Adaptive Newton Method for Empirical Risk Minimization to Statistical Accuracy

Aryan Mokhtari
University of Pennsylvania
aryanm@seas.upenn.edu

Hadi Daneshmand
ETH Zurich, Switzerland
hadi.daneshmand@inf.ethz.ch

Aurelien Lucchi
ETH Zurich, Switzerland
aurelien.lucchi@inf.ethz.ch

Thomas Hofmann
ETH Zurich, Switzerland
thomas.hofmann@inf.ethz.ch

Alejandro Ribeiro
University of Pennsylvania
aribeiro@seas.upenn.edu

Abstract

We consider empirical risk minimization for large-scale datasets. We introduce Ada Newton as an adaptive algorithm that uses Newton’s method with adaptive sample sizes. The main idea of Ada Newton is to increase the size of the training set by a factor larger than one in a way that the minimization variable for the current training set is in the local neighborhood of the optimal argument of the next training set. This allows to exploit the quadratic convergence property of Newton’s method and reach the statistical accuracy of each training set with only one iteration of Newton’s method. We show theoretically that we can iteratively increase the sample size while applying single Newton iterations without line search and staying within the statistical accuracy of the regularized empirical risk. In particular, we can double the size of the training set in each iteration when the number of samples is sufficiently large. Numerical experiments on various datasets confirm the possibility of increasing the sample size by factor 2 at each iteration which implies that Ada Newton achieves the statistical accuracy of the full training set with about two passes over the dataset.

1 Introduction

A hallmark of empirical risk minimization (ERM) on large datasets is that evaluating descent directions requires a complete pass over the dataset. Since this is undesirable due to the large number of training samples, stochastic optimization algorithms with descent directions estimated from a subset of samples are the method of choice. First order stochastic optimization has a long history [19, 17] but the last decade has seen fundamental progress in developing alternatives with faster convergence. A partial list of this consequential literature includes Nesterov acceleration [16, 2], stochastic averaging gradient [20, 6], variance reduction [10, 26], and dual coordinate methods [23, 24].

When it comes to stochastic second order methods the first challenge is that while evaluation of Hessians is as costly as evaluation of gradients, the stochastic estimation of Hessians has proven more challenging. This difficulty is addressed by incremental computations in [9] and subsampling in [7] or circumvented altogether in stochastic quasi-Newton methods [21, 12, 13, 11, 14]. Despite this incipient progress it is nonetheless fair to say that the striking success in developing stochastic first order methods is not matched by equal success in the development of stochastic second order methods. This is because even if the problem of estimating a Hessian is solved there are still four challenges left in the implementation of Newton-like methods in ERM:

1 The first two authors have contributed equally in this work.
(i) Global convergence of Newton’s method requires implementation of a line search subroutine and line searches in ERM require a complete pass over the dataset.

(ii) The quadratic convergence advantage of Newton’s method manifests close to the optimal solution but there is no point in solving ERM problems beyond their statistical accuracy.

(iii) Newton’s method works for strongly convex functions but loss functions are not strongly convex for many ERM problems of practical importance.

(iv) Newton’s method requires inversion of Hessians which is costly in large dimensional ERM.

Because stochastic Newton-like methods can’t use line searches [cf. (i)], must work on problems that may be not strongly convex [cf. (iii)], and never operate very close to the optimal solution [cf. (ii)], they never experience quadratic convergence. They do improve convergence constants and, if efforts are taken to mitigate the cost of inverting Hessians [cf. (iv)] as in [21, 12, 7, 18] they result in faster convergence. But since they still converge at linear rates they do not enjoy the foremost benefits of Newton’s method.

In this paper we attempt to circumvent (i)-(iv) with the Ada Newton algorithm that combines the use of Newton iterations with adaptive sample sizes [5]. Say the total number of available samples is $N$, consider subsets of $n \leq N$ samples, and suppose the statistical accuracy of the ERM associated with $n$ samples is $V_n$ (Section 2). In Ada Newton we add a quadratic regularization term of order $V_n$ to the empirical risk — so that the regularized risk also has statistical accuracy $V_n$ — and assume that for a certain initial sample size $m_0$, the problem has been solved to its statistical accuracy $V_{m_0}$. The sample size is then increased by a factor $\alpha > 1$ to $n = \alpha m_0$. We proceed to perform a single Newton iteration with unit stepsize and prove that the result of this update solves this extended ERM problem to its statistical accuracy (Section 3). This permits a second increase of the sample size by a factor $\alpha$ and a second Newton iteration that is likewise guaranteed to solve the problem to its statistical accuracy. Overall, this permits minimizing the empirical risk in $\alpha/(\alpha - 1)$ passes over the dataset and inverting $\log_\alpha N$ Hessians. Our theoretical results provide a characterization of the values of $\alpha$ that are admissible with respect to different problem parameters (Theorem 1). In particular, we show that asymptotically on the number of samples $n$ and with proper parameter selection we can set $\alpha = 2$ (Proposition 2). In such case we can optimize to within statistical accuracy in about 2 passes over the dataset and after inversion of about $3.32 \log_{10} N$ Hessians. Our numerical experiments verify that $\alpha = 2$ is a valid factor for increasing the size of the training set at each iteration while performing a single Newton iteration for each value of the sample size.

2 Empirical risk minimization

We aim to solve ERM problems to their statistical accuracy. To state this problem formally consider an argument $w \in \mathbb{R}^p$, a random variable $Z$ with realizations $z$ and a convex loss function $f(w; z)$. We want to find an argument $w^*$ that minimizes the statistical average loss $L(w) := \mathbb{E}_Z[f(w, Z)]$.

$$w^* := \arg\min_w L(w) = \arg\min_w \mathbb{E}_Z[f(w, Z)]. \quad (1)$$

The loss in (1) can’t be evaluated because the distribution of $Z$ is unknown. We have, however, access to a training set $T = \{z_1, \ldots, z_N\}$ containing $N$ independent samples $z_1, \ldots, z_N$ that we can use to estimate $L(w)$. We therefore consider a subset $S_n \subseteq T$ and settle for minimization of the empirical risk $L_n(w) := (1/n) \sum_{k=1}^n f(w, z_k)$.

$$w_n^* := \arg\min_w L_n(w) = \arg\min_w \frac{1}{n} \sum_{k=1}^n f(w, z_k), \quad (2)$$

where, without loss of generality, we have assumed $S_n = \{z_1, \ldots, z_n\}$ contains the first $n$ elements of $T$. The difference between the empirical risk in (2) and the statistical loss in (1) is a fundamental quantities in statistical learning. We assume here that there exists a constant $V_n$, which depends on the number of samples $n$, that upper bounds their difference for all $w$ with high probability (w.h.p),

$$\sup_w |L(w) - L_n(w)| \leq V_n, \quad \text{w.h.p.} \quad (3)$$

That the statement in (3) holds with w.h.p means that there exists a constant $\delta$ such that the inequality holds with probability at least $1 - \delta$. The constant $V_n$ depends on $\delta$ but we keep that dependency
are convex and their gradients are Lipschitz continuous with constant $L$. Assumption 3. Assumption 2. Assumption 1. over, their gradients are satisfied.

$w$ extra computation. The main contribution of this paper is to derive a condition that guarantees that $w$ is also of order $O(n^{1/2})$. Indeed, if we find a variable $w$ for which $L_n(\epsilon_n^*) - L_n(w^*) \leq V_n$, finding a better approximation of $w^*$ is moot because (3) implies that this is not necessarily a better approximation of the minimizer $w$ of the statistical loss. We say the variable $w_n$ solves the ERM problem in (2) to within its statistical accuracy. In particular, this implies that adding a regularization of order $V_n$ to (2) yields a problem that is essentially equivalent. We can then consider a quadratic regularizer of the form $cV_n/2\|w\|^2$ to define the regularized empirical risk $R_n(w) := L_n(w) + (cV_n/2)\|w\|^2$ and the corresponding optimal argument

$$w_n^* := \arg\min_w R_n(w) = \arg\min_w L_n(w) + \frac{cV_n}{2}\|w\|^2. \quad (4)$$

Since the regularization in (4) is of order $V_n$ and (3) holds, the difference between $R_n(w_n^*)$ and $L(w^*)$ is also of order $V_n$ – this may be not as immediate as it seems; see [22]. Thus, we can say that a variable $w_n$ satisfying $R_n(w_n) - R_n(w_n^*) \leq V_n$ solves the ERM problem to within its statistical accuracy. We accomplish this goal in this paper with the Ada Newton algorithm which we introduce in the following section.

### 3 Ada Newton

To solve (4) suppose the problem has been solved to within its statistical accuracy for a set $\mathcal{S}_m \subset \mathcal{S}_n$ with $m = n/\alpha$ samples where $\alpha > 1$. Therefore, we have found a variable $w_m$ for which $R_m(w_m) - R_m(w_m^*) \leq V_m$. Our goal is to update $w_m$ using the Newton step in a way that the updated variable $w_n$ estimates $w_n^*$ with accuracy $V_n$. To do so compute the gradient of the risk $R_n$ evaluated at $w_m$

$$\nabla R_n(w_m) = \frac{1}{n} \sum_{k=1}^n \nabla f(w_m, z_k) + cV_n w_m, \quad (5)$$

as well as the Hessian $H_n$ of $R_n$ evaluated at $w_m$

$$H_n := \nabla^2 R_n(w_m) = \frac{1}{n} \sum_{k=1}^n \nabla^2 f(w_m, z_k) + cV_n I, \quad (6)$$

and update $w_m$ with the Newton step of the regularized risk $R_n$ to compute

$$w_n = w_m - H_n^{-1} \nabla R_n(w_m). \quad (7)$$

Note that the stepsize of the Newton update in (7) is 1, which avoids line search algorithms requiring extra computation. The main contribution of this paper is to derive a condition that guarantees that $w_n$ solves $R_n$ to within its statistical accuracy $V_n$. To do so, we first assume the following conditions are satisfied.

**Assumption 1.** The loss functions $f(w, z)$ are convex with respect to $w$ for all values of $z$. Moreover, their gradients $\nabla f(w, z)$ are Lipschitz continuous with constant $M$

$$\|\nabla f(w, z) - \nabla f(w', z)\| \leq M\|w - w'\|, \quad \text{for all } z. \quad (8)$$

**Assumption 2.** The loss functions $f(w, z)$ are self-concordant with respect to $w$ for all $z$.

**Assumption 3.** The difference between the gradients of the empirical loss $L_n$ and the statistical average loss $L$ is bounded by $V_n^{1/2}$ for all $w$ with high probability,

$$\sup_w \|\nabla L(w) - \nabla L_n(w)\| \leq V_n^{1/2}, \quad \text{w.h.p.} \quad (9)$$

The conditions in Assumption 1 imply that the average loss $L(w)$ and the empirical loss $L_n(w)$ are convex and their gradients are Lipschitz continuous with constant $M$. Thus, the empirical risk
Algorithm 1 Ada Newton

1: **Parameters:** Sample size increase constants $\alpha_0 > 1$ and $0 < \beta < 1$.
2: **Input:** Initial sample size $n = m_0$ and argument $w_n = w_{m_0}$ with $\|\nabla R_n(w_n)\| < (\sqrt{2c})V_n$
3: **while** $n \leq N$ **do** (**main loop**)
4: \hspace{1em} Update argument and index: $w_m = w_n$ and $m = n$. Reset factor $\alpha = \alpha_0$.
5: \hspace{1em} **repeat** (**sample size backtracking loop**)
6: \hspace{2em} Increase sample size: $n = \min\{\alpha m, N\}$.
7: \hspace{2em} Compute gradient [cf. (5)]: $\nabla R_n(w_m) = (1/n) \sum_{k=1}^n \nabla f(w_m, z_k) + cV_n w_m$
8: \hspace{2em} Compute Hessian [cf. (6)]: $H_n = (1/n) \sum_{k=1}^n \nabla^2 f(w_m, z_k) + cV_n I$
9: \hspace{2em} Newton Update [cf. (7)]: $w_n = w_m - H_n^{-1} \nabla R_n(w_m)$
10: \hspace{2em} Compute gradient [cf. (5)]: $\nabla R_n(w_n) = (1/n) \sum_{k=1}^n \nabla f(w_n, z_k) + cV_n w_n$
11: \hspace{2em} Backtrack sample size increase $\alpha = \beta \alpha$.
12: **until** $\|\nabla R_n(w_n)\| < (\sqrt{2c})V_n$
13: **end while**

$R_n(w)$ is strongly convex with constant $cV_n$ and its gradients $\nabla R_n(w)$ are Lipschitz continuous with parameter $M + cV_n$. Likewise, the condition in Assumption 2 implies that the average loss $L_n(w)$, the empirical loss $L_n(w)$, and the empirical risk $R_n(w)$ are also self-concordant. The condition in Assumption 3 says that the gradients of the empirical risk converge to their statistical average at a rate of order $V_n^{1/2}$. If the constant $V_n$ in condition (3) is of order not faster than $O(1/n)$ the condition in Assumption 3 holds if the gradients converge to their statistical average at a rate of order $V_n^{1/2} = O(1/\sqrt{n})$. This is a conservative rate for the law of large numbers.

In the following theorem, given Assumptions 1-3, we state a condition that guarantees the variable $w_n$ evaluated as in (7) solves $R_n$ to within its statistical accuracy $V_n$.

**Theorem 1.** Consider the variable $w_m$ as a $V_m$-optimal solution of the risk $R_m$, i.e., a solution such that $R_m(w_m) - R_m(w_m^*) \leq V_m$. Let $n = \alpha m > m$, consider the risk $R_n$ associated with sample set $S_n \supset S_m$, and suppose assumptions 1 - 3 hold. If the sample size $n$ is chosen such that

$$
\left( \frac{2(M + cV_m)V_m}{cV_n} \right)^{1/2} + \frac{2(n - m)}{nc^{1/2}} + \frac{((2 + \sqrt{2})c^{1/2} + c\|w^*\|)(V_m - V_n)}{(cV_n)^{1/2}} \leq \frac{1}{4}
$$

and

$$
144 \left( V_m + \frac{2(n - m)}{n} (V_{n-m} + V_m) \right) + 2(V_m - V_n) + \frac{c(V_m - V_n)}{2} \|w^*\|^2 \leq V_n
$$

are satisfied, then the variable $w_n$, which is the outcome of applying one Newton step on the variable $w_m$ as in (7), has sub-optimality error $V_n$ with high probability, i.e.,

$$R_n(w_n) - R_n(w_n^*) \leq V_n, \quad w.h.p.
$$

**Proof.** See Section 4.

**Proposition 2.** Consider a learning problem in which the statistical accuracy satisfies $V_m \leq \alpha V_n$ for $n = \alpha m$ and $\lim_{n \to \infty} V_n = 0$. If the regularization constant $c$ is chosen so that

$$
\left( \frac{2\alpha M}{c} \right)^{1/2} + \frac{2(\alpha - 1)}{\alpha c^{1/2}} \leq \frac{1}{4}
$$

then, there exists a sample size $\bar{m}$ such that (10) and (11) are satisfied for all $m > \bar{m}$ and $n = \alpha m$.

In particular, if $\alpha = 2$ we can satisfy (10) and (11) with $c > 16(2\sqrt{M} + 1)^2$. 

4
Proposition 3. Consider the sets $S_m$ and $S_n$ as subsets of the training set $T$ such that $S_m \subset S_n \subset T$. We assume that the number of samples in the sets $S_m$ and $S_n$ are $m$ and $n$, respectively. Further, define $\mathbf{w}_m$ as an $V_m$ optimal solution of the risk $R_m$, i.e., $R_m(\mathbf{w}_m) - R_m(\mathbf{w}_n^*) \leq V_m$. In addition, define $\lambda_n(\mathbf{w}) := (\nabla R_n(\mathbf{w})^T \nabla^2 R_n(\mathbf{w})^{-1} \nabla R_n(\mathbf{w}))^{1/2}$ as the Newton decrement of variable $\mathbf{w}$.
associated with the risk $R_n$. If Assumption 1-3 hold, then Newton’s method at point $w_m$ is in the quadratic convergence phase for the objective function $R_n$, i.e., $\lambda_n(w_m) < 1/4$, if we have
\[
\left(\frac{2(M + cV_m)V_m}{cV_n}\right)^{1/2} + \frac{(2(n - m)/n)V_n^{1/2} + (\sqrt{2c} + 2\sqrt{c} + c\|w^*\|)(V_m - V_n)}{(cV_n)^{1/2}} \leq \frac{1}{4} \text{ w.h.p.}
\]
(15)

Proof. See Section 7.1 in the supplementary material.

From the analysis of Newton’s method we know that if the Newton decrement $\lambda_n(w)$ is smaller than 1/4, the variable $w$ is in the local neighborhood of Newton’s method; see e.g., Chapter 9 of [4]. From the result in Proposition 3, we obtain a sufficient condition to guarantee that $\lambda_n(w_m) < 1/4$ which implies that $w_m$, which is a $V_m$ optimal solution for the regularized empirical loss $R_m$, i.e., $R_m(w_m) - R_m(w_m^*) \leq V_m$, is in the local neighborhood of the optimal argument of $R_n$ that Newton’s method converges quadratically.

Unfortunately, the quadratic convergence of Newton’s method for self-concordant functions is in terms of the Newton decrement $\lambda_n(w)$ and it does not necessary guarantee quadratic convergence in terms of objective function error. To be more precise, we can show that $\lambda_n(w_m) \leq \gamma \lambda_n(w_m)^2$; however, we can not conclude that the quadratic convergence of Newton’s method implies $R_n(w_n) - R_n(w_n^*) \leq (\gamma R_n(w_m) - R_n(w_m^*))^2$. In the following proposition we try to characterize an upper bound for the error $R_n(w_n) - R_n(w_n^*)$ in terms of the squared error $(R_n(w_m) - R_n(w_m^*))^2$ using the quadratic convergence property of Newton decrement.

**Proposition 4.** Consider $w_m$ as a variable that is in the local neighborhood of the optimal argument of the risk $R_n$, where Newton’s method has a quadratic convergence rate, i.e., $\lambda_n(w_m) \leq 1/4$. Recall the definition of the variable $w_n$ in (7) as the updated variable using Newton step. If Assumption 1 and 2 hold, then the difference $R_n(w_n) - R_n(w_n^*)$ is upper bounded by
\[
R_n(w_n) - R_n(w_n^*) \leq 144(R_n(w_m) - R_n(w_m^*))^2.
\]
(16)

Proof. See Section 7.2 in the supplementary material.

The result in Proposition 4 provides an upper bound for the sub-optimality $R_n(w_m) - R_n(w_m^*)$ in terms of the sub-optimality of variable $w_m$ for the risk $R_n$, i.e., $R_n(w_m) - R_n(w_m^*)$. Recall that we know that $w_m$ is in the statistical accuracy of $R_m$, i.e., $R_m(w_m) - R_m(w_m^*) \leq V_m$, and we aim to show that the updated variable $w_n$ stays in the statistical accuracy of $R_n$, i.e., $R_n(w_n) - R_n(w_n^*) \leq V_n$. This can be done by showing that the upper bound for $R_n(w_m) - R_n(w_m^*)$ in (16) is smaller than $V_n$. We proceed to derive an upper bound for the sub-optimality $R_n(w_m) - R_n(w_m^*)$ in the following proposition.

**Proposition 5.** Consider the sets $S_m$ and $S_n$ as subsets of the training set $T$ such that $S_m \subset S_n \subset T$. We assume that the number of samples in the sets $S_m$ and $S_n$ are $m$ and $n$, respectively. Further, define $w_m$ as an $V_m$ optimal solution of the risk $R_m$, i.e., $R_m(w_m) - R_m(w_m^*) \leq V_m$. If Assumption 1-3 hold, then the empirical risk error $R_n(w_m) - R_n(w_m^*)$ of the variable $w_m$ corresponding to the set $S_n$ is bounded above by
\[
R_n(w_m) - R_n(w_m^*) \leq V_m + \frac{2(n - m)}{n} (V_m - V_n) + 2(V_m - V_n) + \frac{c(V_m - V_n)}{2} \|w^*\|^2 \text{ w.h.p.}
\]
(17)

Proof. See Section 7.3 in the supplementary material.

The result in Proposition 5 characterizes the sub-optimality of the variable $w_m$, which is an $V_m$ sub-optimal solution for the risk $R_m$, with respect to the empirical risk $R_n$ associated with the set $S_n$.

The results in Proposition 3, 4, and 5 lead to the result in Theorem 1. To be more precise, from the result in Proposition 3 we obtain that the condition in (10) implies that $w_m$ is in the local neighborhood of the optimal argument of $R_n$ and $\lambda_n(w_m) \leq 1/4$. Hence, the hypothesis of Proposition 4 is satisfied and we have $R_n(w_m) - R_n(w_m^*) \leq 144(R_n(w_m) - R_n(w_m^*))^2$. This result paired with the result in Proposition 5 shows that if the condition in (11) is satisfied we can conclude that $R_n(w_n) - R_n(w_n^*) \leq V_n$ which completes the proof of Theorem 1.
5 Experiments

In this section, we study the performance of Ada Newton and compare it with state-of-the-art in solving a large-scale classification problem. In the main paper we only use the protein homology dataset provided on KDD cup 2004 website. Further numerical experiments on various datasets can be found in Section 7.4 in the supplementary material. The protein homology dataset contains $N = 145751$ samples and the dimension of each sample is $p = 74$. We consider three algorithms to compare with the proposed Ada Newton method. One of them is the classic Newton’s method with backtracking line search. The second algorithm is Stochastic Gradient Descent (SGD) and the last one is the SAGA method introduced in [6]. In our experiments, we use logistic loss and set the regularization parameters as $c = 200$ and $V_n = 1/n$.

The stepsize of SGD in our experiments is $2 \times 10^{-2}$. Note that picking larger stepsize leads to faster but less accurate convergence and choosing smaller stepszie improves the accuracy convergence with the price of slower convergence rate. The stepsize for SAGA is hand-optimized and the best performance has been observed for $\alpha = 0.2$ which is the one that we use in the experiments. For Newton’s method, the backtracking line search parameters are $\alpha = 0.4$ and $\beta = 0.5$. In the implementation of Ada Newton we increase the size of the training set by factor 2 at each iteration, i.e., $\alpha = 2$ and we observe that the condition $\| \nabla R_n(w_n) \| < (\sqrt{2c})V_n$ is always satisfied and there is no need for reducing the factor $\alpha$. Moreover, the size of initial training set is $m_0 = 124$. For the warmup step that we need to get into to the quadratic neighborhood of Newton’s method we use the gradient descent method. In particular, we run gradient descent with stepsize $10^{-5}$ for 100 iterations. Note that since the number of samples is very small at the beginning, $m_0 = 124$, and the regularizer is very large, the condition number of problem is very small. Thus, gradient descent is able to converge to a good neighborhood of the optimal solution in a reasonable time. Notice that the computation of this warm up process is very low and is equal to 12400 gradient evaluations. This number of samples is less than 10% of the full training set. In other words, the cost is less than 10% of one pass over the dataset. Although, this cost is negligible, we consider it in comparison with SGD, SAGA, and Newton’s method. We would like to mention that other algorithms such as Newton’s method and stochastic algorithms can also be used for the warm up process; however, the gradient descent method sounds the best option since the gradient evaluation is not costly and the problem is well-conditioned for a small training set.

The left plot in Figure 1 illustrates the convergence path of SGD, SAGA, Newton, and Ada Newton for the protein homology dataset. Note that the $x$ axis is the total number of samples used divided by the size of the training set $N = 145751$ which we call number of passes over the dataset. As we observe, The best performance among the four algorithms belongs to Ada Newton. In particular, Ada Newton is able to achieve the accuracy of $R_N(w) - R_N^* < 1/N$ by 2.4 passes over the dataset which is very close to theoretical result in Theorem 1 that guarantees accuracy of order $O(1/N)$ after $\alpha/(\alpha - 1) = 2$ passes over the dataset. To achieve the same accuracy of $1/N$ Newton’s method requires 7.5 passes over the dataset, while SAGA needs 10 passes. The SGD algorithm can not achieve the statistical accuracy of order $O(1/N)$ even after 25 passes over the dataset.

Although, Ada Newton and Newton outperform SAGA and SGD, their computational complexity are different. We address this concern by comparing the algorithms in terms of runtime. The right
plot in Figure 1 demonstrates the convergence paths of the considered methods in terms of runtime. As we observe, Newton’s method requires more time to achieve the statistical accuracy of $1/N$ relative to SAGA. This observation justifies the belief that Newton’s method is not practical for large-scale optimization problems, since by enlarging $p$ or making the initial solution worse the performance of Newton’s method will be even worse than the ones in Figure 1. Ada Newton resolves this issue by starting from small sample size which is computationally less costly. Ada Newton also requires Hessian inverse evaluations, but the number of inversions is proportional to $\log N$. Moreover, the performance of Ada Newton doesn’t depend on the initial point and the warm up process is not costly as we described before. We observe that Ada Newton outperforms SAGA significantly. In particular it achieves the statistical accuracy of $1/N$ in less than 25 seconds, while SAGA achieves the same accuracy in 62 seconds. Note that since the variable $w_N$ is in the quadratic neighborhood of Newton’s method for $R_N$ the convergence path of Ada Newton becomes quadratic eventually when the size of the training set becomes equal to the size of the full dataset. It follows that the advantage of Ada Newton with respect to SAGA is more significant if we look for a sub-optimality less than $V_n$. We have observed similar performances for other datasets such as A9A, W8A, COVTYPE, and SUSY – see Section 7.4 in the supplementary material.

6 Discussions

As explained in Section 4, Theorem 1 holds because condition (10) makes $w_m$ part of the quadratic convergence region of $R_n$. From this fact, it follows that the Newton iteration makes the suboptimality gap $R_n(w_n) - R_n(w^*_n)$ the square of the suboptimality gap $R_n(w_m) - R_n(w^*_n)$. This yields condition (11) and is the fact that makes Newton steps valuable in increasing the sample size. If we replace Newton iterations by any method with linear convergence rate, the orders of both sides on condition (11) are the same. This would make aggressive increase of the sample size unlikely.

In Section 1 we pointed out four reasons that challenge the development of stochastic Newton methods. It would not be entirely accurate to call Ada Newton a stochastic method because it doesn’t rely on stochastic descent directions. It is, nonetheless, a method for ERM that makes pithy use of the dataset. The challenges listed in Section 1 are overcome by Ada Newton because:

(i) Ada Newton does not use line searches. Optimality improvement is guaranteed by increasing the sample size.

(ii) The advantages of Newton’s method are exploited by increasing the sample size at a rate that keeps the solution for sample size $m$ in the quadratic convergence region of the risk associated with sample size $n = \alpha m$. This allows aggressive growth of the sample size.

(iii) The ERM problem is not necessarily strongly convex. A regularization of order $V_n$ is added to construct the empirical risk $R_n$.

(iv) Ada Newton inverts approximately $\log N$ Hessians. To be more precise, the total number of inversion could be larger than $\log N$ because of the backtracking step. However, the backtracking step is bypassed when the number of samples is sufficiently large.

It is fair to point out that items (ii) and (iv) are true only to the extent that the damped phase in Algorithm 1 is not significant. Our numerical experiments indicate that this is true but the conclusion is not warranted by out theoretical bounds except when the dataset is very large. This suggests the bounds are loose and that further research is warranted to develop tighter bounds.

References


