrees with our theoretical prediction of a phase transition into standard Newton's method after $b > \frac{n}{\alpha \log(n)}$. In the extreme case of $b = n$, `Mb-SVRN` performs only one inner iteration and is the same as `SN`. As `SN` uses the exact gradient at every iteration, its convergence rate per data pass is very sensitive to the Hessian approximation quality, which makes it substantially worse than `Mb-SVRN` for small-to-moderate Hessian sample sizes h.

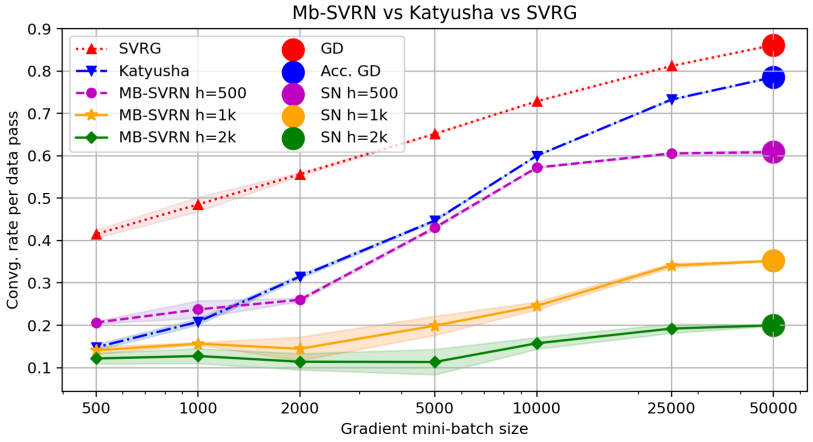
5.3 Robustness to step size

In addition to the demonstrated robustness to gradient mini-batch size, `Mb-SVRN` exhibits empirical resilience to step size variations. As depicted in Figure 6, the convergence rate for small b values closely aligns with the optimal rate (red dot) when the step size approaches the optimal range. In contrast, the convergence rate of Subsampled Newton ($b = n$) sharply increases near its optimal step size. This suggests that the convergence rate of `Mb-SVRN` with small-to-moderate gradient mini-batch size is more robust to changes in step size as compared to using very large gradient mini-

5. For `Mb-SVRN` we include the component Hessian evaluations as well in the computational budget



(a) Convergence rate per data pass of `Mb-SVRN` on `EMNIST` dataset.



(b) Convergence rate per data pass of `Mb-SVRN` on `CIFAR10` dataset.

Figure 5: Experiments on `EMNIST` and `CIFAR10` datasets, as we vary gradient mini-batch size b and Hessian sample size h, showing the robustness of `Mb-SVRN` to gradient mini-batch size and phase transition into standard Newton's method for large mini-batches.

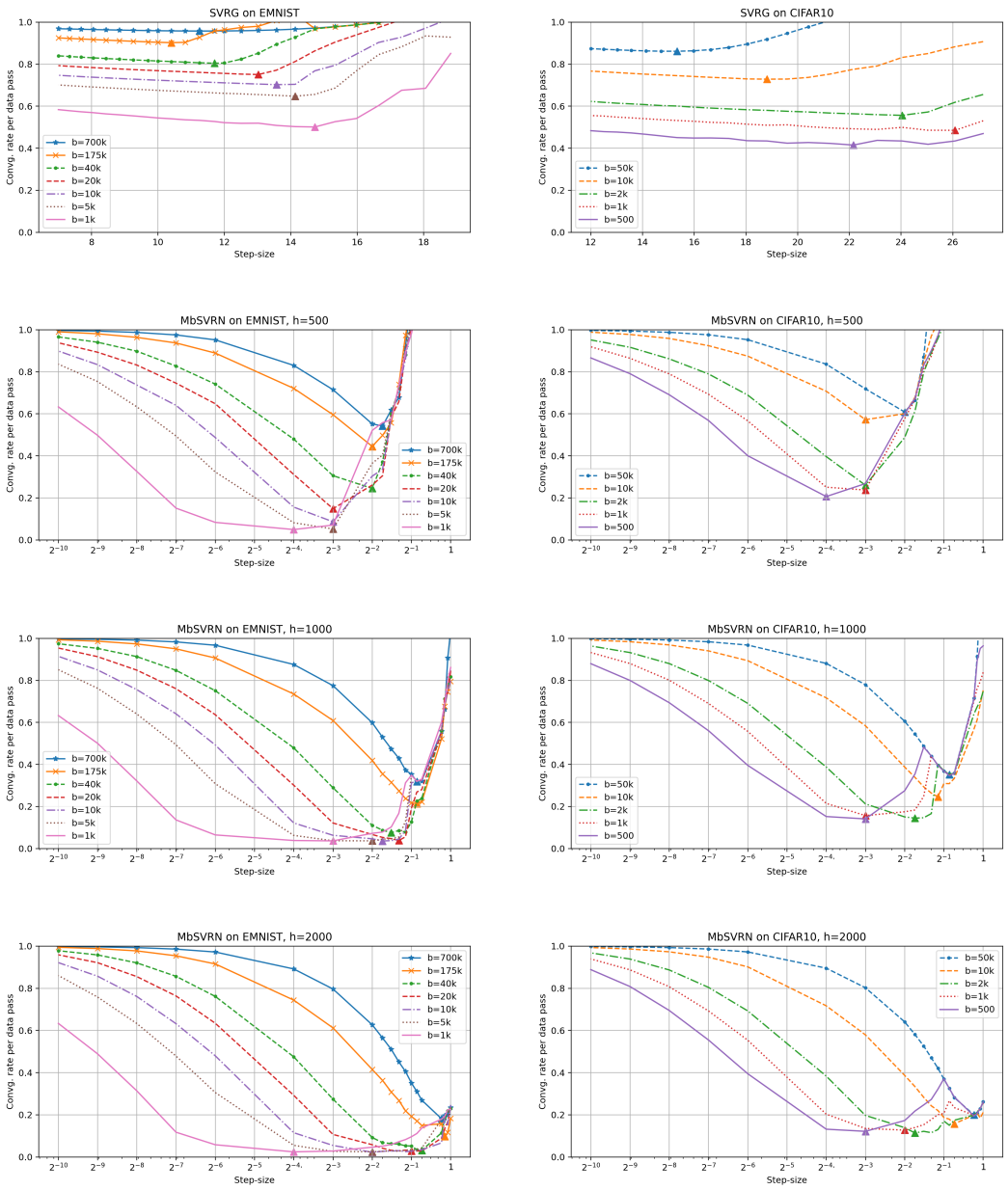


Figure 6: Experiments with logistic regression on `EMNIST` on `CIFAR10` datasets. The big dot on every curve marks the respective optimal convergence rate attained at the optimal step size. The bottom six plots demonstrate the performance of `Mb-SVRN` with different Hessian sample sizes h. The top two plots demonstrate `SVRG`'s performance.

batches or the full gradients. Intuitively, with smaller b values, the algorithm performs numerous inner iterations, relying more on variance reduction. This advantage offsets the impact of step size changes. Conversely, with very large b values, `Mb-SVRN` selects larger step sizes, reducing the number of optimal inner iterations and limiting the variance reduction advantage. On the other hand, for first-order methods like `SVRG` and `Katyusha`, tuning over the hyperparameters η and τ_1 can lead to a huge computational burden (which we observed in our experiments especially in the case of `Katyusha`), as these hyperparameters depend on unknown problem parameters (like smoothness constant).

6. Conclusions and Future Directions

We have shown that incorporating second-order information into a stochastic variance-reduced method allows it to scale more effectively, and to scale to very large mini-batches. We have demonstrated this by analyzing the convergence of `Mb-SVRN`, a prototypical stochastic second-order method with variance reduction, and have shown that its associated convergence rate per data pass remains optimal for a very wide range of gradient mini-batch sizes (up to $n/\alpha \log(n)$). Our main theoretical result provides a convergence guarantee robust to the gradient mini-batch size with high probability through a novel martingale concentration argument. Furthermore, empirically we have shown the robustness of `Mb-SVRN` not only to mini-batch size, but also to the step size and the Hessian approximation quality. Our algorithm, analysis, and implementation uses `SVRG`-type variance reduced gradients, and as such, a natural question pertains to whether the algorithm can be extended to use other (perhaps biased) variance reduction techniques. Another interesting future direction is to investigate the effect of using alternate sampling methods while selecting component gradients, as well as the effect of incorporating acceleration into the method.

Acknowledgments

This material is based upon work supported by the Office of Naval Research under award number N00014-21-1-2532, and by the National Science Foundation under award number CCF-2338655.

Appendix A. Ommitted Proofs

A.1 Proof of Lemma 7.

Proof As each ψ_i is convex and λ-smooth, the following relation holds (see Theorem 2.1.5 Nesterov et al., 2018),

$$\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}})\|^2 \leq 2\lambda \cdot (\psi_i(\tilde{\mathbf{x}}) - \psi_i(\mathbf{x}_t) - \langle \nabla\psi_i(\mathbf{x}_t), \tilde{\mathbf{x}} - \mathbf{x}_t\rangle).$$

Consider the variance of the stochastic gradient if we use just one sample, ψ_i for calculating the stochastic gradient. The variance is given as $\mathbb{E}[\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}} - \mathbf{g}_t\|^2]$.

$$\begin{aligned}
\mathbb{E}_t[\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}} - \mathbf{g}_t\|^2] &= \mathbb{E}_t[\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}}) - (\mathbf{g}_t - \tilde{\mathbf{g}})\|^2] \\
&\leq \mathbb{E}_t[\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}})\|^2] \\
&\leq 2\lambda \cdot \mathbb{E}_t[(\psi_i(\tilde{\mathbf{x}}) - \psi_i(\mathbf{x}_t) - \langle \nabla\psi_i(\mathbf{x}_t), \tilde{\mathbf{x}} - \mathbf{x}_t\rangle)] \\
&= 2\lambda \cdot (f(\tilde{\mathbf{x}}) - f(\mathbf{x}_t) - \langle \mathbf{g}_t, \tilde{\mathbf{x}} - \mathbf{x}_t\rangle). \qquad (24)
\end{aligned}$$

Similarly, if we use b samples and upper bound $\mathbb{E}\|\bar{\mathbf{g}}_t - \mathbf{g}_t\|^2$,

$$\begin{aligned}\mathbb{E}_t\|\bar{\mathbf{g}}_t - \mathbf{g}_t\|^2 &= \mathbb{E}_t\|\frac{1}{b}\sum_{t=1}^{b}\nabla\psi_{i_t}(\mathbf{x}_t) - \frac{1}{b}\sum_{t=1}^{b}\nabla\psi_{i_t}(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}} - \mathbf{g}_t\|^2 \\ &= \frac{1}{b^2}\mathbb{E}_t\|\sum_{t=1}^{b}\nabla\psi_{i_t}(\mathbf{x}_t) - \sum_{t=1}^{b}\nabla\psi_{i_t}(\tilde{\mathbf{x}}) + b\cdot\tilde{\mathbf{g}} - b\cdot\mathbf{g}_t\|^2 \\ &= \frac{1}{b^2}\mathbb{E}_t\|\sum_{t=1}^{b}(\nabla\psi_{i_t}(\mathbf{x}_t) - \nabla\psi_{i_t}(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}} - \mathbf{g}_t)\|^2.\end{aligned}$$

Now as all the indices i_t are chosen independently, we can write the variance of the sum as the sum of individual variances.

$$\begin{aligned}\mathbb{E}_t\|\bar{\mathbf{g}}_t - \mathbf{g}_t\|^2 &= \frac{1}{b^2}\sum_{t=1}^{b}\mathbb{E}_t\|(\nabla\psi_{i_t}(\mathbf{x}_t) - \nabla\psi_{i_t}(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}} - \mathbf{g}_t)\|^2 \\ &= \frac{1}{b}\mathbb{E}_t\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}} - \mathbf{g}_t\|^2 \\ &\leq \frac{2\lambda}{b}(f(\tilde{\mathbf{x}}) - f(\mathbf{x}_t) - \langle\mathbf{g}_t, \tilde{\mathbf{x}} - \mathbf{x}_t\rangle), \end{aligned} \tag{25}$$

where in the last inequality, we used (24). Since f has continuous first and second-order derivatives, we can use the quadratic Taylor's expansion for f around $\mathbf{x}_t$. For vectors $\mathbf{a}$ and $\mathbf{v}$, $\exists\ \theta \in [0,1]$ such that,

$$f(\mathbf{a}+\mathbf{v}) = f(\mathbf{a}) + \langle\nabla f(\mathbf{a}), \mathbf{v}\rangle + \frac{1}{2}\mathbf{v}^\top\nabla^2 f(\mathbf{a}+\theta\mathbf{v})\mathbf{v}.$$

Let $\mathbf{a} = \mathbf{x}_t$, $\mathbf{v} = \tilde{\mathbf{x}} - \mathbf{x}_t$, we get,

$$f(\tilde{\mathbf{x}}) - f(\mathbf{x}_t) - \langle\mathbf{g}_t, (\tilde{\mathbf{x}} - \mathbf{x}_t)\rangle = \frac{1}{2}(\tilde{\mathbf{x}} - \mathbf{x}_t)^\top\nabla^2 f(\mathbf{x}_t + \theta(\tilde{\mathbf{x}} - \mathbf{x}_t))(\tilde{\mathbf{x}} - \mathbf{x}_t).$$

Using the assumption $\tilde{\mathbf{x}}, \mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$, we have $\mathbf{x}_t + \theta(\tilde{\mathbf{x}} - \mathbf{x}_t) \in \mathcal{U}_f(c\epsilon_0\eta)$. Take $\mathbf{z} = \mathbf{x}_t + \theta(\tilde{\mathbf{x}} - \mathbf{x}_t)$, we have $\frac{1}{1+c\epsilon_0\eta}\cdot\mathbf{H} \preceq \nabla^2 f(\mathbf{z}) \preceq (1+c\epsilon_0\eta)\cdot\mathbf{H}$ (see Proof of Lemma 1 in Dereziński (2025)). We get,

$$\begin{aligned}\frac{1}{2}(\tilde{\mathbf{x}} - \mathbf{x}_t)^\top\nabla^2 f(\mathbf{x}_t + \theta(\tilde{\mathbf{x}} - \mathbf{x}_t))^\top(\tilde{\mathbf{x}} - \mathbf{x}_t) &= \frac{1}{2}\|\mathbf{x}_t - \tilde{\mathbf{x}}\|^2_{\nabla^2 f(\mathbf{z})} \\ &\leq \frac{1}{2}(1 + c\epsilon_0\eta)\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|^2_{\mathbf{H}}.\end{aligned}$$

implying,

$$f(\tilde{\mathbf{x}}) - f(\mathbf{x}_t) - \langle\mathbf{g}_t, (\tilde{\mathbf{x}} - \mathbf{x}_t)\rangle \leq \frac{1}{2}(1 + c\epsilon_0\eta)\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|^2_{\mathbf{H}}.$$

Substitute the above relation in (25), we get,

$$\mathbb{E}_t\|\bar{\mathbf{g}}_t - \mathbf{g}_t\|^2 \leq (1 + c\epsilon_0\eta)\frac{\lambda}{b}\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|^2_{\mathbf{H}}.$$

■

A.2 Proof of Lemma 8.

Proof For the sake of this proof we abuse the notation and write $\mathbf{x}_{\mathbf{AN}}$ as $\mathbf{x}_{t+1}$. However we make clear that `Mb-SVRN` uses $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t$. In this proof we denote $\mathbf{x}_{t+1} = \mathbf{x}_{\mathbf{AN}} = \mathbf{x}_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t$, $\Delta_{t+1} = \mathbf{x}_{\mathbf{AN}} - \mathbf{x}^*$, and $\Delta_t = \mathbf{x}_t - \mathbf{x}^*$.

$$\begin{aligned}
\Delta_{t+1} &= \Delta_t - \eta\hat{\mathbf{H}}^{-1}\mathbf{g}_t \\
&= \Delta_t - \eta\hat{\mathbf{H}}^{-1}(\mathbf{g}_t - \mathbf{g}^*) \\
&= \Delta_t - \eta\hat{\mathbf{H}}^{-1}\int_0^1 \nabla^2 f(\mathbf{x}^* + \theta\Delta_t)\Delta_t d\theta \\
&= \Delta_t - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}\Delta_t \\
\Rightarrow \left\|\Delta_{t+1}\right\|_{\bar{\mathbf{H}}} &= \left\|(\mathbf{I} - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}})\Delta_t\right\|_{\bar{\mathbf{H}}},
\end{aligned}$$

where $\bar{\mathbf{H}} = \int_0^1 \nabla^2 f(\mathbf{x}^* + \theta\Delta_t)d\theta$. We upper bound $\left\|\Delta_{t+1}\right\|_{\bar{\mathbf{H}}}$ as,

$$\begin{aligned}
\left\|\Delta_{t+1}\right\|_{\bar{\mathbf{H}}} &= \left\|(\mathbf{I} - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}})\Delta_t\right\|_{\bar{\mathbf{H}}} \\
&= \left\|\bar{\mathbf{H}}^{1/2}(\mathbf{I} - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}})\Delta_t\right\| \\
&= \left\|(\mathbf{I} - \eta\bar{\mathbf{H}}^{1/2}\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}^{1/2})\bar{\mathbf{H}}^{1/2}\Delta_t\right\| \\
&\leq \left\|\mathbf{I} - \eta\bar{\mathbf{H}}^{1/2}\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}^{1/2}\right\| \cdot \left\|\Delta_t\right\|_{\bar{\mathbf{H}}}.
\end{aligned}$$

As $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$, for any $0 \leq \theta \leq 1$, we have $\mathbf{x}^* + \theta\Delta_t \in \mathcal{U}_f(c\epsilon_0\eta)$, and therefore $\frac{1}{1+c\epsilon_0\eta}\mathbf{H} \preceq \bar{\mathbf{H}} \preceq (1 + c\epsilon_0\eta)\mathbf{H}$. Also we have $\frac{1}{\sqrt{\alpha}}\tilde{\mathbf{H}} \preceq \hat{\mathbf{H}} \preceq \sqrt{\alpha}\tilde{\mathbf{H}}$ and $\frac{1}{1+\epsilon_0\eta}\mathbf{H} \preceq \tilde{\mathbf{H}} \preceq (1 + \epsilon_0\eta)\mathbf{H}$. Combining these positive semidefinite orderings for $\tilde{\mathbf{H}}$ and $\hat{\mathbf{H}}$ along with $1 + \epsilon_0\eta < 2$, we have $\frac{1}{2\sqrt{\alpha}}\mathbf{H} \preceq \hat{\mathbf{H}} \preceq 2\sqrt{\alpha}\mathbf{H}$. We get,

$$\begin{aligned}
&\frac{1}{2\sqrt{\alpha}(1 + c\epsilon_0\eta)}\bar{\mathbf{H}}^{-1} \preceq \hat{\mathbf{H}}^{-1} \preceq 2\sqrt{\alpha}(1 + c\epsilon_0\eta)\bar{\mathbf{H}}^{-1}, \\
&\frac{\eta}{2\sqrt{\alpha}(1 + c\epsilon_0\eta)}\mathbf{I} \preceq \eta\bar{\mathbf{H}}^{1/2}\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}^{1/2} \preceq 2\eta\sqrt{\alpha}(1 + c\epsilon_0\eta)\mathbf{I},
\end{aligned}$$

implying that,

$$\left\|\mathbf{I} - \eta\bar{\mathbf{H}}^{1/2}\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}^{1/2}\right\| \leq \max\{1 - \frac{\eta}{2\sqrt{\alpha}(1 + c\epsilon_0\eta)}, 2\eta\sqrt{\alpha}(1 + c\epsilon_0\eta) - 1\}.$$

Since $\eta < \frac{1}{4\sqrt{\alpha}}$, and $\epsilon_0 < \frac{1}{8c\sqrt{\alpha}}$, maximum value would be $1 - \frac{\eta}{2\sqrt{\alpha}(1+c\epsilon_0\eta)}$. We get,

$$\left\|\mathbf{I} - \eta\bar{\mathbf{H}}^{1/2}\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}^{1/2}\right\| \leq 1 - \frac{\eta}{4\sqrt{\alpha}},$$

and hence,

$$\left\|\Delta_{t+1}\right\|_{\bar{\mathbf{H}}} \leq \left(1 - \frac{\eta}{4\sqrt{\alpha}}\right)\left\|\Delta_t\right\|_{\bar{\mathbf{H}}}.$$

Changing $\|\cdot\|_{\bar{\mathbf{H}}}$ to $\|\cdot\|_{\mathbf{H}}$ we get,

$$\|\Delta_{t+1}\|_{\mathbf{H}} \leq \sqrt{1+c\epsilon_0\eta}\,\|\Delta_{t+1}\|_{\bar{\mathbf{H}}} \leq \sqrt{1+c\epsilon_0\eta}\cdot\left(1-\frac{\eta}{4\sqrt{\alpha}}\right)\|\Delta_t\|_{\bar{\mathbf{H}}} \leq (1+c\epsilon_0\eta)\cdot\left(1-\frac{\eta}{4\sqrt{\alpha}}\right)\|\Delta_t\|_{\mathbf{H}}$$
$$\leq \left(1-\eta\left(\frac{1}{4\sqrt{\alpha}}-c\epsilon_0\right)-\frac{c\epsilon_0\eta^2}{4\sqrt{\alpha}}\right)\|\Delta_t\|_{\mathbf{H}}.$$

Using $\epsilon_0 < \frac{1}{8c\sqrt{\alpha}}$, we conclude,

$$\|\Delta_{t+1}\|_{\mathbf{H}} \leq \left(1-\frac{\eta}{8\sqrt{\alpha}}\right)\|\Delta_t\|_{\mathbf{H}}.$$

■

A.3 Proof of Lemma 9.

In the proof, we will use a result from Minsker (2011), stated as the following,

Lemma 27 (Matrix Bernstein: Corollary 4.1 from Minsker (2011)) *Let $\mathbf{Y}_1, \mathbf{Y}_2, ..\mathbf{Y}_m \in \mathbb{C}^d$ be a sequence of random vectors such that $\mathbb{E}\mathbf{Y}_i = 0$, $\|\mathbf{Y}_i\| < U$ almost surely $\forall 1 \leq i \leq m$. Denote $\sigma^2 := \sum_{i=1}^m \mathbb{E}\|\mathbf{Y}_i\|^2$. Then $\forall t^2 > \sigma^2 + \frac{tU}{3}$,*

$$\Pr\left(\left\|\sum_{i=1}^m \mathbf{Y}_i\right\|_2 > t\right) \leq 28exp\left[-\frac{t^2/2}{\sigma^2+tU/3}\right].$$

We now return to the proof of Lemma 9.

Proof Define random vectors $\mathbf{v}_i$, as $\mathbf{v}_i = \nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}} - \mathbf{g}_t$. Then $\mathbb{E}[\mathbf{v}_i] = 0$. Also,

$$\begin{aligned}
\|\mathbf{v}_i\|^2 &\leq 2\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}})\|^2 + 2\|\mathbf{g}_t - \tilde{\mathbf{g}}\|^2 \\
&\leq 4\lambda\left(\psi_i(\tilde{\mathbf{x}}) - \psi_i(\mathbf{x}_t) - \langle\nabla\psi_i(\mathbf{x}_t), \tilde{\mathbf{x}} - \mathbf{x}_t\rangle\right) + 4\lambda\left(f(\tilde{\mathbf{x}}) - f(\mathbf{x}_t) - \langle\mathbf{g}_t, \tilde{\mathbf{x}} - \mathbf{x}_t\rangle\right) \\
&\leq 2\lambda^2\|\tilde{\mathbf{x}} - \mathbf{x}_t\|^2 + 2\lambda^2\|\tilde{\mathbf{x}} - \mathbf{x}_t\|^2 \\
&= 4\lambda^2\|\tilde{\mathbf{x}} - \mathbf{x}_t\|^2.
\end{aligned}$$

We also have the following upper bound on the variance of $\mathbf{v}_i$:

$$\begin{aligned}
\mathbb{E}[\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}}) + \tilde{\mathbf{g}} - \mathbf{g}_t\|^2] &= \mathbb{E}[\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}}) - (\mathbf{g}_t - \tilde{\mathbf{g}})\|^2] \\
&\leq \mathbb{E}[\|\nabla\psi_i(\mathbf{x}_t) - \nabla\psi_i(\tilde{\mathbf{x}})\|^2] \\
&\leq 2\lambda\cdot\mathbb{E}[(\psi_i(\tilde{\mathbf{x}}) - \psi_i(\mathbf{x}_t) - \langle\nabla\psi_i(\mathbf{x}_t), \tilde{\mathbf{x}} - \mathbf{x}_t\rangle)] \\
&= 2\lambda\cdot(f(\tilde{\mathbf{x}}) - f(\mathbf{x}_t) - \langle\mathbf{g}_t, \tilde{\mathbf{x}} - \mathbf{x}_t\rangle).
\end{aligned}$$

Since f is twice continuously differentiable, there exists a $\theta \in [0,1]$ such that,

$$f(\tilde{\mathbf{x}}) - f(\mathbf{x}_t) - \langle\mathbf{g}_t, (\tilde{\mathbf{x}} - \mathbf{x}_t)\rangle = \frac{1}{2}(\tilde{\mathbf{x}} - \mathbf{x}_t)^\top\nabla^2 f(\mathbf{x}_t + \theta(\tilde{\mathbf{x}} - \mathbf{x}_t))^\top(\tilde{\mathbf{x}} - \mathbf{x}_t).$$

Using the assumption that $\mathbf{x}_t, \tilde{\mathbf{x}} \in \mathcal{U}_f(c\epsilon_0\eta)$, with $\mathbf{z} = \mathbf{x}_t + \theta(\tilde{\mathbf{x}} - \mathbf{x}_t)$ we have $\mathbf{z}_t \in \mathcal{U}_f(c\epsilon_0\eta)$. With $\epsilon_0 < \frac{1}{8c\sqrt{\alpha}}$, and $\eta < \frac{1}{4\sqrt{\alpha}}$,

$$\begin{aligned}\frac{1}{2}(\tilde{\mathbf{x}} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t + \theta(\tilde{\mathbf{x}} - \mathbf{x}_t))^\top (\tilde{\mathbf{x}} - \mathbf{x}_t) &= \frac{1}{2}\|\mathbf{x}_t - \tilde{\mathbf{x}}\|^2_{\nabla^2 f(\mathbf{z})} \\ &\leq \frac{1}{2}(1 + c\epsilon_0\eta)\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|^2_{\mathbf{H}} \\ &\leq \left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|^2_{\mathbf{H}}.\end{aligned}$$

We get,

$$\mathbb{E}\|\mathbf{v}_i\|^2 \leq 2\lambda\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|^2_{\mathbf{H}}.$$

So we take $U = 2\lambda\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|$ and $\sigma^2 = 2b\lambda\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|^2_{\mathbf{H}}$. Also note that, $U \leq 2\frac{\lambda}{\sqrt{\mu}}\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|_{\mathbf{H}}$. Look for t such that,

$$\exp\left(-\frac{t^2/2}{\sigma^2 + tU/3}\right) < \delta\frac{b^2}{n^2}$$

$$\begin{aligned}&\iff \frac{t^2/2}{\sigma^2 + tU/3} > 2\ln(n/b\delta) \\ &\iff t^2 > 4\sigma^2\ln(n/b\delta) + \frac{4tU}{3}\ln(n/b\delta) \\ &\iff \frac{t^2}{2} + \frac{t^2}{2} > 4\sigma^2\ln(n/b\delta) + \frac{4tU}{3}\ln(n/b\delta).\end{aligned}$$

Now in the regime of $b < \frac{8}{9}\kappa$ we have $2\sigma\ln(n/b\delta) < \frac{4}{3}U\ln(n/b\delta)$. Consider $\mathbf{Y}_i = \mathbf{v}_i$, $t = \frac{8}{3}U\ln(n/b\delta) = \frac{4}{3}.\frac{2\lambda}{\sqrt{\mu}}\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|_{\mathbf{H}}\ln(\frac{n}{b\delta})$. For this value of t, we get with probability $1 - \delta\frac{b^2}{n^2}$,

$$\left\|\sum_{i=1}^{b}\mathbf{Y}_i\right\| \leq \frac{8}{3}.\frac{2\lambda}{\sqrt{\mu}}\left\|\tilde{\mathbf{x}} - \mathbf{x}_t\right\|_{\mathbf{H}}\ln(\frac{n}{b\delta}).$$

This means with probability $1 - \delta\frac{b^2}{n^2}$,

$$\left\|\mathbf{g}_t - \bar{\mathbf{g}}_t\right\| \leq \frac{16\lambda}{3b\sqrt{\mu}}\ln(n/b\delta)\left\|\mathbf{x}_t - \tilde{\mathbf{x}}\right\|_{\mathbf{H}}.$$

Moreover, in the regime of $b \geq \frac{8}{9}\kappa$, we have $2\sigma\ln(n/b\delta) \geq \frac{4}{3}U\ln(n/b\delta)$. In this case set $t = 2\sqrt{2}\sigma\ln(n/b\delta) = 4\sqrt{b\lambda}\left\|\mathbf{x}_t - \tilde{\mathbf{x}}\right\|_{\mathbf{H}}\ln(n/b\delta)$. We conclude,

$$\left\|\mathbf{g}_t - \bar{\mathbf{g}}_t\right\| \leq \frac{4\sqrt{\lambda}}{\sqrt{b}}\ln(n/b\delta)\left\|\mathbf{x}_t - \tilde{\mathbf{x}}\right\|_{\mathbf{H}}.$$

■

A.4 Proof of Lemma 10.

Proof Since $\mathbf{g}^* = 0$, we have $\left\|\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}} = \left\|\hat{\mathbf{H}}^{-1}(\mathbf{g}_t - \mathbf{g}^*)\right\|_{\mathbf{H}}$. Consider the following,

$$\left\|\hat{\mathbf{H}}^{-1}\mathbf{g}_t\right\|_{\mathbf{H}} = \left\|\hat{\mathbf{H}}^{-1}(\mathbf{g}_t - \mathbf{g}^*)\right\|_{\mathbf{H}} = \left\|\hat{\mathbf{H}}^{-1}\int_0^1 \nabla^2 f(\mathbf{x}^* + \theta(\mathbf{x}_t - \mathbf{x}^*))\Delta_t \cdot d\theta\right\|_{\mathbf{H}}.$$

Denoting $\bar{\mathbf{H}} = \int_0^1 \nabla^2 f(\mathbf{x}^* + \theta(\mathbf{x}_t - \mathbf{x}^*)) \cdot d\theta$, we get,

$$\begin{aligned}
\left\|\hat{\mathbf{H}}^{-1}(\mathbf{g}_t - \mathbf{g}^*)\right\|_{\mathbf{H}} &= \left\|\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}\Delta_t\right\|_{\mathbf{H}} \\
&= \left\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}\Delta_t\right\| \\
&= \left\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}\mathbf{H}^{-1/2}\mathbf{H}^{1/2}\Delta_t\right\| \\
&\leq \left\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\bar{\mathbf{H}}\mathbf{H}^{-1/2}\right\| \|\Delta_t\|_{\mathbf{H}} \\
&= \left\|\underbrace{\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\mathbf{H}^{1/2}}\underbrace{\mathbf{H}^{-1/2}\bar{\mathbf{H}}\mathbf{H}^{-1/2}}\right\| \|\Delta_t\|_{\mathbf{H}} \\
&\leq \left\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\mathbf{H}^{1/2}\right\| \cdot \left\|\mathbf{H}^{-1/2}\bar{\mathbf{H}}\mathbf{H}^{-1/2}\right\| \cdot \|\Delta_t\|_{\mathbf{H}}.
\end{aligned}$$

Since we have $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}} \tilde{\mathbf{H}}$ and $\tilde{\mathbf{x}} \in \mathcal{U}_f(\epsilon_0\eta)$, we have $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}(1+\epsilon_0\eta)} \mathbf{H}$. Furthermore for $\mathbf{x}_t \in \mathcal{U}_f(c\epsilon_0\eta)$, we have $\bar{\mathbf{H}} \approx_{1+c\epsilon_0\eta} \mathbf{H}$, because for all $\theta, \mathbf{x}^* + \theta(\mathbf{x}_t - \mathbf{x}^*) \in \mathcal{U}_f(c\epsilon_0\eta)$. So we use the results $\left\|\mathbf{H}^{1/2}\hat{\mathbf{H}}^{-1}\mathbf{H}^{1/2}\right\| \leq \sqrt{\alpha}(1+\epsilon_0\eta)$ and $\left\|\mathbf{H}^{-1/2}\bar{\mathbf{H}}\mathbf{H}^{-1/2}\right\| \leq (1+c\epsilon_0\eta)$ and get,

$$\begin{aligned}
\left\|\hat{\mathbf{H}}^{-1}(\mathbf{g}_t - \mathbf{g}^*)\right\|_{\mathbf{H}} &\leq \sqrt{\alpha}(1+\epsilon_0\eta)(1+c\epsilon_0\eta)\|\Delta_t\|_{\mathbf{H}} \\
&< 2\sqrt{\alpha}\|\Delta_t\|_{\mathbf{H}}.
\end{aligned}$$

■

A.5 Proof of Theorem 13.

In our proof of Theorem 13, we use a Master tail bound for adapted sequenced from Tropp (2011). Let $\mathcal{F}_k$ be a filtration and random process $(Y_k)_{k\geq 0}$ be $\mathcal{F}_k$ measurable. Also let another random process V_k such that V_k is $\mathcal{F}_{k-1}$ measurable. Consider the difference sequence for $k \geq 1$,

$$X_k = Y_k - Y_{k-1}.$$

Also, assume the following relation holds for a function $g : (0,\infty) \to [0,\infty]$:

$$\mathbb{E}_{k-1}e^{\theta\mathbf{X}_k} \leq e^{g(\theta)V_k}.$$

Then we have,

Theorem 28 (Master tail bound for adapted sequences from Tropp (2011)) *For all $\zeta, w \in \mathbb{R}$, we have,*

$$\Pr\left(\exists k \geq 0 : Y_k \geq Y_0 + \zeta \;\; \textit{and} \;\; \sum_{i=1}^{k} V_i \leq \sigma^2\right) \leq \inf_{\theta>0} e^{-\zeta\theta + g(\theta)\sigma^2}.$$

In our proof, we also use the following Lemma from Tropp (2011). The original version of the result assumes that $\mathbb{E}[X] = 0$, however, we show below that the proof also holds for the case when $\mathbb{E}[X] < 0$.

Lemma 29 (Freedman MGF, Lemma 6.7 from Tropp (2012)) *Let X be a random variable such that $\mathbb{E}X \leq 0$ and $X \leq R$ almost surely. Then for any $\theta > 0$ and $h(R) = \frac{e^{\theta R} - \theta R - 1}{R^2}$,*

$$\mathbb{E}e^{\theta X} \leq e^{h(R)\mathbb{E}[X^2]}.$$

Proof Consider the Taylor series expansion of $e^{\theta x}$,

$$\begin{aligned} e^{\theta x} &= 1 + x + \frac{x^2}{2!} + \cdots \\ &= 1 + x + x^2 h(x). \end{aligned}$$

Replace x with the random variable X and take expectation on both sides, we get,

$$\mathbb{E}e^{\theta X} \leq 1 + \mathbb{E}X + \mathbb{E}[X^2 \cdot h(X)].$$

On the second term use $\mathbb{E}X \leq 0$ and on the third term use $X \leq R$ to get $h(X) \leq h(R)$ almost surely, we get,

$$\mathbb{E}e^{\theta X} \leq 1 + h(R) \cdot \mathbb{E}[X^2] \leq e^{h(R) \cdot \mathbb{E}[X^2]}.$$

■

We now return to the proof of Theorem 13.

Proof If Y_k is a submartingale we have,

$$\mathbb{E}_{k-1} X_k \leq 0,$$

and also we know that $X_k \leq R$. For $k \geq 1$, consider $V_k = \mathbb{E}_{k-1}(X_k^2)$. Clearly, V_k is $\mathcal{F}_{k-1}$ measurable. We now establish the relation that $\mathbb{E}_{k-1} e^{\theta \mathbf{X}_k} \leq e^{g(\theta)V_k}$. For any $\theta > 0$, consider a function $h(x) : [0, \infty] \to [\frac{\theta^2}{2}, \infty)$ defined as follows,

$$h(x) = \frac{e^{\theta x} - \theta x - 1}{x^2}, \text{ and } \; h(0) = \frac{\theta^2}{2}.$$

It is easy to show that $h(x)$ is an increasing function of x. Using Lemma 29 we get $\mathbb{E}_{k-1} e^{\theta \mathbf{X}_k} \leq e^{g(\theta)V_k}$ for $g(\theta) = \frac{e^{\theta R} - \theta R - 1}{R^2}$. Now we can use the Master tail bound Theorem 28. The only thing remaining to analyze is,

$$\inf_{\theta > 0} e^{\theta\zeta + g(\theta)\cdot\sigma^2}.$$

Doing little calculus shows that, $\theta = \frac{1}{R} \ln\left(1 + \frac{\zeta R}{\sigma^2}\right)$ minimizes $e^{\theta\zeta + g(\theta)\cdot\sigma^2}$ and the minimum value is $e^{-\frac{1}{2}\left(\frac{\zeta}{\sigma^2} + \frac{\zeta}{R}\right)}$. Finally observe that $e^{-\frac{1}{2}\left(\frac{\zeta}{\sigma^2} + \frac{\zeta}{R}\right)} \leq e^{-\frac{1}{4}\min\left(\frac{\zeta}{\sigma^2}, \frac{\zeta}{R}\right)}$. This completes the proof. ■

A.6 Proof of Theorem 26.

Proof Using the λ-smoothness of f, we know that

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) + \mathbf{g}_t^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{\lambda}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2.$$

Substitute $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t$, we get,

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) - \eta\mathbf{g}_t^\top\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t + \eta^2\frac{\lambda}{2}\|\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t\|^2 \\ &= f(\mathbf{x}_t) - \eta\mathbf{g}_t^\top\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t + \eta^2\frac{\lambda}{2}\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t - \mathbf{g}_t + \mathbf{g}_t\right)\|^2. \end{aligned}$$

Take total expectation $\mathbb{E}$ on both sides, by which we mean expectation conditioned only on known $\tilde{\mathbf{x}}$. We get,

$$\begin{aligned} \mathbb{E}f(\mathbf{x}_{t+1}) &\leq \mathbb{E}\left(f(\mathbf{x}_t) - \eta\mathbf{g}_t^\top\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t + \eta^2\frac{\lambda}{2}\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t - \mathbf{g}_t + \mathbf{g}_t\right)\|^2\right) \\ &= \mathbb{E}\left(\mathbb{E}_t\left(f(\mathbf{x}_t) - \eta\mathbf{g}_t^\top\hat{\mathbf{H}}^{-1}\bar{\mathbf{g}}_t + \eta^2\frac{\lambda}{2}\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t - \mathbf{g}_t + \mathbf{g}_t\right)\|^2\right)\right). \end{aligned}$$

Here in the last inequality, $\mathbb{E}_t$ means conditional expectation, conditioned on known $\mathbf{x}_t$. Analyzing the inner expectation $\mathbb{E}_t$, we get,

$$\mathbb{E}f(\mathbf{x}_{t+1}) \leq \mathbb{E}\left(f(\mathbf{x}_t) - \eta\mathbf{g}_t^\top\hat{\mathbf{H}}^{-1}\mathbf{g}_t + \eta^2\frac{\lambda}{2}\mathbb{E}_t\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t - \mathbf{g}_t + \mathbf{g}_t\right)\|^2\right),$$

where we obtained that second term due to the unbiasedness of stochastic gradient i.e., $\mathbb{E}_t[\bar{\mathbf{g}}_t] = \mathbf{g}_t$. Furthermore using unbiasedness we have $\mathbb{E}_t\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t - \mathbf{g}_t + \mathbf{g}_t\right)\|^2 = \mathbb{E}_t\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t - \mathbf{g}_t\right)\|^2 + \|\hat{\mathbf{H}}^{-1}\mathbf{g}_t\|^2$. We get,

$$\mathbb{E}f(\mathbf{x}_{t+1}) \leq \mathbb{E}\left(f(\mathbf{x}_t) - \eta\mathbf{g}_t^\top\hat{\mathbf{H}}^{-1}\mathbf{g}_t + \eta^2\frac{\lambda}{2}\|\hat{\mathbf{H}}^{-1}\mathbf{g}_t\|^2 + \eta^2\frac{\lambda}{2}\mathbb{E}_t\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t - \mathbf{g}_t\right)\|^2\right).$$

Since $\|\hat{\mathbf{H}}^{-1}\mathbf{g}_t\|^2 \leq \left\|\hat{\mathbf{H}}^{-1}\right\| \cdot \|\hat{\mathbf{H}}^{-1/2}\mathbf{g}_t\|^2 \leq \frac{\sqrt{\alpha}}{\mu}\|\hat{\mathbf{H}}^{-1/2}\mathbf{g}_t\|^2 = \frac{\sqrt{\alpha}}{\mu}\mathbf{g}_t^\top\hat{\mathbf{H}}^{-1}\mathbf{g}_t$, where we used $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}} \tilde{\mathbf{H}}$ to write $\left\|\hat{\mathbf{H}}^{-1}\right\| \leq \frac{\sqrt{\alpha}}{\mu}$. Substituting we get,

$$\mathbb{E}f(\mathbf{x}_{t+1}) \leq \mathbb{E}\left(f(\mathbf{x}_t) - \eta\left(1 - \frac{\eta\kappa\sqrt{\alpha}}{2}\right)\mathbf{g}_t^\top\hat{\mathbf{H}}^{-1}\mathbf{g}_t + \eta^2\frac{\lambda}{2}\mathbb{E}_t\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t - \mathbf{g}_t\right)\|^2\right).$$

Again using $\hat{\mathbf{H}} \approx_{\sqrt{\alpha}} \tilde{\mathbf{H}}$, we get $\mathbf{g}_t^\top\hat{\mathbf{H}}^{-1}\mathbf{g}_t \geq \frac{1}{\lambda\sqrt{\alpha}}\|\mathbf{g}_t\|^2$. Also note that $\eta < \frac{2}{\kappa\sqrt{\alpha}}$ and hence $\left(1 - \frac{\sqrt{\alpha}\eta\kappa}{2}\right) > 0$. We get,

$$\mathbb{E}f(\mathbf{x}_{t+1}) \leq \mathbb{E}\left(f(\mathbf{x}_t) - \frac{\eta}{\lambda\sqrt{\alpha}}\left(1 - \frac{\eta\kappa\sqrt{\alpha}}{2}\right)\|\mathbf{g}_t\|^2 + \eta^2\frac{\lambda}{2}\mathbb{E}_t\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t - \mathbf{g}_t\right)\|^2\right).$$

Now use $\mathbb{E}_t\|\hat{\mathbf{H}}^{-1}\left(\bar{\mathbf{g}}_t-\mathbf{g}_t\right)\|^2 \le \frac{\alpha}{\mu^2}\mathbb{E}_t\|\bar{\mathbf{g}}_t-\mathbf{g}_t\|^2 \le \frac{8\lambda\alpha}{b\mu^2}\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)+f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right)$. Also due to μ-strong convexity of f we have $\|\mathbf{g}_t\|^2 \ge 2\mu\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)\right)$. We get,

$$\begin{aligned}\mathbb{E}f(\mathbf{x}_{t+1}) &\le \mathbb{E}\left(f(\mathbf{x}_t)-\frac{2\eta\mu}{\lambda\sqrt{\alpha}}\left(1-\frac{\eta\kappa\sqrt{\alpha}}{2}\right)\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)\right)+\eta^2\frac{4\lambda^2\alpha}{b\mu^2}\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)+f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right)\right)\\ &= \mathbb{E}\left(f(\mathbf{x}_t)-\frac{2\eta}{\kappa\sqrt{\alpha}}\left(1-\frac{\eta\kappa\sqrt{\alpha}}{2}\right)\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)\right)+\eta^2\frac{4\kappa^2\alpha}{b}\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)+f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right)\right).\end{aligned}$$

Subtract $f(\mathbf{x}^*)$ from both sides we get,

$$\begin{aligned}\mathbb{E}f(\mathbf{x}_{t+1})-f(\mathbf{x}^*) &\le \mathbb{E}\left(\underbrace{\left[1-\frac{2\eta}{\kappa\sqrt{\alpha}}+\eta^2\left(1+\frac{4\kappa^2\alpha}{b}\right)\right]}_{\xi}\cdot\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)\right)+\eta^2\frac{4\kappa^2\alpha}{b}\cdot\left(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right)\right)\\ &= \xi\cdot\mathbb{E}\left(f(\mathbf{x}_t)-f(\mathbf{x}^*)\right)+\eta^2\frac{4\kappa^2\alpha}{b}\left(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right) \qquad (26)\end{aligned}$$

We denote $\xi := 1-\frac{2\eta}{\kappa\sqrt{\alpha}}+\eta^2\left(1+\frac{4\kappa^2\alpha}{b}\right)$. Since we perform $\frac{n}{b}$ inner iterations before updating $\tilde{\mathbf{x}}$, we recursively unfold the relation (26) for $\frac{n}{b}$ times. This provides the following,

$$\begin{aligned}\mathbb{E}f(\mathbf{x}_{n/b})-f(\mathbf{x}^*) &\le \xi^{\frac{n}{b}}\left(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right)+\eta^2\frac{4\kappa^2\alpha}{b}\cdot\left(1+\xi+\xi^2+\cdots+\xi^{\frac{n}{b}-1}\right)\left(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right)\\ &= \xi^{\frac{n}{b}}\left(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right)+\eta^2\frac{4\kappa^2\alpha}{b}\cdot\left(\frac{1-\xi^{\frac{n}{b}}}{1-\xi}\right)\cdot\left(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right).\end{aligned}$$

For $\eta < \frac{b}{5\kappa^3\alpha^{3/2}}$, we have $\xi < 1-\frac{\eta}{\kappa\sqrt{\alpha}}$. Substituting upper bound for ξ we get,

$$\mathbb{E}f(\mathbf{x}_{n/b})-f(\mathbf{x}^*) \le \left[\left(1-\frac{\eta}{\kappa\sqrt{\alpha}}\right)^{n/b}+\frac{4\eta\kappa^3\alpha^{3/2}}{b}\cdot\left[1-\left(1-\frac{\eta}{\kappa\sqrt{\alpha}}\right)^{n/b}\right]\right]\cdot\left(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right).$$

Let $\eta < \frac{b}{8\kappa^3\alpha^{3/2}}$, we get,

$$\mathbb{E}f(\mathbf{x}_{n/b})-f(\mathbf{x}^*) \le \left(\frac{1}{2}\left(1-\frac{\eta}{\kappa\sqrt{\alpha}}\right)^{\frac{n}{b}}+\frac{1}{2}\right)\cdot\left(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right).$$

Now if $\left(1-\frac{\eta}{\kappa\sqrt{\alpha}}\right)^{\frac{n}{b}} < \frac{1}{2}$ we get,

$$\mathbb{E}f(\mathbf{x}_{n/b})-f(\mathbf{x}^*) < \frac{3}{4}\cdot\left(f(\tilde{\mathbf{x}})-f(\mathbf{x}^*)\right),$$

otherwise if $\left(1 - \frac{\eta}{\kappa\sqrt{\alpha}}\right)^{\frac{n}{b}} \geq \frac{1}{2}$, then we consider two cases based on the value of η. If $\eta = \frac{2}{\kappa\sqrt{\alpha}}$ then we get,

$$\mathbb{E} f(\mathbf{x}_{n/b}) - f(\mathbf{x}^*) < \left(1 - \frac{1}{\alpha\kappa^2}\right) \cdot \left(f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right),$$

and if $\eta = \frac{b}{8\kappa^3\alpha^{3/2}}$ then,

$$\mathbb{E} f(\mathbf{x}_{n/b}) - f(\mathbf{x}^*) < \left(1 - \frac{b}{16\kappa^4\alpha^2}\right) \cdot \left(f(\tilde{\mathbf{x}}) - f(\mathbf{x}^*)\right).$$

This completes the proof. ■

Appendix B. Approximate Hessian via Subsampling

Here, we reference a guarantee for the approximation factor of the sub-sampled Hessian estimate, which is a straightforward application of the matrix Bernstein's inequality.

Lemma 30 (E.g., Lemma 8 from Dereziński (2025)) *Suppose Assumption 1 holds and let* $\psi i_1, \psi_{i_2}, \cdots \psi_{i_h}$ *be i.i.d uniform samples from* $\psi_1, \psi_2, \cdots, \psi_n$. *There is an absolute constant* c *such that for any* $\mathbf{x} \in \mathbb{R}^d$, *with probability* $1 - \delta$, *the matrix*

$$\hat{\mathbf{H}} = \left(1 + \frac{\gamma}{\mu}\right)^{-1/2} \left(\frac{1}{h}\sum_{j=1}^{h} \nabla^2 \psi_{i_j}(\mathbf{x}) + \gamma \mathbf{I}\right), \quad with \;\; \gamma = \max\{12\lambda \log(2d/\delta)/h, \mu\},$$

satisfies

$$\frac{1}{\sqrt{\alpha}} \nabla^2 f(\mathbf{x}) \preceq \hat{\mathbf{H}} \preceq \sqrt{\alpha} \nabla^2 f(\mathbf{x}),$$

with $\alpha = 1 + \mathcal{O}\left(\kappa \log(d/\delta)/h + \sqrt{\kappa \log(d/\delta)/h}\right)$.

References

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Zeyuan Allen-Zhu. Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization. *arXiv preprint arXiv:1802.03866*, 2018.

Albert S Berahas, Jorge Nocedal, and Martin Takác. A multi-batch l-bfgs method for machine learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Albert S Berahas, Raghu Bollapragada, and Jorge Nocedal. An investigation of newton-sketch and subsampled newton methods. *Optimization Methods and Software*, 35(4):661–680, 2020.

Albert S Berahas, Jiahao Shi, Zihong Yi, and Baoyu Zhou. Accelerating stochastic sequential quadratic programming for equality constrained optimization using predictive variance reduction. *Computational Optimization and Applications*, 86:79—-116, 2023.

Raghu Bollapragada, Jorge Nocedal, Dheevatsa Mudigere, Hao-Jun Shi, and Ping Tak Peter Tang. A progressive batching L-BFGS method for machine learning. In *International Conference on Machine Learning*, pages 620–629. PMLR, 2018.

Raghu Bollapragada, Richard H Byrd, and Jorge Nocedal. Exact and inexact subsampled newton methods for optimization. *IMA Journal of Numerical Analysis*, 39(2):545–578, 2019.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Michał Dereziński. Stochastic variance-reduced newton: Accelerating finite-sum minimization with large batches. *Transactions on Machine Learning Research*, January 2025.

Michal Derezinski, Dhruv Mahajan, S Sathiya Keerthi, SVN Vishwanathan, and Markus Weimer. Batch-expansion training: an efficient optimization framework. In *International Conference on Artificial Intelligence and Statistics*, pages 736–744. PMLR, 2018.

Derek Driggs, Matthias J Ehrhardt, and Carola-Bibiane Schönlieb. Accelerating variance-reduced stochastic gradient methods. *Mathematical Programming*, pages 1–45, 2022.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Murat A Erdogdu and Andrea Montanari. Convergence rates of sub-sampled newton methods. *Advances in Neural Information Processing Systems*, 28, 2015.

Miria Feng, Zachary Frangella, and Mert Pilanci. Cronos: Enhancing deep learning with scalable gpu accelerated convex neural networks. *Advances in Neural Information Processing Systems*, 37:102973–103004, 2024.

Alexander Gasnikov, Pavel Dvurechensky, Eduard Gorbunov, Evgeniya Vorontsova, Daniil Selikhanovych, César A Uribe, Bo Jiang, Haoyue Wang, Shuzhong Zhang, Sébastien Bubeck, et al. Near optimal methods for minimizing convex functions with lipschitz p-th derivatives. In *Conference on Learning Theory*, pages 1392–1393. PMLR, 2019.

Alon Gonen, Francesco Orabona, and Shai Shalev-Shwartz. Solving ridge regression using sketched preconditioned svrg. In *International conference on machine learning*, pages 1397–1405. PMLR, 2016.

Robert Gower, Donald Goldfarb, and Peter Richtárik. Stochastic block bfgs: Squeezing more curvature out of data. In *International Conference on Machine Learning*, pages 1869–1878. PMLR, 2016.

Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. *Advances in Neural Information Processing Systems*, 28, 2015.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jakub Konecnỳ and Peter Richtárik. Semi-stochastic gradient descent methods. *arXiv preprint arXiv:1312.1666*, 2013.

Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pages 451–467. PMLR, 2020.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal catalyst for first-order optimization. *Advances in neural information processing systems*, 28, 2015.

Zhenwei Lin, Zikai Xiong, Dongdong Ge, and Yinyu Ye. Pdcs: A primal-dual large-scale conic programming solver with gpu enhancements. *arXiv preprint arXiv:2505.00311*, 2025.

Ji Liu, Steve Wright, Christopher Ré, Victor Bittorf, and Srikrishna Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. In *International Conference on Machine Learning*, pages 469–477. PMLR, 2014.

Yanli Liu, Fei Feng, and Wotao Yin. Acceleration of svrg and katyusha x by inexact preconditioning. In *International Conference on Machine Learning*, pages 4003–4012. PMLR, 2019.

Aurelien Lucchi, Brian McWilliams, and Thomas Hofmann. A variance reduced stochastic newton method. *arXiv preprint arXiv:1503.08316*, 2015.

Stanislav Minsker. On some extensions of bernstein's inequality for self-adjoint operators. *arXiv preprint arXiv:1112.5448*, 2011.

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

Philipp Moritz, Robert Nishihara, and Michael Jordan. A linearly-convergent stochastic l-bfgs algorithm. In *Artificial Intelligence and Statistics*, pages 249–258. PMLR, 2016.

Sen Na, Michał Dereziński, and Michael W Mahoney. Hessian averaging in stochastic newton methods achieves superlinear convergence. *Mathematical Programming*, 201(1):473–520, 2023.

Ion Necoara and Nitesh Kumar Singh. Stochastic subgradient for composite convex optimization with functional constraints. *Journal of Machine Learning Research*, 23(265):1–35, 2022.

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

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

Philippe Rigollet and Xin Tong. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 2011.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled newton methods. *Mathematical Programming*, 174(1):293–326, 2019.

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in neural information processing systems*, 25, 2012.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162:83–112, 2017.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(1), 2013.

William F Sharpe. Mean-variance analysis in portfolio choice and capital markets, 1989.

Virginia Smith, Simone Forte, Chenxin Ma, Martin Takáč, Michael I Jordan, and Martin Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18(230):1–49, 2018.

Panos Toulis and Edoardo M Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. 2017.

Joel Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Zhe Wang, Yi Zhou, Yingbin Liang, and Guanghui Lan. Stochastic variance-reduced cubic regularization for nonconvex optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2731–2740. PMLR, 2019.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Junyu Zhang, Lin Xiao, and Shuzhong Zhang. Adaptive stochastic variance reduction for subsampled newton method with cubic regularization. *INFORMS Journal on Optimization*, 4(1):45–64, 2022.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic variance-reduced cubic regularization methods. *Journal of Machine Learning Research*, 20(134):1–47, 2019.