Efficient Nonconvex Empirical Risk Minimization via Adaptive Sample Size Methods

Aryan Mokhtari
MIT

Asuman Ozdaglar
MIT

Ali Jadbabaie
MIT

Abstract

In this paper, we are interested in finding a local minimizer of an empirical risk minimization (ERM) problem where the loss associated with each sample is possibly a nonconvex function. Unlike traditional deterministic and stochastic algorithms that attempt to solve the ERM problem for the full training set, we propose an adaptive sample size scheme to reduce the overall computational complexity of finding a local minimum. To be more precise, we first find an approximate local minimum of the ERM problem corresponding to a small number of samples and use the uniform convergence theory to show that if the population risk is a Morse function, by properly increasing the size of training set the iterates generated by the proposed procedure always stay close to a local minimum of the corresponding ERM problem. Therefore, eventually the proposed procedure finds a local minimum of the ERM corresponding to the full training set which happens to also be a local minimum of the expected risk minimization problem with high probability. We formally state the conditions on the size of the initial sample set and characterize the required accuracy for obtaining an approximate local minimum to ensure that the iterates always stay in a neighborhood of a local minimum and do not get attracted to saddle points.

1 Introduction

A crucial problem in learning is the gap between the optimal solution of Statistical Risk Minimization (SRM), which is the problem that we aim to solve, and Empirical Risk Minimization (ERM), which is the problem that we can solve in practice. The goal of SRM is to come up with a classifier or learner by solving a stochastic program with respect to the distribution of the data. As data generating distribution is often unknown, one has to settle for N independent samples from the distribution to create a dataset – also called training set – and find a classifier or learner that performs well on the data. Indeed, the gap between these two solutions is a decreasing function of the number of acquired samples for the ERM problem.

Depending on whether the risk function used for evaluating the performance of a learner (classifier) is convex or not, the ERM problem boils down to a convex or nonconvex finite sum minimization problem. For the convex case, there exist various deterministic methods such gradient descent, accelerated gradient methods, quasi-Newton algorithms, and Newton's method that can be used to solve the problem [Bertsekas, 1999, Boyd and Vandenberghe, 2004, Nesterov, 2013, Wright and Nocedal, 1999]. However, each iteration of these methods requires a pass over the training set which is computationally prohibitive when N is large. Stochastic and incremental first-order and second-order methods have been deeply studied for the ERM problem when the risk is convex [Defazio et al., 2014a,b, Gürbüzbalaban et al., 2015, 2017, Le Roux et al., 2012, Mairal, 2015, Mokhtari et al., 2018, Vanli et al., 2018].

For nonconvex risk functions, deterministic [Agarwal et al., 2017, Carmon et al., 2017a,b,c, 2018, Nesterov, 2013] and stochastic [Allen Zhu and Hazan, 2016, Lei et al., 2017, Reddi et al., 2016a,b] methods can be used to reach a first-order stationary point (a critical point) of the ERM problem. Since a critical point could be a saddle point, a better convergence criteria would be to ensure convergence to a second-order stationary point of ERM. This goal can be achieved by escaping from saddle points via injecting a properly chosen noise [Daneshmand et al., 2018, Ge et al., 2015, Jin et al., 2017a,b], or using the eigenvector corresponding to the smallest eigenvalue of the Hessian to obtain an

Preliminary work. Under review by AISTATS 2019. Do not distribute.

escape direction [Agarwal et al., 2017, Allen-Zhu, 2017, Carmon et al., 2018, Paternain et al., 2017, Reddi et al., 2018, Royer and Wright, 2018, Xu and Yang, 2017]. Most of the methods that converge to a second-order stationary point are able to converge to a local minimum of the objective function when the saddle points are non-degenerate as they can escape from strict saddles.

However, most of the existing algorithms for solving the ERM problem, both in convex and nonconvex settings, do not exploit the connection between statistical and empirical risk minimization and are designed for a general finite sum minimization problem. While this is not necessarily a drawback, but it is nonetheless true that not exploiting the connection between SRM and ERM may leave some performance gains on the tab. A new line of research which is based on solving a sequence of ERM problems with geometrically increasing samples attempted to collect these gains exhaustively [Daneshmand et al., 2016, Mokhtari and Ribeiro, 2017, Mokhtari et al., 2016]. The main idea of adaptive sample size methods is to first obtain a good solution for an ERM problem corresponding to a small subset of the training set, which is computationally cheaper to solve. This is followed by increasing the size of the working samples set by adding new samples to the current set and using the most recent solution as a warm start for the new ERM problem. The key idea is that since the samples are drawn from the same distribution the solution for a smaller set should be a good approximate for the solution of the enlarged set (containing the smaller set). In the convex setting, the sequence of ERM problems are convex and one can solve them arbitrary close to their global minimum. In fact, recent works showed that for the convex case, if we use first-order [Mokhtari and Ribeiro, 2017] or second-order [Mokhtari et al., 2016] methods to solve the subproblem the overall complexity significantly reduces compared to solving the ERM problem for the full training set using deterministic or stochastic methods.

For the nonconvex case, it is typically hard to reach a global minimizer and the best that one can hope is converging to a local minimum under the assumption that the stationary points are non-degenerate. In this paper, our goal is to exploit the connection between ERM and SRM via adaptive sample size schemes to improve the overall computational complexity for reaching an approximate local minimum of ERM and consequently SRM. In our proposed approach, we first find an approximate local minimum of the ERM problem corresponding to a small number of samples. Then, based on the uniform convergence theory, we show that if the population risk is a Morse function, i.e., its saddles are non-degenerate, by properly increasing the size of training set the approximate local minimum for the smaller problem is within a neighborhood of the local minimum of the ERM problem corresponding to the enlarged set. We further show that by following simple gradient or Newton steps the sequence of iterates approaches a local minimum of the new ERM problem without getting attracted to any saddle points or local maxima. We formally characterize how accurate the subproblems should be solved to ensure that the iterates always stay in a local neighborhood of a local minimum and do not get attracted to saddle points. By following this scheme and doubling the size of the training set at the end of each stage we finally reach a local minimum of the ERM problem for the full training set, which is an approximate local minimum of the SRM problem with high probability.

To better highlight the advantage of the proposed adaptive sample size methods in nonconvex settings, note that the main challenge in converging to a second-order stationary point or to a local minimum when saddles are non-degenerate is escaping from saddle points. To do so, these algorithms typically require computation of the objective function Hessian which has a computational complexity of $O(Nd^2)$, where d is the dimension of the problem, as well as a direction that leads to descent direction by computing an approximate of the eigenvector corresponding to the minimum eigenvalue of the Hessian. This process can be computationally expensive as the number of saddle points visited before reaching to a neighborhood of a local minimum could be very large. The adaptive sample size allows us to do this only for an ERM problem with a small number of samples, and then stay within a neighborhood of a sequence of local minima as we enlarge the size of the training set. This procedure, indeed, leads to a significantly lower complexity for reaching a local minimum of the ERM problem, if the extra cost for staying close to local minimum while we increase the size of the training set is negligible. In particular, we show that, given an approximate local minimum of the initial training set, the proposed adaptive sample size approach with an accelerated gradient descent update reaches a local minimum of the full training set after at most $\mathcal{O}(N\sqrt{\kappa})$ gradient evaluations, where κ can be interpreted as the condition number of the population risk at its critical points. Moreover, if we have access to second order information, the proposed scheme obtains a local minimum of the ERM problem after at most $2N$ gradient and Hessian evaluations and $\log N$ Hessian inverse computations.

Outline. We start the paper by reviewing the problem formulations for statistical risk minimization and empirical risk minimization (ERM) as well as recapping the uniform convergence results for non-convex loss

functions (Section 2). Then, we describe the details of the proposed adaptive sample size scheme for obtaining an approximate local minimum of the ERM problem (Section 3). Theoretical convergence guarantees for our proposed framework is then presented (Section 4). In particular, we characterize the overall computation cost of running the algorithm until reaching a local minimum of the ERM problem when we update the iterates in subproblems according to gradient descent, Nesterov's accelerated gradient, or Newton's method. We also compare the theoretical convergence guarantees for the proposed adaptive sample size scheme with state-of-the-art algorithms (Section 5). We finally close the paper with concluding remarks (Section 6).

Notation. Vectors are written as lowercase $\mathbf{x} \in \mathbb{R}^p$ and matrices as uppercase $\mathbf{A} \in \mathbb{R}^{p\times p}$. We use $\|\mathbf{x}\|$ to denote the l_2 norm of the vector $\mathbf{x}$. Given a matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, we denote by $\lambda_i(\mathbf{A})$ its i-th largest eigenvalue, and by $\|\mathbf{A}\|_{op}$ its operator norm which is defined as $\|\mathbf{A}\|_{op} := \max\{\lambda_1(\mathbf{A}), -\lambda_p(\mathbf{A})\}$ where λ_1 and λ_p are the largest and smallest eigenvalues of $\mathbf{A}$, respectively. The inner product of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ is denoted by $\langle \mathbf{x}, \mathbf{y}\rangle := \sum_{i=1}^p x_i y_i$. Given a function f its gradient and Hessian at point $\mathbf{x}$ are denoted as $\nabla f(\mathbf{x})$ and $\nabla^2 f(\mathbf{x})$, respectively. We use $B^p(r) = \{\mathbf{x} \in \mathbb{R}^p \mid \|\mathbf{x}\|_2 \leq r\}$ to denote the Euclidean ball with radius r in p dimensions.

2 Preliminaries

Consider a decision vector $\mathbf{w} \in \mathbb{R}^p$, a random variable $\mathbf{Z}$ with realizations $\mathbf{z} \in \mathbb{R}^d$ and a loss function $\ell : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$. We aim to solve

$$\min_{\mathbf{w}} R(\mathbf{w}) = \min_{\mathbf{w}} \mathbb{E}_{\mathbf{Z}}[\ell(\mathbf{w}, \mathbf{Z})] = \min_{\mathbf{w}} \int_{\mathbf{Z}} \ell(\mathbf{w}, \mathbf{Z}) P(d\mathbf{z}), \tag{1}$$

where $R(\mathbf{w}) := \mathbb{E}_{\mathbf{Z}}[\ell(\mathbf{w}, \mathbf{Z})]$ is defined as the statistical risk, and P is the probability distribution of the random variable $\mathbf{Z}$. In the rest of the paper, we also refer to R as the expected risk or the population risk. Note that the loss function ℓ is not necessarily convex with respect to $\mathbf{w}$ and could be nonconvex. Even under the assumption that the loss function ℓ is convex, the optimization problem in (1) cannot be solved accurately since the distribution P is unknown. However, in most learning problems we have access to a training set $\mathcal{T} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ containing N independent samples $\mathbf{z}_1, \ldots, \mathbf{z}_N$ drawn from P. Therefore, we attempt to minimize the empirical risk associated with the training set $\mathcal{T} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, which is equivalent to solving the following optimization program

$$\min_{\mathbf{w}} R_n(\mathbf{w}) = \min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, \mathbf{z}_i), \tag{2}$$

for $n = N$. Note that in (2) we defined $R_n(\mathbf{w}) := (1/n)\sum_{i=1}^n \ell(\mathbf{w}, \mathbf{z}_i)$ as the empirical risk corresponding to the realizations $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$.

For the case that the loss function ℓ is convex, there is a rich literature on bounds for the difference between the expected risk R and the empirical risk R_n which is also referred to as *estimation error* [Bartlett et al., 2006, Bottou, 2010, Bottou and Bousquet, 2007, Frostig et al., 2015, Vapnik, 2013]. In particular, it has been shown that if the population risk R is convex, then with high probability for a sufficiently large number of samples n the gap between the expected risk and empirical risk is bounded by

$$\sup_{\mathbf{w}\in\mathbb{R}^p} |R(\mathbf{w}) - R_n(\mathbf{w})| \leq \mathcal{O}\left(n^{-\alpha}\right), \tag{3}$$

where α can be a constant in the interval $[0.5, 1]$ depending on the regularity conditions that the loss function ℓ satisfies [Bartlett et al., 2006, Vapnik, 2013].

In a recent paper, Mei et al. [2018] extended this result to the nonconvex setting under the assumptions that gradients and Hessians of the population risk R satisfy some regularity conditions (we state them formally below). To simplify the analysis they focused on the problem in which the decision variable $\mathbf{w}$ belongs to a bounded set and they assumed that the bounded set is large enough to contain all the stationary points. To formally state their result, we first present their assumptions.

Assumption 1. *The loss function gradient $\nabla\ell$ is τ^2-sub-Gaussian, i.e., for any $\mathbf{y} \in \mathbb{R}^p$ and $\mathbf{w} \in B^p(r)$*

$$\mathbb{E}[\exp\left(\langle \mathbf{y}, \nabla\ell(\mathbf{w}, \mathbf{Z}) - \mathbb{E}[\nabla\ell(\mathbf{w}, \mathbf{Z})]\rangle\right)] \leq \exp\left[\frac{\tau^2\|\mathbf{y}\|^2}{2}\right]. \tag{4}$$

Assumption 2. *The Hessian $\nabla^2\ell$ for the vectors in the unit sphere is τ^2-sub-exponential, i.e., for any $\mathbf{y} \in B^p(1)$ and $\mathbf{w} \in B^p(r)$ we have*

$$\mathbb{E}\left[\exp\left[\frac{1}{\tau^2}\left|\langle \mathbf{y}, \nabla^2\ell(\mathbf{w}, \mathbf{Z})\mathbf{y}\rangle - \langle \mathbf{y}, \mathbb{E}[\nabla^2\ell(\mathbf{w}, \mathbf{Z})]\mathbf{y}\rangle\right|\right]\right] \leq 2. \tag{5}$$

Assumption 3. *The gradients $\nabla\ell$ and Hessians $\nabla^2\ell$ are Lipschitz continuous with constants $M = \tau^2 p^q$, and $L = \tau^3 p^q$, respectively, for some constant q.*

Note that the stochastic gradient $\nabla\ell(\mathbf{w}, \mathbf{Z})$ and stochastic Hessian $\nabla^2\ell(\mathbf{w}, \mathbf{Z})$ are unbiased estimators of the risk gradient $\mathbb{E}[\nabla\ell(\mathbf{w}, \mathbf{Z})]$ and Hessian $\mathbb{E}[\nabla^2\ell(\mathbf{w}, \mathbf{Z})]$, and, therefore, conditions in Assumptions 1 and 2 can be considered as bounds on the variations of these estimators. Note that the constants for Lipschitz continuity of gradients and Hessians are defined properly as a function of τ and p in a way that $r\tau$ and q are dimensionless.

The uniform convergence result for the nonconvex ERM problem is stated in the following theorem.

Theorem 1 (Mei et al. [2018]). *Under Assumptions 1-3, there exists a universal constant C_0, such that letting $C = C_0(\max\{q, \log(r\tau/\delta), 1\})$, for $n \geq Cp\log p$ with probability at least $1-\delta$ the following inequalities hold:*

$$\sup_{\mathbf{w}\in B_r^p} \|\nabla R_n(\mathbf{w}) - \nabla R(\mathbf{w})\|_2 \leq \tau\sqrt{\frac{Cp\log n}{n}},$$
$$\sup_{\mathbf{w}\in B_r^p} \left\|\nabla^2 R_n(\mathbf{w}) - \nabla^2 R(\mathbf{w})\right\|_{op} \leq \tau^2\sqrt{\frac{Cp\log n}{n}}. \quad (6)$$

Theorem 1 shows that the difference between gradients and Hessians of the population and empirical risks are within $\mathcal{O}(\sqrt{p\log n/n})$ of each other with high probability. Hence, the landscape of stationary points for the empirical risk ∇R_n is similar to the one for the expected risk $\nabla R(\mathbf{w})$ if the number of samples n is sufficiently large. Moreover, it immediately follows from this result that there is no gain in finding a local minimum for the risk R_n that has a gradient smaller than $\mathcal{O}(\sqrt{p\log n/n})$. Additionally, to ensure that the eigenvalues of the population risk Hessian $\nabla^2 R(\mathbf{w}^\dagger)$ are strictly positive definite at a point $\mathbf{w}^\dagger$, we need to ensure that all the eigenvalues of the empirical risk Hessian $\nabla^2 R_n(\mathbf{w}^\dagger)$ are larger than the statistical gap $\mathcal{O}(\sqrt{p\log n/n})$. Based on these observations, given a training set of size N, our goal is to find an approximate local minimum $\mathbf{w}_N$ of the risk R_N satisfying the following conditions

$$\|\nabla R_N(\mathbf{w}_N)\| \leq \zeta_N, \quad \nabla^2 R_N(\mathbf{w}_N) \succeq \gamma_N \mathbf{I}, \quad (7)$$

where $\zeta_N = \mathcal{O}(\tau\sqrt{Cp\log N/N})$ and $\gamma_N = \mathcal{O}(\tau^2\sqrt{Cp\log N/N})$. Indeed, if we find a point $\mathbf{w}_N$ satisfying the conditions in (7), based on the result in Theorem 1, $\mathbf{w}_N$ is an approximate local minimum of the population risk R with high probability.

To derive our theoretical results, besides the conditions in Assumptions 1-3, we need to assume that the population risk R is strongly Morse as we formally define in the following assumption.

Assumption 4. *The population risk R is (α, β)-strongly Morse if for any point $\mathbf{w}^\dagger$ that satisfies the condition $\|\nabla R(\mathbf{w}^\dagger)\| \leq \alpha$, it holds that $|\lambda_i(\nabla^2 R(\mathbf{w}^\dagger))| \geq \beta$ for all $i \in \{1, \ldots, p\}$.*

The definition of (α, β)-strongly Morse functions is borrowed from [Mei et al., 2018]. Note that in this definition α and β are positive constants. This condition ensures that all the critical points of the population risk R are non-degenerate and in a neighborhood of each of them the absolute value of the eigenvalues of the Hessian $\nabla^2 R$ are strictly larger than 0. Note that the (α, β)-strongly Morse condition can be relaxed to the assumption that all the critical point are nondegenerate, i.e., if $\mathbf{w}^\dagger$ is a critical point then $|\lambda_i(\nabla^2 R(\mathbf{w}^\dagger))| \geq \beta'$. Indeed, this condition in conjunction with Lipschitz continuity of gradients and Hessians implies that there exist α and β such that the condition in Assumption 4 holds.

3 An Adaptive Sample Size Scheme for Nonconvex Problems

In this section, we aim to design an adaptive sample size mechanism which builds on the uniform convergence theory for ERM problems to find a local minimum of the empirical risk R_N upto its statistical accuracy faster than traditional (stochastic and deterministic) methods. The main steps of the proposed scheme can be explained as follows. We first find a local minimizer of the ERM problem corresponding to m_0 samples which is significantly smaller than N. Then, we increase the size of the training set such that the current iterate, which is an approximate local min for ERM with m_0 samples, stays in a neighborhood of a local minimum of the ERM problem corresponding to the enlarged training set. Our theory suggests that by doubling the size of training set this condition is satisfied. After adding more samples to the active training set, we update the iterate according to a first-order or second-order method until the norm of gradient becomes sufficiently small and the iterate becomes very close to a local minimum of the enlarged ERM problem. This procedure continues until the training set becomes identical to the full training set $\mathcal{T}$ which contains N samples. At the end of procedure, the output is a point $\mathbf{w}_N$ which is close to a local minimum of R_N.

To be more specific, consider the training set $\mathcal{S}_m$ with m samples as a subset of the full training set $\mathcal{T}$, i.e., $\mathcal{S}_m \subset \mathcal{T}$. Assume that we found a point $\mathbf{w}_m$ which is close to one of the local minimizers of the risk R_m, i.e., $\mathbf{w}_m$ satisfies $\|\nabla R_m(\mathbf{w}_m)\| \leq \epsilon_m$ and $\nabla^2 R_m(\mathbf{w}_m) \succ \mathbf{0}$ for some positive constant ϵ_m. The fundamental question at hand is under what conditions on ϵ_m and the initial size of the training set m_0 we can ensure that the iterate $\mathbf{w}_m$ is within a neighborhood of a local minimum of the ERM problem corresponding to a larger set $\mathcal{S}_n$ which has $n = 2m$ samples and contains the previous set, i.e., $\mathcal{S}_m \subset \mathcal{S}_n \subset \mathcal{T}$. We formally answer this question in the following section. We further derive an upper bound on the overall computational complexity for reaching a local minimum of the ERM problem corresponding to the full training set $\mathcal{T}$ for different choices of iterative methods used to solve the subproblems.

The steps of the proposed adaptive sample size scheme are summarized in Algorithm 1. We assume that for

Algorithm 1 Adaptive Sample Size Mechanism

1: **Input:** Initial sample size $n = m_0$ and argument $\mathbf{w}_n = \mathbf{w}_{m_0}$
2: **while** $n \leq N$ **do** {main loop}
3: Set $\mathbf{w}_m \leftarrow \mathbf{w}_n$ and $m \leftarrow n$.
4: Increase sample size: $n \leftarrow \min\{2m, N\}$.
5: Set the initial variable: $\tilde{\mathbf{w}} \leftarrow \mathbf{w}_m$.
6: **while** $\|\nabla R_n(\tilde{\mathbf{w}})\| > \epsilon_n$ **do**
7: $\tilde{\mathbf{w}} \leftarrow$ FO-update($\tilde{\mathbf{w}}$,$\nabla R_n(\tilde{\mathbf{w}})$) or $\tilde{\mathbf{w}} \leftarrow$ SO-update($\tilde{\mathbf{w}}$,$\nabla R_n(\tilde{\mathbf{w}})$,$\nabla^2 R_n(\tilde{\mathbf{w}})$)
8: **end while**
9: Set $\mathbf{w}_n \leftarrow \tilde{\mathbf{w}}$.
10: **end while**

the initial training set with m_0 samples, we have access to a point $\mathbf{w}_{m_0}$ which is close (we formalize the measure of closeness later) to one of the local minima of the risk R_{m_0}. Note that in steps 5-8 we use the iterate $\mathbf{w}_m$, which is an approximate local minimum for R_m, as the initial point and update it by following a first-order update (FO-update), e.g., gradient descent, or a second-order update (SO-update), e.g., Newton's method, until we reach a point that satisfies the stop condition in step 6. Then, the output $\mathbf{w}_n$ is an approximate local minimum for the ERM problem with $n = 2m$ samples. This process continuous until we reach the full training set $n = N$. The parameter ϵ_n used in step 6 depends on the choice of descent algorithm that we use for updating the iterates in step 7. In the following section, we formally state how ϵ_n should be chosen.

4 Main Result

In this section, we study the overall computational complexity of the adaptive sample size scheme outlined in Algorithm 1 to reach a local minimum of the ERM problem in (2). We study different cases where we use the gradient descent algorithm, accelerated gradient descent method, or Newton's method to solve the subproblems at each stage. Although in this section we focus on the complexity analysis of these three methods only, other deterministic algorithms, e.g., quasi-Newton methods, and stochastic methods, e.g., SVRG, SAG, SAGA, can also be used to update the iterates in Step 7 of Algorithm 1.

We first use the result of uniform convergence theorem (Theorem 1) as well as a crucial property of the proposed adaptive sample size scheme that the enlarged training set $\mathcal{S}_n$ at each stage is a superset of the previous set $\mathcal{S}_m$, i.e., $\mathcal{S}_n \supset \mathcal{S}_m$ $(n > m)$, to show that the gap between gradients and Hessians of the risks R_m and R_n is proportional to $\frac{n-m}{n}$.

Proposition 1. *Consider the sets $\mathcal{S}_m$ and $\mathcal{S}_n$ as subsets of the training set $\mathcal{T}$ such that $\mathcal{S}_m \subset \mathcal{S}_n \subset \mathcal{T}$, where the number of samples in the sets $\mathcal{S}_m$ and $\mathcal{S}_n$ are m and n, respectively. Furthermore, recall the definition of C in Theorem 1. If Assumptions 1-3 hold and $\min\{m, n-m\} \geq Cp \log p$, then with probability at least $1 - 2\delta$ the gradient variation is bounded by*

$$\sup_{\mathbf{w} \in B_r^p} \|\nabla R_n(\mathbf{w}) - \nabla R_m(\mathbf{w})\|_2 \leq \frac{n-m}{n} \tau \sqrt{Cp} \left(\sqrt{\frac{\log(n-m)}{n-m}} + \sqrt{\frac{\log m}{m}} \right), \quad (8)$$

and the Hessian variation is bounded by

$$\sup_{\mathbf{w} \in B_r^p} \left\|\nabla^2 R_n(\mathbf{w}) - \nabla^2 R_m(\mathbf{w})\right\|_{op} \leq \frac{n-m}{n} \tau^2 \sqrt{Cp} \left(\sqrt{\frac{\log(n-m)}{n-m}} + \sqrt{\frac{\log m}{m}} \right). \quad (9)$$

The result in Proposition 1 establishes an upper bound on the difference between gradients and Hessians of the risk functions R_m and R_n corresponding to the sample sets $\mathcal{S}_m$ and $\mathcal{S}_n$, respectively, when $\mathcal{S}_m \subset \mathcal{S}_n$. As one would expect, the gap is proportional to the difference between the number of samples in the sample sets, i.e., $n - m$. We would like to highlight that these results only hold if the larger set $\mathcal{S}_n$ contains the smaller set $\mathcal{S}_m$, and for general subsets of the full training set $\mathcal{C}$ these results may not hold.

In the following proposition, we use the uniform convergence theorem to show that, when the number of samples n is sufficiently large, the corresponding empirical risk R_n is strongly Morse if the population risk is strongly Morse.

Proposition 2. *Suppose the conditions in Assumptions 1-3 are satisfied. Furthermore, recall the definition of C in Theorem 1. If the number of samples n satisfies the condition*

$$n \geq \log n \max\left\{ \frac{Cp\tau^2}{\alpha^2}, \frac{Cp\tau^4}{\beta^2} \right\}, \quad (10)$$

then the empirical risk R_n corresponding to the set of realizations $\mathcal{S}_n$ is (α_n, β_n)-strongly Morse with probability at least $1 - \delta$, where

$$\alpha_n = \alpha - \tau \sqrt{\frac{Cp \log n}{n}}, \quad \beta_n = \beta - \tau^2 \sqrt{\frac{Cp \log n}{n}}. \quad (11)$$

The above immediately follows from the uniform convergence result in Theorem 1 and the assumption that the population risk R is (α, β)-strongly Morse. Indeed, the result is meaningful when the constants α_n and β_n are strictly positive which requires the number of samples n to be larger than the threshold stated in Proposition 2. In the following subsections, we formally state our theoretical results.

4.1 Gradient descent algorithm

We first state the result for the case that gradient descent method is used in the proposed adaptive sample size scheme. To be more precise, consider $\mathbf{w}_m$ as an approximate local minimum of R_m. We focus on the case that in step 7 of Algorithm 1, we update the iterates using the gradient descent (GD) algorithm. If we initialize the sequence $\tilde{\mathbf{w}}$ as $\tilde{\mathbf{w}}^0 = \mathbf{w}_m$, the approximate local minimum $\mathbf{w}_n$ for the risk R_n is the outcome of the update

$$\tilde{\mathbf{w}}^{k+1} = \tilde{\mathbf{w}}^k - \eta_n \nabla R_n(\tilde{\mathbf{w}}^k) \tag{12}$$

after s_n iterations, i.e., $\mathbf{w}_n = \tilde{\mathbf{w}}^{s_n}$, where η is a properly chosen stepsize. The parameter η_n is indexed by n since it depends on the number of samples.

In the following theorem, we explicitly express the required condition on the accuracy ϵ_n at each stage and characterize an upper bound on the number of gradient iterations s_n at each stage. Using these results we derive an upper bound on the overall computational complexity of the algorithm when the iterates are updated by GD.

Theorem 2. *Consider the adaptive sample size method outlined in Algorithm 1. Suppose Assumptions 1-3 hold, and recall the definition of C in Theorem 1. Let $\mathcal{S}_{m_0}$ be the initial set with m_0 samples such that*

$$m_0 \geq Cp \log p, \quad \frac{m_0}{\log m_0} \geq \max\left\{\frac{9Cp\tau^2}{\alpha^2}, \frac{4Cp\tau^4}{\beta^2}\right\}. \tag{13}$$

Assume that we have access to an approximate local minimum $\mathbf{w}_{m_0}$ of the initial ERM problem with cost R_{m_0} satisfying the conditions $\|\nabla R_{m_0}(\mathbf{w}_{m_0})\| \leq \epsilon_{m_0}$ and $\nabla^2 R_{m_0}(\mathbf{w}_{m_0}) \succ \mathbf{0}$, where ϵ_n for any positive integer n is defined as $\epsilon_n := \tau\sqrt{\frac{Cp\log n}{n}}$. If at each stage of the adaptive sample size scheme we use the update of gradient descent with the stepsize $\eta_n = \min\left\{\frac{\beta_n}{\alpha_n L}, \frac{2}{\beta_n + M}\right\}$ to reach a point satisfying $\|\nabla R_n(\mathbf{w}_n)\| \leq \epsilon_n$, then with high probability the total number of gradient evaluations to reach a local minimum of the full training set $\mathcal{T}$ satisfying $\|\nabla R_N(\mathbf{w}_N)\| \leq \epsilon_N$ and $\nabla^2 R_N(\mathbf{w}_N) \succ \beta_N \mathbf{I}$ is at most

$$2N \log 4 \ \max\left\{\frac{8\alpha L}{\beta^2}, 1 + \frac{2M}{\beta}\right\}. \tag{14}$$

The result in Theorem 2 shows that after evaluating $\mathcal{O}(N\frac{M}{\beta} + N\frac{\alpha L}{\beta^2})$ gradients or equivalently after operating on $\mathcal{O}(N\frac{M}{\beta} + N\frac{\alpha L}{\beta^2})$ sample points we reach a local minimizer of the ERM corresponding to the full training set if we start from a local minimizer of the ERM associated with a small subset of data points. Note that the condition on ϵ_n ensures that we solve each subproblem up to its statistical accuracy.

The proof of Theorem 2 can be divided into three main steps. First, we show that the variable $\mathbf{w}_m$ is within a local neighborhood of a local minimum of R_n. Second, we prove that if the iterate $\mathbf{w}_m$ is in a local neighborhood of a local minimum of R_n by following the gradient update the iterates always stay close to the local minimum and do not get attracted to the saddle points of R_n. Third, we derive an upper bound on the number of GD steps s_n that should be run at each stage s_n which indeed depends on the required accuracy ϵ_n. By combining these steps we can show that the output of the algorithm is an approximate local minimum of R_N and the overall number of gradient evaluations or processed samples is bounded above by the expression in (14). These points are described in detail in the proof of Theorem 2 which is available in the supplementary material.

4.2 Accelerated gradient descent algorithm

In this section, we study the theoretical guarantees for the proposed adaptive sample size mechanism when the accelerated gradient descent (AGD) method is used for updating the iterates. In particular, If we initialize the sequences $\tilde{\mathbf{w}}$ as $\tilde{\mathbf{w}}^0 = \tilde{\mathbf{y}}^0 = \mathbf{w}_m$, where $\mathbf{w}_m$ is an approximate local minimizer of R_m, the approximate local minimum $\mathbf{w}_n$ for the risk R_n is the outcome of the updates

$$\tilde{\mathbf{w}}^{k+1} = \tilde{\mathbf{y}}^k - \eta_n \nabla R_n(\tilde{\mathbf{y}}^k) \tag{15}$$

$$\tilde{\mathbf{y}}^{k+1} = \tilde{\mathbf{w}}^{k+1} + \xi_n(\tilde{\mathbf{w}}^{k+1} - \tilde{\mathbf{w}}^k) \tag{16}$$

after s_n steps, i.e., $\mathbf{w}_n = \tilde{\mathbf{w}}^{s_n}$. The parameters η_n and ξ_n will be formally defined in the following theorem where we state our results for the case that AGD is used for updating the iterates.

Theorem 3. *Consider the adaptive sample size method outlined in Algorithm 1. Suppose Assumptions 1-3 hold, and recall the definition of C in Theorem 1. Let $\mathcal{S}_{m_0}$ be the initial set with m_0 samples such that*

$$m_0 \geq Cp \log p, \ \frac{m_0}{\log m_0} \geq \max\left\{\frac{\rho Cp\tau^2}{\alpha^2}, \frac{4Cp\tau^4}{\beta^2}\right\}. \tag{17}$$

where $\rho = (1 + 4\sqrt{M/\beta})^2$. Assume that we have access to an approximate local minimum $\mathbf{w}_{m_0}$ of the initial ERM problem with cost R_{m_0} satisfying the conditions $\|\nabla R_{m_0}(\mathbf{w}_{m_0})\| \leq \epsilon_{m_0}$ and $\nabla^2 R_{m_0}(\mathbf{w}_{m_0}) \succ \mathbf{0}$, where ϵ_n for any positive integer n is defined as $\epsilon_n := \tau\sqrt{\frac{Cp\log n}{n}}$. If at each stage of the adaptive sample size scheme we use the update of accelerated gradient descent with the parameters $\eta_n = 1/M$ and $\xi_n = (\sqrt{M} - \sqrt{\beta_n})/(\sqrt{M} + \sqrt{\beta_n})$ to reach a point satisfying $\|\nabla R_n(\mathbf{w}_n)\| \leq \epsilon_n$,

then with high probability the total number of gradient evaluations to reach a local minimum of the full training set $\mathcal{T}$ satisfying the conditions $\|\nabla R_N(\mathbf{w}_N)\| \leq \epsilon_N$ and $\nabla^2 R_N(\mathbf{w}_N) \succ \beta_N \mathbf{I}$ is at most

$$2N\sqrt{\frac{2M}{\beta}}\log\left(\frac{16M}{\beta}\right). \tag{18}$$

The result in Theorem 3 shows that the total number of gradient evaluations when we use AGD to update the iterates is of $\mathcal{O}(N\sqrt{M/\beta})$, which is better than the result for adaptive sample size GD in (14). Although the overall computational complexity of adaptive sample size AGD is lower than the one for GD, the condition on the initial size of the training set m_0 for AGD is stronger than that of GD. To be more precise, as the ratio M/β is larger than 1, it can be verified that the factor ρ in (17) is larger than 25, and, therefore, the lower bound on the size of the initial training set in (17) is larger the lower bound for GD in (13).

4.3 Newton's method

We proceed by studying the case that Newton's update is used to solve each subproblem up to a point that the norm of gradient is small enough and the Hessian stays positive definite. In particular, we show that a single iteration of the Newton's method with stepsize $\eta = 1$ is sufficient to move from $\mathbf{w}_m$ to $\mathbf{w}_n$ which are approximate local minimums of R_m and R_n, respectively. In other words, given $\mathbf{w}_m$ which is an approximate local minimum of R_m satisfying $\nabla R_m(\mathbf{w}_m) \leq \epsilon_m$ and $\nabla^2 R_m(\mathbf{w}_m) \succ 0$, we obtain the next approximate local minimum using the update

$$\mathbf{w}_n = \mathbf{w}_m - \nabla^2 R_n(\mathbf{w}_m)^{-1}\nabla R_n(\mathbf{w}_m). \tag{19}$$

Hence, for the case that we use Newton's step the number of intermediate iterations is only $s_n = 1$ at each stage. In the following theorem, we formally state the conditions on m_0 and ϵ_n when Newton's update used in the adaptive sample size scheme and characterize the overall number of gradient and Hessian evaluations for obtaining an approximate local minimum of R_N.

Theorem 4. *Consider the adaptive sample size mechanism outlined in Algorithm 1. Suppose the conditions in Assumptions 1-3 are satisfied. Further, recall the definition of C in Theorem 1. Let $\mathcal{S}_{m_0}$ be the initial set with m_0 samples such that $m_0 \geq Cp\log p$ and*

$$\frac{m_0}{\log m_0} \geq \max\left\{\frac{9Cp\tau^2}{\alpha^2}, \frac{4Cp\tau^4}{\beta^2}, \frac{4Cp(\tau^2+2L\tau)^2}{\beta^4}\right\}. \tag{20}$$

Assume that we have access to a point $\mathbf{w}_{m_0}$ satisfying the conditions $\|\nabla R_{m_0}(\mathbf{w}_{m_0})\| \leq \epsilon_{m_0}$ and $\nabla^2 R_{m_0}(\mathbf{w}_{m_0}) \succ \mathbf{0}$, where $\epsilon_n := \tau\sqrt{\frac{Cp\log n}{n}}$. Then by running a single iteration of Newton's method with stepsize 1 at each stage, we reach a local minimum of the full training set $\mathcal{T}$ satisfying $\|\nabla R_N(\mathbf{w}_N)\| \leq \epsilon_N$ and $\nabla^2 R_N(\mathbf{w}_N) \succ \beta_N \mathbf{I}$ after computing at most $2N$ gradients and Hessians and $\log_2(N/m_0)$ matrix inversions.

The result in Theorem 4 shows that the total number of gradient and Hessian evaluations for the adaptive sample size scheme with Newton's method is at most $2N$ which is independent of the problem parameters including gradients Lipschitz continuity parameter M and the Morse condition parameters α and β. However, this come at the cost of evaluating $\log_2(N/m_0)$ matrix inversions.

Remark 1. *Note that the result at each stage holds with probability $1-2\delta$. Since the number of times that we double the size of training set is $\log(N/m_0)$, the final result holds with a high probability of at least $(1-2\delta)^{\log(N/m_0)}$. This observation implies that the results in Theorems 2-4 hold with probability at least $1-\delta'$ if we set $C = C_0(\max\{q, \log(r\tau/(2\delta'\log(N/m_0))), 1\})$.*

Remark 2. *One may have the concern that the implementation of the proposed scheme requires access to the constant C. However, we would like to highlight that our theoretical results hold if the constant C used for the choice of m_0 is larger than the one defined in Theorem 1. Therefore, the implementation of the algorithm only requires access to an upper bound on the parameter C defined in Theorem 1.*

The results in Theorems 2-4 guarantee that the output of the adaptive sample size procedure $\mathbf{w}_N$ is such that the gradient corresponding to the risk of the full training set is small, i.e., $\|\nabla R_N(\mathbf{w}_N)\| \leq \tau\sqrt{\frac{Cp\log N}{N}}$, and the Hessian of the risk is strictly positive definite, i.e., $\nabla^2 R_N(\mathbf{w}_N) \succ \beta_N \mathbf{I}$. By using the result in Theorem 1 it can be further shown that with high probability the point $\mathbf{w}_N$ is also close to a local minimum of the expected risk R with high probability.

5 Complexity Comparison

In this section, our goal is to highlight the advantage of our proposed adaptive sample size scheme over state-of-the-art methods for obtaining a local minimum when the objective function is Morse. In particular, we focus on the Accelerated-Nonconvex-Method (ANM) proposed in [Carmon et al., 2018] where the authors show that if we assume saddle points of the empirical risk R_N are σ-strict, then it is possible to find a point $\mathbf{w}$ satisfying $\|\nabla R_N(\mathbf{w})\| \leq \mathcal{O}(1/\sqrt{N})$ and $\nabla^2 R_N(\mathbf{w}) \succeq \sigma\mathbf{I}$

after at most

$$\mathcal{O}\left(\frac{M^{1/2}L^2\log(\sigma^{-2})}{\sigma^{7/2}}+\frac{M^{1/2}}{\sigma^{1/2}}\log\left(\frac{\sqrt{N}}{\sigma}\right)\right) \quad (21)$$

gradient evaluations or Hessian-vector product computations. Note that strongly Morse functions are a subclass of functions with strict saddle and therefore their result also holds when the function R_N is (σ^2,σ)-strongly Morse with $\sigma=\min\{\sqrt{\alpha_N},\beta_N\}$. Since each gradient or Hessian evaluation for the full objective function requires access to the full training set, the number of overall gradient or Hessian evaluations is

$$\mathcal{O}\left(\frac{NM^{1/2}L^2\log(\sigma^{-2})}{\sigma^{7/2}}+\frac{NM^{1/2}}{\sigma^{1/2}}\log\left(\frac{\sqrt{N}}{\sigma}\right)\right) \quad (22)$$

On the other hand, our proposed adaptive sample size scheme requires access to an approximate local minimum of the initial risk function R_{m_0} which corresponds to a set with m_0 samples where possibly $m_0 << N$. Since the function R_{m_0} is $(\alpha_{m_0},\beta_{m_0})$ strongly Morse, by defining $\sigma_0:=\min\{\sqrt{\alpha_{m_0}},\beta_{m_0}\}$, it can be shown that using ANM one can find a point $\mathbf{w}_{m_0}$ satisfying the conditions $\|\nabla R_{m_0}(\mathbf{w}_{m_0})\|\leq\epsilon_{m_0}$ and $\nabla^2 R_{m_0}(\mathbf{w}_{m_0})\succeq\mathbf{0}$ after at most

$$\mathcal{O}\left(\frac{m_0M^{1/2}L^2\log(\sigma_0^{-2})}{\sigma_0^{7/2}}+\frac{m_0M^{1/2}}{\sigma_0^{1/2}}\log\left(\frac{1}{\sigma_0\epsilon_{m_0}}\right)\right) \quad (23)$$

gradient and Hessian vector product evaluations. Note that for the proposed adaptive sample size scheme we have $\epsilon_{m_0}=\mathcal{O}(1/\sqrt{m_0})$. Further, based on the conditions on m_0 it can be shown that $\alpha_{m_0}\geq\frac{3}{4}\alpha$ and $\beta_{m_0}\geq\frac{1}{2}\beta$ (see (62)) and therefore $\sigma_0\geq\sigma/2$. Hence, if we use a GD update for our proposed adaptive sample size scheme the overall number of gradient evaluations is

$$\mathcal{O}\Bigg(\frac{m_0M^{1/2}L^2\log(\sigma^{-2})}{\sigma^{7/2}}+\frac{m_0M^{1/2}\log\left(\frac{\sqrt{m_0}}{\sigma}\right)}{\sigma^{1/2}} + N\ \max\left\{\frac{8\alpha L}{\beta^2},1+\frac{2M}{\beta}\right\}\Bigg). \quad (24)$$

If we use the accelerate gradient method then the overall number of gradient evaluations is

$$\mathcal{O}\Bigg(\frac{m_0M^{1/2}L^2\log(\sigma^{-2})}{\sigma^{7/2}}+\frac{m_0M^{1/2}\log\left(\frac{\sqrt{m_0}}{\sigma}\right)}{\sigma^{1/2}} + N\sqrt{\frac{2M}{\beta}}\log\left(\frac{2M}{\beta}\right)\Bigg). \quad (25)$$

	Complexity
ANM	$\mathcal{O}(N\sigma^{-7/2})$
ANM + adaptive GD	$\mathcal{O}(m_0\sigma^{-7/2}+N\sigma^{-2})$
ANM + adaptive AGD	$\mathcal{O}(m_0\sigma^{-7/2}+N\sigma^{-1/2})$
ANM + adaptive Newton	$\mathcal{O}(m_0\sigma^{-7/2}+N)$

Table 1: Overall number of processed samples for achieving a local minimum of the ERM problem with N samples is (σ^2,σ)-strongly Morse. Here, m_0 is the size of the initial training set for adaptive methods.

If we use a second-order based update for solving the ERM subproblems, the overall number of gradient evaluations is

$$\mathcal{O}\left(\frac{m_0M^{1/2}L^2\log(\sigma^{-2})}{\sigma^{7/2}}+\frac{m_0M^{1/2}\log\left(\frac{\sqrt{m_0}}{\sigma}\right)}{\sigma^{1/2}}+N\right), \quad (26)$$

while the number of Hessian and Hessian inverse evaluations are $2N$ and $\log(N)$, respectively.

Comparing the theoretical bound in (22) with the ones in (24)-(26) demonstrates the advantage of our proposed adaptive sample size scheme when $N >> m_0$. We state the simplified versions of the bounds, in Table 1. Note that to simplify the bounds presented in Table 1, we replace β by its lower bound σ. The adaptive sample size framework reduces the overall number of samples processed when the total number of samples N is significantly larger than the size of the initial training set m_0. This is indeed the case in many real applications with $N >> p$ as the lower bound on m_0 is almost proportional to the dimension of the problem p. In this section, we only studied the effect of applying the adaptive sample size scheme on the ANM method, but, indeed, similar conclusions will be achieved if it is applied to other state-of-the-art methods for finding second-order stationary points.

6 Conclusions

In this paper we proposed an adaptive sample size scheme which exploits statistical properties of the empirical risk minimization problem to obtain one of its local minimums efficiently, under the assumption that the expected risk is strongly Morse. Our theoretical results suggest that if the dimension of the problem p is significantly smaller than the total number of samples N, the overall the proposed scheme for finding an approximate local minimum of ERM is substantially lower than existing fixed sample size methods.

References

N. Agarwal, Z. Allen Zhu, B. Bullins, E. Hazan, and T. Ma. Finding approximate local minima faster than gradient descent. In *STOC*, pages 1195–1199, 2017.

Z. Allen-Zhu. Natasha 2: Faster non-convex optimization than SGD. *CoRR*, abs/1708.08694, 2017.

Z. Allen Zhu and E. Hazan. Variance reduction for faster non-convex optimization. In *ICML*, pages 699–707, 2016.

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

D. P. Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems 20, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 161–168, 2007.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *ICML*, pages 654–663, 2017a.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points i. *arXiv preprint arXiv:1710.11606*, 2017b.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points ii: First-order methods. *arXiv preprint arXiv:1711.00841*, 2017c.

Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

H. Daneshmand, A. Lucchi, and T. Hofmann. Starting small-learning with adaptive sample sizes. In *International conference on machine learning*, pages 1463–1471, 2016.

H. Daneshmand, J. M. Kohler, A. Lucchi, and T. Hofmann. Escaping saddles with stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 1163–1172, 2018.

A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014a.

A. Defazio, J. Domke, et al. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pages 1125–1133, 2014b.

R. Frostig, R. Ge, S. M. Kakade, and A. Sidford. Competing with the empirical risk minimizer in a single pass. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 728–763, 2015. URL `http://jmlr.org/proceedings/papers/v40/Frostig15.html`.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points - online stochastic gradient for tensor decomposition. In *COLT*, pages 797–842, 2015.

M. Gürbüzbalaban, A. Ozdaglar, and P. Parrilo. A globally convergent incremental newton method. *Mathematical Programming*, 151(1):283–313, 2015.

M. Gürbüzbalaban, A. Ozdaglar, and P. A. Parrilo. On the convergence rate of incremental aggregated gradient algorithms. *SIAM Journal on Optimization*, 27(2):1035–1048, 2017.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *ICML*, pages 1724–1732, 2017a.

C. Jin, P. Netrapalli, and M. I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. *CoRR*, abs/1711.10456, 2017b.

N. Le Roux, M. W. Schmidt, F. R. Bach, et al. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2672–2680, 2012.

L. Lei, C. Ju, J. Chen, and M. I. Jordan. Non-convex finite-sum optimization via SCSG methods. In *Advances in Neural Information Processing Systems 30*, pages 2345–2355, 2017.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

S. Mei, Y. Bai, A. Montanari, et al. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

A. Mokhtari and A. Ribeiro. First-order adaptive sample size methods to reduce complexity of empirical risk minimization. In *Advances in Neural Information Processing Systems*, pages 2060–2068, 2017.

A. Mokhtari, H. Daneshmand, A. Lucchi, T. Hofmann, and A. Ribeiro. Adaptive Newton method for empirical risk minimization to statistical accuracy. In *Advances in Neural Information Processing Systems*, pages 4062–4070, 2016.

A. Mokhtari, M. Gürbüzbalaban, and A. Ribeiro. Surpassing gradient descent provably: A cyclic incremental method with linear convergence rate. *SIAM Journal on Optimization*, 28(2):1420–1447, 2018.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

S. Paternain, A. Mokhtari, and A. Ribeiro. A second order method for nonconvex optimization. *arXiv preprint arXiv:1707.08028*, 2017.

S. J. Reddi, A. Hefny, S. Sra, B. Póczos, and A. J. Smola. Stochastic variance reduction for nonconvex optimization. In *ICML*, pages 314–323, 2016a.

S. J. Reddi, S. Sra, B. Póczos, and A. J. Smola. Fast incremental method for smooth nonconvex optimization. In *IEEE Conference on Decision and Control, CDC*, pages 1971–1977, 2016b.

S. J. Reddi, M. Zaheer, S. Sra, B. Póczos, F. Bach, R. Salakhutdinov, and A. J. Smola. A generic approach for escaping saddle points. In *AISTATS*, pages 1233–1242, 2018.

C. W. Royer and S. J. Wright. Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477, 2018.

N. D. Vanli, M. Gürbüzbalaban, and A. Ozdaglar. Global convergence rate of proximal incremental aggregated gradient methods. *SIAM Journal on Optimization*, 28(2):1282–1300, 2018.

V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

S. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

Y. Xu and T. Yang. First-order stochastic algorithms for escaping from saddle points in almost linear time. *arXiv preprint arXiv:1711.01944*, 2017.

7 Supplementary Material

7.1 Proof of Proposition 1

We only prove the result in (8) which shows an upper bound on the difference between gradients of the risk functions R_m and R_n associated with the sets $\mathcal{S}_m$ and $\mathcal{S}_n$, respectively. Indeed, by following the same steps one can prove the result in (9) for the Hessians.

To do so, consider the difference

$$\nabla R_n(\mathbf{w}) - \nabla R_m(\mathbf{w}) = \frac{1}{n}\sum_{i\in\mathcal{S}_n} \nabla \ell(\mathbf{w}, \mathbf{z}_i) - \frac{1}{m}\sum_{i\in\mathcal{S}_m} \nabla \ell(\mathbf{w}, \mathbf{z}_i). \tag{27}$$

Notice that the set $\mathcal{S}_m$ is a subset of the set $\mathcal{S}_n$ and we can write $\mathcal{S}_n = \mathcal{S}_m \cup \mathcal{S}_{n-m}$. Thus, we can rewrite the right hand side of (27) as

$$\begin{aligned}\nabla R_n(\mathbf{w}) - \nabla R_m(\mathbf{w}) &= \frac{1}{n}\left[\sum_{i\in\mathcal{S}_m} \nabla \ell(\mathbf{w}, \mathbf{z}_i) + \sum_{i\in\mathcal{S}_{n-m}} \nabla \ell(\mathbf{w}, \mathbf{z}_i)\right] - \frac{1}{m}\sum_{i\in\mathcal{S}_m} \ell(\mathbf{w}, \mathbf{z}_i) \\ &= \frac{1}{n}\sum_{i\in\mathcal{S}_{n-m}} \nabla \ell(\mathbf{w}, \mathbf{z}_i) - \frac{n-m}{mn}\sum_{i\in\mathcal{S}_m} \nabla \ell(\mathbf{w}, \mathbf{z}_i). \end{aligned} \tag{28}$$

Factoring $(n-m)/n$ from the terms on the right hand side of (28) yields

$$\nabla R_n(\mathbf{w}) - \nabla R_m(\mathbf{w}) = \frac{n-m}{n}\left[\frac{1}{n-m}\sum_{i\in\mathcal{S}_{n-m}} \nabla \ell(\mathbf{w}, \mathbf{z}_i) - \frac{1}{m}\sum_{i\in\mathcal{S}_m} \nabla \ell(\mathbf{w}, \mathbf{z}_i)\right]. \tag{29}$$

Now add and subtract the gradient of the population risk $\nabla R(\mathbf{w})$ to obtain

$$\begin{aligned}\|\nabla R_n(\mathbf{w}) - \nabla R_m(\mathbf{w})\| &= \frac{n-m}{n}\left\|\frac{1}{n-m}\sum_{i\in\mathcal{S}_{n-m}} \nabla \ell(\mathbf{w}, \mathbf{z}_i) - \nabla R(\mathbf{w}) + \nabla R(\mathbf{w}) - \frac{1}{m}\sum_{i\in\mathcal{S}_m} \nabla \ell(\mathbf{w}, \mathbf{z}_i)\right\| \\ &\leq \frac{n-m}{n}\left(\|\nabla R_{n-m} - \nabla R(\mathbf{w})\| + \|\nabla R_m - \nabla R(\mathbf{w})\|\right), \end{aligned} \tag{30}$$

where the last inequality follows by using the triangle inequality. By using the result in Theorem 1 twice, the claim in (8) follows from the inequality in (30).

7.2 Proof of Theorem 2

Consider $\mathbf{w}_m$ as an approximate local minimum of the risk R_m corresponding to the set $\mathcal{S}_m$. Based on the hypothesis of the theorem we know that

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \epsilon_m := \tau\sqrt{\frac{Cp\log m}{m}}. \tag{31}$$

Further based on the condition on m_0 in (13) it can be verified that for any $m \geq m_0$ it holds that

$$\alpha - 2\tau\sqrt{\frac{Cp\log m}{m}} \geq \tau\sqrt{\frac{Cp\log m}{m}} \tag{32}$$

and therefore

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \alpha - 2\tau\sqrt{\frac{Cp\log m}{m}}. \tag{33}$$

Further, using the result of Proposition 1 for $n = 2m$, we know that with high probability (at least $1-\delta$) the following inequality holds

$$\begin{aligned}\|\nabla R_n(\mathbf{w}_m) - \nabla R_m(\mathbf{w}_m)\|_2 &\leq \frac{n-m}{n}\ \tau\sqrt{Cp}\left(\sqrt{\frac{\log(n-m)}{n-m}} + \sqrt{\frac{\log m}{m}}\right) \\ &\leq \tau\sqrt{\frac{Cp\log m}{m}}. \end{aligned} \tag{34}$$

By combining the bounds in (31), (33) and (34) we obtain that with probability at least $1-\delta$ the norm of gradient $\|\nabla R_n(\mathbf{w}_m)\|$ is bounded above by

$$\|\nabla R_n(\mathbf{w}_m)\| \leq \min\left\{\alpha - \tau\sqrt{\frac{Cp\log m}{m}}\ ,\ 2\tau\sqrt{\frac{Cp\log m}{m}}\right\}. \tag{35}$$

Note that for any $n > m \geq 1$ it holds that

$$\frac{\log m}{m} \geq \frac{\log n}{n}. \tag{36}$$

By using the inequalities in (35) and (36) it can be shown that

$$\|\nabla R_n(\mathbf{w}_m)\| \leq \min\left\{\alpha - \tau\sqrt{\frac{Cp\log 2m}{2m}}\ ,\ 2\tau\sqrt{\frac{Cp\log m}{m}}\right\}. \tag{37}$$

Replace $2m$ by n and use the defitinion of α_n in (11) to obtain that with probability at least $1-\delta$ it holds

$$\|\nabla R_n(\mathbf{w}_m)\| \leq \min\left\{\alpha_n\ ,\ 2\tau\sqrt{\frac{Cp\log m}{m}}\right\}. \tag{38}$$

The result in (38) in conjunction with the Morse property of the risk R_n implies that the absolute value of the eigenvalues of the Hessian $\nabla^2 R_n(\mathbf{w}_m)$ are larger than β_n, i.e.,

$$|\lambda_i(\nabla^2 R_n(\mathbf{w}_m))| \geq \beta_n, \qquad \text{for all } i = 1,\dots,p. \tag{39}$$

However, it does not imply that $\mathbf{w}_m$ is close to a local minimum since at least one of the eigenvalues of the Hessian $\nabla^2 R_n(\mathbf{w}_m)$ might be negative. Hence, we need to ensure that the Hessian $\nabla^2 R_n(\mathbf{w}_m)$ is at least positive semidefinite.

For any unit size vector $\mathbf{v} \in \mathbb{R}^p$ it holds

$$\mathbf{v}^\top \nabla^2 R_n(\mathbf{w}_m)\mathbf{v} \geq \mathbf{v}^\top \nabla^2 R_m(\mathbf{w}_m)\mathbf{v} - \|\nabla^2 R_n(\mathbf{w}_m) - \nabla^2 R_m(\mathbf{w}_m)\|_{op} \tag{40}$$

Further, using the result of Proposition 1 for $n = 2m$ we can write

$$\begin{aligned}\|\nabla^2 R_n(\mathbf{w}_m) - \nabla^2 R_m(\mathbf{w}_m)\|_{op} &\leq \frac{n-m}{n}\tau^2\sqrt{Cp}\left(\sqrt{\frac{\log(n-m)}{n-m}} + \sqrt{\frac{\log m}{m}}\right)\\ &\leq \tau^2\sqrt{\frac{Cp\log m}{m}}\end{aligned} \tag{41}$$

with high probability. Using this result, the lower bound in (40) can be simplified to

$$\mathbf{v}^\top \nabla^2 R_n(\mathbf{w}_m)\mathbf{v} \geq \mathbf{v}^\top \nabla^2 R_m(\mathbf{w}_m)\mathbf{v} - \tau^2\sqrt{\frac{Cp\log m}{m}}. \tag{42}$$

According to the condition in (33), if we know that $\|\nabla R_m(\mathbf{w}_m)\| \leq \alpha_m$ then $\|\nabla^2 R_m(\mathbf{w}_m)\| \geq \beta_m$. As $\mathbf{w}_m$ is an approximate local minimum of the risk R_m we obtain that $\nabla^2 R_m(\mathbf{w}_m) \succeq \mathbf{0}$. By combining these two observations we obtain that for any $\mathbf{v} \in \mathbb{R}^p$ it holds

$$\mathbf{v}^\top \nabla^2 R_m(\mathbf{w}_m)\mathbf{v} \geq \beta_m = \beta - \tau^2\sqrt{\frac{Cp\log m}{m}}. \tag{43}$$

Applying this substitution into (42) implies that for any $\mathbf{v} \in \mathbb{R}^p$

$$\mathbf{v}^\top \nabla^2 R_n(\mathbf{w}_m)\mathbf{v} \geq \beta - 2\tau^2\sqrt{\frac{Cp\log m}{m}}. \tag{44}$$

Note that based on the size of the initial training set $\mathcal{S}_{m_0}$ in (13) we know that

$$\frac{m_0}{\log m_0} \geq \frac{4Cp\tau^4}{\beta^2} \iff \beta \geq 2\tau^2\sqrt{\frac{Cp\log m_0}{m_0}} \tag{45}$$

This inequality together with the one in (36) leads to the conclusion that the expression on right hand side of (44) is non-negative, and, therefore, the Hessian $\nabla^2 R_n(\mathbf{w}_m)$ is positive semidefinite, i.e., $\nabla^2 R_n(\mathbf{w}_m) \succeq 0$. This result and the expression in (39) lead to the conclusion that

$$\nabla^2 R_n(\mathbf{w}_m) \succeq \beta_n \mathbf{I}. \tag{46}$$

Therefore, $\mathbf{w}_m$ is within a neighborhood of a local minimum of R_n.

Now we need to use the point $\mathbf{w}_m$ to reach a point $\mathbf{w}_n$ that satisfies the conditions

$$\|\nabla R_n(\mathbf{w}_n)\| \leq \epsilon_n, \quad \nabla^2 R_n(\mathbf{w}_n) \succeq \mathbf{0}. \tag{47}$$

As we stated in this proof so far, the conditions in (47) ensure that the variable $\mathbf{w}_n$ is within a neighborhood of a local minimum of the risk R_{2n}.

There are two important questions that we need to address. First, if we start with the point $\mathbf{w}_m$ which satisfies the conditions in (37) and (46), can we show that all the iterates stay close to the local minimum and do not converge to a saddle point? Second, how many iterations of gradient descent method is needed to reach a point $\mathbf{w}_n$ satisfying the conditions in (47)?

Note that so far we have shown that $\mathbf{w}_m$ is close to a local minimum of R_n that we denote by $\mathbf{w}_n^*$. We define the set $\mathcal{C}_{\mathbf{w}_n^*}$ as a neighborhood around the local minimum $\mathbf{w}_n^*$ that all points in this neighborhood satisfy the conditions $\|\nabla R_n(\mathbf{w})\| \leq \alpha_n$ and $\nabla^2 \nabla R_n(\mathbf{w}) \succeq \beta_n \mathbf{I}$. Indeed, based on our definition both $\mathbf{w}_m$ and $\mathbf{w}_n^*$ belong to the set $\mathcal{C}_{\mathbf{w}_n^*}$. If we update the variable $\mathbf{w}_m$ using a first-order update and guarantee that the norm of gradient does not increase, it follows from the Morse property of the risk R_n that he updated variables stay in the set $\mathcal{C}_{\mathbf{w}_n^*}$ and therefore the Hessian at the updated point stays positive definite. Although the standard update of gradient descent around a local minimum ensures that the objective function value decreases, it does not guarantee that the norm of gradient does not grow. Hence, by the standard analysis of the gradient descent algorithm we cannot ensure that the updated variable stays in the set $\mathcal{C}_{\mathbf{w}_n^*}$ and has a gradient norm smaller than α_n, and, consequently, its Hessian may not be positive definite. Therefore, here we choose the stepsize of the gradient descent update different from the standard $1/M$ or $2/(M+\beta_n)$, to ensure that the norm of gradient decreases at each step and the iterates do not leave the neighborhood of the local minimum close to $\mathbf{w}_m$.

Set $\tilde{\mathbf{w}}^0 = \mathbf{w}_m$ and follow the update

$$\tilde{\mathbf{w}}^{i+1} = \tilde{\mathbf{w}}^i - \eta_n \nabla R_n(\tilde{\mathbf{w}}^i). \tag{48}$$

with the stepsize $\eta_n = \min\left\{\frac{\beta_n}{\alpha_n L}, \frac{2}{\beta_n + M}\right\}$ where L is the constant for the Lipschitz continuity of the Hessians as defined in Assumption 3.

Assume that the iterate $\tilde{\mathbf{w}}^i$ belongs to the set $\mathcal{C}_{\mathbf{w}_n^*}$. Using the Taylor's expansion of the function R_n we can write

$$\begin{aligned}
\|\nabla R_n(\tilde{\mathbf{w}}^{i+1})\| &\leq \|\nabla R_n(\tilde{\mathbf{w}}^i) + \nabla^2 R_n(\tilde{\mathbf{w}}^i)(\tilde{\mathbf{w}}^{i+1} - \tilde{\mathbf{w}}^i)\| + \frac{L}{2}\|(\tilde{\mathbf{w}}^{i+1} - \tilde{\mathbf{w}}^i)\|^2 \\
&= \|\nabla R_n(\tilde{\mathbf{w}}^i) - \eta_n \nabla^2 R_n(\tilde{\mathbf{w}}^i)\nabla R_n(\tilde{\mathbf{w}}^i)\| + \frac{L\eta_n^2}{2}\|\nabla R_n(\tilde{\mathbf{w}}^i)\|^2 \\
&\leq \|\mathbf{I} - \eta_n \nabla^2 R_n(\tilde{\mathbf{w}}^i)\|\|\nabla R_n(\tilde{\mathbf{w}}^i)\| + \frac{L\eta_n^2}{2}\|\nabla R_n(\tilde{\mathbf{w}}^i)\|^2 \\
&\leq \max\{|1 - \eta_n\beta_n|, |1 - \eta_n M|\}\|\nabla R_n(\tilde{\mathbf{w}}^i)\| + \frac{L\eta_n^2}{2}\|\nabla R_n(\tilde{\mathbf{w}}^i)\|^2
\end{aligned} \tag{49}$$

As $\eta_n \leq \frac{2}{\beta_n + M}$ we obtain that $\max\{|1 - \eta_n\beta_n|, |1 - \eta_n M|\} \leq (1 - \eta_n\beta_n)$. Therefore,

$$\begin{aligned}
\|\nabla R_n(\tilde{\mathbf{w}}^{i+1})\| &\leq (1 - \eta_n\beta_n)\|\nabla R_n(\tilde{\mathbf{w}}^i)\| + \frac{L\eta_n^2}{2}\|\nabla R_n(\tilde{\mathbf{w}}^i)\|^2 \\
&= (1 - \eta_n\beta_n + \frac{L\eta_n^2}{2}\|\nabla R_n(\tilde{\mathbf{w}}^i)\|)\|\nabla R_n(\tilde{\mathbf{w}}^i)\| \\
&\leq (1 - \eta_n\beta_n + \frac{L\alpha_n\eta_n^2}{2})\|\nabla R_n(\tilde{\mathbf{w}}^i)\|,
\end{aligned} \tag{50}$$

where the last inequality holds due to the fact that $\tilde{\mathbf{w}}^i \in \mathcal{C}_{\mathbf{w}_n^*}$ and therefore we know $\|\nabla R_n(\tilde{\mathbf{w}}^i)\| \leq \alpha_n$. Since the stepsize η_n satisfies the inequality $\eta_n \leq \frac{\beta_n}{\alpha_n L}$ we can simplify the upper bound in (50) to

$$\begin{aligned}\|\nabla R_n(\tilde{\mathbf{w}}^{i+1})\| &\leq \left(1 - \frac{\eta_n \beta_n}{2}\right) \|\nabla R_n(\tilde{\mathbf{w}}^i)\| \\ &\leq \left(1 - \min\left\{\frac{\beta_n^2}{2\alpha_n L}, \frac{\beta_n}{\beta_n + M}\right\}\right) \|\nabla R_n(\tilde{\mathbf{w}}^i)\|. \end{aligned} \tag{51}$$

Therefore, the norm of gradient decreases at each step and the updated iterate $\tilde{\mathbf{w}}^{i+1}$ also belongs to the neighborhood $\mathcal{C}_{\mathbf{w}_n^*}$. Using an induction argument it can be easily verified if the initial iterate which is $\mathbf{w}_m$ is in $\mathcal{C}_{\mathbf{w}_n^*}$, then by following the update of gradient descent all the iterates stay in the set $\mathcal{C}_{\mathbf{w}_n^*}$ as they approach the local minimum $\mathbf{w}_n^*$.

Now we study the number of gradient iterations s_n needed to reach a point $\mathbf{w}_n = \tilde{\mathbf{w}}^{s_n}$ with a gradient norm smaller than ϵ_n. If we consider $\mathbf{w}_n$ as the output of running gradient descent for s_n iterations with the initial iterate $\mathbf{w}_m$, then by applying the expression in (51) recursively we can write

$$\begin{aligned}\|\nabla R_n(\mathbf{w}_n)\| &\leq \left[1 - \min\left\{\frac{\beta_n^2}{2\alpha_n L}, \frac{\beta_n}{\beta_n + M}\right\}\right]^{s_n} \|\nabla R_n(\mathbf{w}_m)\| \\ &\leq \left[1 - \min\left\{\frac{\beta_n^2}{2\alpha_n L}, \frac{\beta_n}{\beta_n + M}\right\}\right]^{s_n} 2\tau\sqrt{\frac{Cp\log m}{m}}, \end{aligned} \tag{52}$$

where the second inequality is implied by the condition in (38). Therefore, to ensure that the following condition is satisfied

$$\|\nabla R_n(\mathbf{w}_n)\| \leq \epsilon_n = \tau\sqrt{\frac{Cp\log n}{n}}, \tag{53}$$

we need to check if the following inequality holds:

$$\tau\sqrt{\frac{Cp\log n}{n}} \geq \left[1 - \min\left\{\frac{\beta_n^2}{2\alpha_n L}, \frac{\beta_n}{\beta_n + M}\right\}\right]^{s_n} 2\tau\sqrt{\frac{Cp\log m}{m}}, \tag{54}$$

where $n = 2m$. This condition is satisfied if the number of gradient descent iterations s_n is larger than

$$\begin{aligned} s_n &\geq \max\left\{\frac{2\alpha_n L}{\beta_n^2}, \frac{\beta_n + M}{\beta_n}\right\} \log \frac{2\tau\sqrt{\frac{Cp\log m}{m}}}{\tau\sqrt{\frac{Cp\log n}{n}}} \\ &= \max\left\{\frac{2\alpha_n L}{\beta_n^2}, \frac{\beta_n + M}{\beta_n}\right\} \log \frac{2\sqrt{2\log m}}{\sqrt{\log 2m}} \end{aligned} \tag{55}$$

As $\frac{2\sqrt{2\log m}}{\sqrt{\log 2m}} \leq 4$, we obtain that

$$s_n \geq \max\left\{\frac{2\alpha_n L}{\beta_n^2}, \frac{\beta_n + M}{\beta_n}\right\} \log 4 \tag{56}$$

is a sufficient condition for the inequality in (54).

This result shows that the after running $\left\{\frac{2\alpha_n L}{\beta_n^2}, \frac{\beta_n + M}{\beta_n}\right\} \log 4$ iterations of gradient descent for the initial variable $\tilde{\mathbf{w}}^0 = \mathbf{w}_m$ we reach a point $\tilde{\mathbf{w}}^{s_n} = \mathbf{w}_n$ that satisfies the condition in (53). Further, since the norm of gradient is always decreasing and $\|\nabla R_n(\mathbf{w}_n)\| \leq \|\nabla R_n(\mathbf{w}_m)\| \leq \alpha_n$ and the iterates $\{\tilde{\mathbf{w}}^i\}_{i=0}^{i=s_n}$ never leave the neighborhood $\mathcal{C}_{\mathbf{w}_n^*}$ we obtain that

$$\nabla^2 R_n(\mathbf{w}_n) \succeq \beta_n \mathbf{I}. \tag{57}$$

So far we showed that if we start with a point $\mathbf{w}_m$ which satisfies the following conditions

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \epsilon_m, \quad \nabla^2 R_m(\mathbf{w}_m) \succeq \beta_m \mathbf{I}, \tag{58}$$

and set $n = 2m$, then after running a sufficient number of gradient descent iterations we reach a point $\mathbf{w}_n$ satisfying

$$\|\nabla R_n(\mathbf{w}_n)\| \leq \epsilon_n, \quad \nabla^2 R_n(\mathbf{w}_n) \succeq \beta_n \mathbf{I}, \tag{59}$$

Hence, when we reach the full training set with N samples the final iterate $\mathbf{w}_N$ is such that

$$\|\nabla R_N(\mathbf{w}_N)\| \leq \epsilon_N, \quad \nabla^2 R_N(\mathbf{w}_N) \succeq \beta_N \mathbf{I}, \tag{60}$$

By considering the conditions in (60) and the definitions of ϵ_N and β_N, the claim follows.

Note that we are minimizing the risk R_n using the steps of gradient descents, each iteration requires n gradient evaluations. Since we run s_n iterations for the risk R_n, the overall number of gradient computations when we operate on the risk R_n is ns_n. Hence, if we define $\mathcal{I} = \{m_0, 2m_0, \ldots, N\}$, then the total number of gradient evaluations for the proposed adaptive sample size scheme when we use gradient descent to solve the subproblems is bounded above by

$$\sum_{n \in \mathcal{I}} n s_n = \log 4 \sum_{n \in \mathcal{I}} n \ \max \left\{ \frac{2\alpha_n L}{\beta_n^2}, 1 + \frac{M}{\beta_n} \right\} \tag{61}$$

In the expression in (61) we can replace α_n by α as $\alpha_n \leq \alpha$. Further, using the condition on m_0 it can be shown that

$$\begin{aligned} \beta_n &= \beta - \tau^2 \sqrt{\frac{Cp \log n}{n}} \\ &\geq \beta - \tau^2 \sqrt{\frac{Cp \log m_0}{m_0}} \\ &\geq \beta - \frac{\beta}{2} = \frac{\beta}{2} \end{aligned} \tag{62}$$

where the last inequality is implied by the result in (45). Therefore we can replace β_n in (61) by its lower bound $\beta/2$. Applying these changes leads to

$$\begin{aligned} \sum_{n \in \mathcal{I}} n s_n &\leq \log 4 \sum_{n \in \mathcal{I}} n \ \max \left\{ \frac{2\alpha L}{\beta^2/4}, 1 + \frac{M}{\beta/2} \right\} \\ &\leq 2N \log 4 \ \max \left\{ \frac{8\alpha L}{\beta^2}, 1 + \frac{2M}{\beta} \right\}. \end{aligned} \tag{63}$$

Therefore, the claim of the theorem follows.

7.3 Proof of Theorem 3

Consider $\mathbf{w}_m$ as an approximate local minimum of the risk R_m corresponding to the set $\mathcal{S}_m$. Based on the hypothesis of the theorem we know that

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \epsilon_m := \tau \sqrt{\frac{Cp \log m}{m}}. \tag{64}$$

Further, based on the condition on m_0

$$\tau \sqrt{\frac{Cp \log m_0}{m_0}} \left(1 + \sqrt{\frac{16M}{\beta}} \right) \leq \alpha \tag{65}$$

Note that $\log m_0/m_0 \geq \log m/m$ for any $m \geq m_0$. Further, based on the condition on m_0 and the argument in (62) we know that $\beta \leq 2\beta_n$. Hence, we can write that

$$\tau \sqrt{\frac{Cp \log m}{m}} \left(1 + \sqrt{\frac{8M}{\beta_n}} \right) \leq \alpha \tag{66}$$

By regrouping the terms we obtain that

$$2\tau\sqrt{\frac{Cp\log m}{m}}\sqrt{\frac{2M}{\beta_n}} \leq \alpha - \tau\sqrt{\frac{Cp\log m}{m}} \tag{67}$$

As $n \geq m$ we can write

$$2\tau\sqrt{\frac{Cp\log m}{m}}\sqrt{\frac{2M}{\beta_n}} \leq \alpha - \tau\sqrt{\frac{Cp\log n}{n}} = \alpha_n \tag{68}$$

and therefore

$$2\tau\sqrt{\frac{Cp\log m}{m}} \leq \alpha_n\sqrt{\frac{\beta_n}{2M}} \tag{69}$$

Subtracting $\tau\sqrt{\frac{Cp\log m}{m}}$ from both sides implies that

$$\tau\sqrt{\frac{Cp\log m}{m}} \leq \alpha_n\sqrt{\frac{\beta_n}{2M}} - \tau\sqrt{\frac{Cp\log m}{m}} \tag{70}$$

Hence, using the bound in (64) we can write

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \alpha_n\sqrt{\frac{\beta_n}{2M}} - \tau\sqrt{\frac{Cp\log m}{m}} \tag{71}$$

Now by using the result in (34) it can be shown that

$$\|\nabla R_n(\mathbf{w}_m)\| \leq \alpha_n\sqrt{\frac{\beta_n}{2M}} \tag{72}$$

Indeed, the result in (34) and (64) also shows that $\|\nabla R_n(\mathbf{w}_m)\| \leq 2\tau\sqrt{\frac{Cp\log m}{m}}$ and therefore we have

$$\|\nabla R_n(\mathbf{w}_m)\| \leq \min\left\{\alpha_n\left(\sqrt{\frac{\beta_n}{2M}}\right),\ 2\tau\sqrt{\frac{Cp\log m}{m}}\right\}. \tag{73}$$

Note that $\sqrt{\beta_n/M} \leq 1$ and therefore $\|\nabla R_n(\mathbf{w}_m)\| \leq \alpha_n$. Hence, all the arguments from (39) to (46) are still valid. From the convergence of accelerated gradient method in [Nesterov, 2013] we know that if the function f is μ-strongly convex and M-smooth, the iterates generated by accelerated gradient method satisfy the inequality

$$f(\tilde{\mathbf{w}}^k) - f(\mathbf{w}^*) \leq \left(1 - \sqrt{\frac{\mu}{M}}\right)^k \left[f(\tilde{\mathbf{w}}^0) - f(\mathbf{w}^*) + \frac{\gamma_0}{2}\|\tilde{\mathbf{w}}^0 - \mathbf{w}^*\|^2\right] \tag{74}$$

where $\mathbf{x}^*$ is the minimum and $\gamma_0 \geq \mu$. By setting $\gamma_0 = \mu$ and using the following inequalities

$$f(\tilde{\mathbf{w}}^k) - f(\mathbf{w}^*) \geq \frac{1}{2M}\|\nabla f(\tilde{\mathbf{w}}^k)\|^2, \qquad \frac{\mu}{2}\|\tilde{\mathbf{w}}^0 - \mathbf{w}^*\|^2 \leq f(\tilde{\mathbf{w}}^0) - f(\mathbf{w}^*), \qquad f(\tilde{\mathbf{w}}^0) - f(\mathbf{w}^*) \leq \frac{1}{2\mu}\|\nabla f(\tilde{\mathbf{w}}^0)\|^2 \tag{75}$$

one can show that

$$\|\nabla f(\tilde{\mathbf{w}}^k)\| \leq \left(1 - \sqrt{\frac{\mu}{M}}\right)^{k/2} \sqrt{\frac{2M}{\mu}}\ \|\nabla f(\tilde{\mathbf{w}}^0)\|. \tag{76}$$

Hence, we can show that the iterates generated by adaptive samples size AGD always have a gradient norm smaller than α_n. To verify this claim note that the initial error at each stage is bounded above by

$$\|\nabla R_n(\tilde{\mathbf{w}}^0)\| = \|\nabla R_n(\mathbf{w}_m)\| \leq \alpha_n\left(\sqrt{\frac{\beta_n}{2M}}\right) \tag{77}$$

By combining this result with the inequality in (76) we obtain that for all intermediate iterates $\tilde{\mathbf{w}}^k$ we can write

$$\|\nabla R_n(\tilde{\mathbf{w}}^k)\| \leq \left(1 - \sqrt{\frac{\beta_n}{M}}\right)^{k/2} \alpha_n \leq \alpha_n \tag{78}$$

Note that we used the fact that $\mu = \beta_n$ in our setting. Therefore, the iterates always stay in the region that norm of gradient is smaller than α_n. This observation simply shows that the iterates always within a neighborhood of a local minimum.

It remains to characterize the number of accelerated gradient iterations required to find a point $\mathbf{w}_n$ that satisfies $\|\nabla R_n(\mathbf{w}_n)\| \leq \tau\sqrt{\frac{Cp\log n}{n}}$. To do so, note that after s_n iterations the output of the process which is $\mathbf{w}_n$ satisfies the inequality

$$\begin{aligned}\|\nabla R_n(\mathbf{w}_n)\| &\leq \sqrt{\frac{2M}{\beta_n}}\left(1-\sqrt{\frac{\beta_n}{M}}\right)^{s_n/2}\|\nabla R_n(\mathbf{w}_m)\| \\ &\leq \sqrt{\frac{2M}{\beta_n}}\left(1-\sqrt{\frac{\beta_n}{M}}\right)^{s_n/2} 2\tau\sqrt{\frac{Cp\log m}{m}}. \end{aligned} \tag{79}$$

Therefore we need to ensure that s_n is large enough that the following condition is satisfied

$$\begin{aligned}&\tau\sqrt{\frac{Cp\log n}{n}} \geq \sqrt{\frac{2M}{\beta_n}}\left(1-\sqrt{\frac{\beta_n}{M}}\right)^{s_n/2} 2\tau\sqrt{\frac{Cp\log m}{m}} \\ &\iff \left(\frac{16M}{\beta_n}\right)\left(1-\sqrt{\frac{\beta_n}{M}}\right)^{s_n} \leq \frac{\log 2m}{\log m}. \end{aligned} \tag{80}$$

As $\frac{\log 2m}{\log m} \geq 1$, this condition is satisfied if

$$\left(\frac{16M}{\beta_n}\right)\left(1-\sqrt{\frac{\beta_n}{M}}\right)^{s_n} \leq 1. \tag{81}$$

This is satisfied if

$$s_n \geq \sqrt{\frac{M}{\beta_n}}\log\left(\frac{16M}{\beta_n}\right). \tag{82}$$

Note that we are minimizing the risk R_n using the steps of AGD, and each iteration requires n gradient evaluations. Since we run s_n iterations for the risk R_n, the overall number of gradient computations when we operate on the risk R_n is ns_n. Hence, if we define $\mathcal{I} = \{m_0, 2m_0, \ldots, N\}$, then the total number of gradient evaluations for the proposed adaptive sample size scheme when we use AGD to solve the subproblems is bounded above by

$$\sum_{n\in\mathcal{I}} ns_n = \sum_{n\in\mathcal{I}} n\sqrt{\frac{M}{\beta_n}}\log\left(9\left(\frac{M}{\beta_n}+M\right)\right) \tag{83}$$

Using the condition on m_0 it can be shown that $\beta_n \geq \beta/2$ and therefore we can replace β_n in (83) by its lower bound $\beta/2$. Applying these changes leads to

$$\begin{aligned}\sum_{n\in\mathcal{I}} ns_n &\leq \sum_{n\in\mathcal{I}} n\sqrt{\frac{2M}{\beta}}\log\left(\frac{16M}{\beta}\right) \\ &\leq 2N\sqrt{\frac{2M}{\beta}}\log\left(\frac{16M}{\beta}\right). \end{aligned} \tag{84}$$

Therefore, the claim of the theorem follows.

7.4 Proof of Theorem 4

Consider $\mathbf{w}_m$ as an approximate local minimum of the risk R_m corresponding to the set $\mathcal{S}_m$. Based on the hypothesis of the theorem we know that

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \tau\sqrt{\frac{Cp\log m}{m}} \tag{85}$$

Further based on the condition on m_0 in (20) it can be verified that for any $m \geq m_0$ it holds that

$$\alpha - 2\tau\sqrt{\frac{Cp\log m}{m}} \geq \tau\sqrt{\frac{Cp\log m}{m}} \tag{86}$$

and therefore

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \alpha - 2\tau\sqrt{\frac{Cp\log m}{m}}. \tag{87}$$

In addition, the condition in (20) implies that for any $m \geq m_0$ we have

$$\beta^2 \geq \sqrt{\frac{Cp\log m}{m}}(2\tau^2 + 4L\tau) \tag{88}$$

Regroup the terms to obtain that

$$\begin{aligned} 4L\tau\sqrt{\frac{Cp\log m}{m}} &\leq \beta^2 - 2\tau^2\sqrt{\frac{Cp\log m}{m}} \\ &\leq \left(\beta - \tau^2\sqrt{\frac{Cp\log m}{m}}\right)^2 \end{aligned} \tag{89}$$

Divide both sides by $2L$ and regroup the terms to obtain

$$\tau\sqrt{\frac{Cp\log m}{m}} \leq \frac{1}{2L}\left(\beta - \tau^2\sqrt{\frac{Cp\log m}{m}}\right)^2 - \tau\sqrt{\frac{Cp\log m}{m}} \tag{90}$$

As $\|\nabla R_m(\mathbf{w}_m)\| \leq \tau\sqrt{\frac{Cp\log m}{m}}$ we obtain that

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \frac{1}{2L}\left(\beta - \tau^2\sqrt{\frac{Cp\log m}{m}}\right)^2 - \tau\sqrt{\frac{Cp\log m}{m}} \tag{91}$$

We can summarize these inequalities as

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \min\left\{\alpha - 2\tau\sqrt{\frac{Cp\log m}{m}}\ ,\ \tau\sqrt{\frac{Cp\log m}{m}}\ ,\ \frac{1}{2L}\left(\beta - \tau^2\sqrt{\frac{Cp\log m}{m}}\right)^2 - \tau\sqrt{\frac{Cp\log m}{m}}\right\}. \tag{92}$$

Using the result in Proposition 1 for $n = 2m$, we know that with high probability (at least $1 - \delta$) we have

$$\|\nabla R_n(\mathbf{w}_m) - \nabla R_m(\mathbf{w}_m)\|_2 \leq \tau\sqrt{\frac{Cp\log m}{m}}. \tag{93}$$

By combining the bounds (92) and (93) we obtain that with probability at least $1 - \delta$ the norm of gradient $\|\nabla R_n(\mathbf{w}_m)\|$ is bounded above by

$$\|\nabla R_n(\mathbf{w}_m)\| \leq \min\left\{\alpha - \tau\sqrt{\frac{Cp\log m}{m}}\ ,\ 2\tau\sqrt{\frac{Cp\log m}{m}}\ ,\ \frac{1}{2L}\left[\beta - \tau^2\sqrt{\frac{Cp\log m}{m}}\right]^2\right\}. \tag{94}$$

Using the definitions of α_n and β_n, and the fact that $m/\log m \leq n/\log n$ for $n \geq m$, we can write

$$\|\nabla R_n(\mathbf{w}_m)\| \leq \min\left\{\alpha_n\ ,\ 2\tau\sqrt{\frac{Cp\log m}{m}}\ ,\ \frac{\beta_n^2}{2L}\right\}. \tag{95}$$

Note that according to the result in (95) the norm of gradient $\|\nabla R_n(\mathbf{w}_m)\|$ is smaller than α_n. By following the argument from (38) to (46) we obtain that $\nabla^2 R_n(\mathbf{w}_m) \succeq \beta_n \mathbf{I}$, and, therefore, $\mathbf{w}_m$ is within a neighborhood of a local minimum of R_n.

Further, the condition in (95) implies that $\|\nabla R_n(\mathbf{w}_m)\| \leq \frac{\beta_n^2}{L}$. We show that this condition is sufficient to ensure that $\mathbf{w}_m$ is within the quadratic convergence region of Newton's method for the objective function R_n. To do so, note that based on L-Lipschitz continuity of the Hessian $\nabla^2 R_n$ we know that the iterate $\mathbf{w}_n = \mathbf{w}_m - \nabla^2 R_n(\mathbf{w}_m)^{-1}\nabla R_n(\mathbf{w}_m)$ satisfies

$$\begin{aligned}\|\nabla R_n(\mathbf{w}_n)\| &\leq \|\nabla R_n(\mathbf{w}_m) + \nabla^2 R_n(\mathbf{w}_m)(\mathbf{w}_n - \mathbf{w}_m)\| + \frac{L}{2}\|\mathbf{w}_n - \mathbf{w}_m\|^2 \\ &= \frac{L}{2}\|\mathbf{w}_n - \mathbf{w}_m\|^2 \\ &= \frac{L}{2}\|\nabla^2 R_n(\mathbf{w}_m)^{-1}\nabla R_n(\mathbf{w}_m)\|^2 \\ &\leq \frac{L}{2\beta_n^2}\|\nabla R_n(\mathbf{w}_m)\|^2 \end{aligned} \tag{96}$$

According to this inequality and the condition $\|\nabla R_n(\mathbf{w}_m)\| \leq \frac{\beta_n^2}{2L} < \frac{2\beta_n^2}{L}$, we obtain that $\mathbf{w}_m$ is within the quadratic convergence region of Newton's method for the risk R_n. Therefore by runnig iterations of Newton's method the norm $\|\nabla R_n\|$ approaches zero quadratically. However, we proceed to show that only a single iteration of Newton's method is enough to ensure that the update variable $\mathbf{w}_n$ satisfies the condition $\|\nabla R_n(\mathbf{w}_n)\| \leq \epsilon_n$. Note that by combining the inequality in (96) with the one in (95) we can write

$$\begin{aligned}\|\nabla R_n(\mathbf{w}_n)\| &\leq \frac{L}{2\beta_n^2}\|\nabla R_n(\mathbf{w}_m)\|^2 \\ &\leq \frac{1}{4}\|\nabla R_n(\mathbf{w}_m)\| \\ &\leq \frac{\tau}{2}\sqrt{\frac{Cp\log m}{m}}. \end{aligned} \tag{97}$$

The last step is to verify the condition

$$\frac{\tau}{2}\sqrt{\frac{Cp\log m}{m}} \leq \tau\sqrt{\frac{Cp\log n}{n}}, \tag{98}$$

which holds since $\log m \leq 2\log 2m$. Considering these conditions and the inequality in (97), we can show that

$$\|\nabla R_n(\mathbf{w}_n)\| \leq \epsilon_n. \tag{99}$$

Further, since $\|\nabla R_n(\mathbf{w}_m)\| \succeq \beta_n \mathbf{I}$ and $\mathbf{w}_n$ stays in the neighborhood of the local minimum with gradient norm smaller than α_n we obtain that $\|\nabla R_n(\mathbf{w}_n)\| \succeq \beta_n \mathbf{I}$.

Hence, we showed that if we start with a point $\mathbf{w}_m$ which satisfies the following conditions

$$\|\nabla R_m(\mathbf{w}_m)\| \leq \epsilon_m, \quad \nabla^2 R_m(\mathbf{w}_m) \succeq \beta_m \mathbf{I}, \tag{100}$$

and set $n = 2m$, then after running a single iteration of Newton's method we reach a point $\mathbf{w}_n$ satisfying

$$\|\nabla R_n(\mathbf{w}_n)\| \leq \epsilon_n, \quad \nabla^2 R_n(\mathbf{w}_n) \succeq \beta_n \mathbf{I}, \tag{101}$$

Therefore, when we reach the full training set with N samples the final iterate $\mathbf{w}_N$ is such that

$$\|\nabla R_N(\mathbf{w}_N)\| \leq \epsilon_N, \quad \nabla^2 R_N(\mathbf{w}_N) \succeq \beta_N \mathbf{I}, \tag{102}$$

As the number of Newton's iteration at each stage is 1, the overall number of Newton direction steps that we need to compute is $\log_2(N/m_0)$. If we define $\mathcal{I} = \{m_0, 2m_0, \ldots, N\}$, then the total number of gradient evaluations for the proposed adaptive sample size scheme when we use Newton's update is bounded above by

$$\sum_{n\in\mathcal{I}} n s_n = \sum_{n\in\mathcal{I}} n \leq 2N. \tag{103}$$