Convex optimization based on global lower second-order models

Nikita Doikov[*]
Catholic University of Louvain,
Louvain-la-Neuve, Belgium
Nikita.Doikov@uclouvain.be

Yurii Nesterov[†]
Catholic University of Louvain,
Louvain-la-Neuve, Belgium
Yurii.Nesterov@uclouvain.be

Abstract

In this paper, we present new second-order algorithms for composite convex optimization, called Contracting-domain Newton methods. These algorithms are affine-invariant and based on global second-order lower approximation for the smooth component of the objective. Our approach has an interpretation both as a second-order generalization of the conditional gradient method, or as a variant of trust-region scheme. Under the assumption, that the problem domain is bounded, we prove $\mathcal{O}(1/k^2)$ global rate of convergence in functional residual, where k is the iteration counter, minimizing convex functions with Lipschitz continuous Hessian. This significantly improves the previously known bound $\mathcal{O}(1/k)$ for this type of algorithms. Additionally, we propose a stochastic extension of our method, and present computational results for solving empirical risk minimization problem.

1 Introduction

Classical Newton method is one of the most popular optimization schemes for solving ill-conditioned problems. The method has very fast quadratic convergence, provided that the starting point is sufficiently close to the optimum [3, 22, 31]. However, the questions related to its global behaviour for a wide class of functions are still open, being in the area of active research.

The significant progress in this direction was made after [33], where Cubic regularization of Newton method with its global complexity bounds were justified. The main idea of [33] is to use a global *upper* approximation model of the objective, which is the second-order Taylor's polynomial augmented by a cubic term. The next point in the iteration process is defined as the minimizer of this model. Cubic Newton attains global convergence for convex functions with Lipschitz continuous Hessian. The rate of convergence in functional residual is of the order $\mathcal{O}(1/k^2)$ (here and later on, k is the iteration counter). This is much faster than the classical $\mathcal{O}(1/k)$ rate of the Gradient Method [31]. Later on, accelerated [27], adaptive [7, 8] and universal [17, 12, 18] second-order schemes based on cubic regularization were developed. Randomized versions of Cubic Newton, suitable for solving high-dimensional problems were proposed in [13, 19].

Another line of results on global convergence of Newton method is mainly related to the framework of self-concordant functions [32, 31]. This class is affine-invariant. From the global perspective, it provides us with an *upper* second-order approximation of the objective, which naturally leads to the Damped Newton Method. Several new results are related to its analysis for generalized self-concordant functions [2, 38], and the notion of Hessian stability [23]. However, for more refined problem classes, we can often obtain much better complexity estimates, by using the cubic regularization technique [14].

[*] Institute of Information and Communication Technologies, Electronics and Applied Mathematics (ICTEAM)
[†] Center for Operations Research and Econometrics (CORE)

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

In this paper, we investigate a different approach, which is motivated by a new global second-order *lower* model of the objective function, introduced in Section 3.

We incorporate this model into a new second-order optimization algorithm, called Contracting-Domain Newton Method (Section 4). At every iteration, it minimizes a lower approximation of the smooth component of the objective, augmented by a composite term. The next point is defined as a convex combination of the minimizer, and the previous point. By its nature, it is similar to the scheme of Conditional Gradient Method (or, Frank-Wolfe algorithm, [15, 30]). Under assumption of boundedness of the problem domain, for convex functions with Hölder continuous Hessian of degree $\nu \in [0, 1]$, we establish its $\mathcal{O}(1/k^{1+\nu})$ global rate of convergence in functional residual. In the case $\nu = 1$, for the class of convex function with Lipschitz continuous Hessian, this gives $\mathcal{O}(1/k^2)$ rate of convergence. As compared with Cubic Newton, the new method is affine-invariant and universal, since it does not depend on the norms and parameters of the problem class. When the composite component is strongly convex (with respect to arbitrary norm), we show $\mathcal{O}(1/k^{2+2\nu})$ rate for a universal scheme. If the parameters of problem class are known, we can prove a global linear convergence. We also provide different trust-region interpretations for our algorithm.

In Section 5, we present aggregated models, which accumulate second-order information into quadratic Estimating Functions [31]. This leads to another optimization process, called Aggregating Newton Method, with the global convergence of the same order $\mathcal{O}(1/k^{1+\nu})$ as for general convex case. The latter method can be seen as a second-order counterpart of the dual averaging gradient schemes [28, 29].

In Section 6, we consider the problem of finite-sum minimization. We propose stochastic extensions of our method. During the iterations of the basic variant, we need to increase the batch size for randomized estimates of gradients and Hessians up to the order $\mathcal{O}(k^4)$ and $\mathcal{O}(k^2)$ respectively. Using the *variance reduction* technique for the gradients, we reduce the batch size up to the level $\mathcal{O}(k^2)$ for both estimates. At the same time, the global convergence rate of the resulting methods is of the order $\mathcal{O}(1/k^2)$, as for general convex functions with Lipschitz continuous Hessian.

Section 7 contains numerical experiments. Section 8 contains some final remarks. All necessary proofs are provided in the supplementary material.

2 Problem formulation and notations

Our goal is to solve the following *composite* convex minimization problem:

$$\min_x F(x) \quad := \quad f(x) + \psi(x), \tag{1}$$

where $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a *simple* proper closed convex function, and function f is convex and twice continuously differentiable at every point $x \in \mathrm{dom}\, \psi$. Let us fix an arbitrary (possibly non-Euclidean) norm $\| \cdot \|$ on $\mathbb{R}^n$. We denote by D the corresponding diameter of $\mathrm{dom}\, \psi$:

$$D \quad := \quad \sup_{x, y \in \mathrm{dom}\, \psi} \|x - y\|. \tag{2}$$

Our main assumption on problem (1) is that $\mathrm{dom}\, \psi$ is bounded:

$$D \quad < \quad +\infty. \tag{3}$$

The most important example of ψ is $\{0, +\infty\}$-indicator of a simple compact convex set $Q = \mathrm{dom}\, \psi$. In particular, for a ball in $\| \cdot \|_p$-norm with $p \geq 1$, this is

$$\psi(x) \quad = \quad \begin{cases} 0, & \|x\|_p := \Bigl(\sum_{i=1}^n |x^{(i)}|^p \Bigr)^{1/p} \leq \frac{D}{2}, \\ +\infty, & \text{else.} \end{cases} \tag{4}$$

From the machine learning perspective, D is usually considered as a *regularization parameter* in this setting. We denote by $\langle \cdot, \cdot \rangle$ the standard scalar product of two vectors, $x, y \in \mathbb{R}^n$:

$$\langle x, y \rangle \quad := \quad \sum_{i=1}^n x^{(i)} y^{(i)}.$$

For function f, we denote its gradient by $\nabla f(x) \in \mathbb{R}^n$, and its Hessian matrix by $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$. Having fixed the norm $\| \cdot \|$ for primal variables $x \in \mathbb{R}^n$, the *dual* norm can be defined in the standard way:

$$\|s\|_* \quad := \quad \sup_{h \in \mathbb{R}^n : \|h\| \leq 1} \langle s, h \rangle.$$

The dual norm is necessary for measuring the size of gradients. For a matrix $A \in \mathbb{R}^{n \times n}$, we use the corresponding induced operator norm, defined as

$$\|A\| \quad := \quad \sup_{h \in \mathbb{R}^n : \|h\| \leq 1} \|Ah\|_*.$$

3 Second-order lower model of objective function

To characterize the complexity of problem (1), we need to introduce some assumptions on the growth of derivatives. Let us assume that the Hessian of f is Hölder continuous of degree $\nu \in [0,1]$ on $\mathrm{dom}\, \psi$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \quad \leq \quad H_\nu \|x - y\|^\nu, \qquad x, y \in \mathrm{dom}\, \psi. \tag{5}$$

The actual parameters of this problem class may be unknown. However, we assume that for *some* $\nu \in [0,1]$ inequality (5) is satisfied with corresponding constant $0 \leq H_\nu < +\infty$. The direct consequence of (5) is the following global bounds for Taylor's approximation, for all $x, y \in \mathrm{dom}\, \psi$

$$\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x)\|_* \quad \leq \quad \tfrac{H_\nu \|y-x\|^{1+\nu}}{1+\nu}, \tag{6}$$

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle - \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y - x\rangle| \quad \leq \quad \tfrac{H_\nu \|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)}. \tag{7}$$

Recall, that in addition to (5), we assume that f is *convex*:

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x \rangle, \qquad x, y \in \mathrm{dom}\, \psi. \tag{8}$$

Employing both smoothness and convexity, we are able to enhance this global lower bound, as follows.

Lemma 1 *For all $x, y \in \mathrm{dom}\, \psi$ and $t \in [0,1]$, it holds*

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{t}{2}\langle \nabla^2 f(x)(y - x), y - x\rangle - \tfrac{t^{1+\nu} H_\nu \|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)}. \tag{9}$$

Note that the right-hand side of (9) is concave in $t \in [0,1]$, and for $t = 0$ we obtain the standard first-order lower bound. The maximization of (9) over t gives

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{\bar{\gamma}_{x,y}}{2}\langle \nabla^2 f(x)(y-x), y - x\rangle, \tag{10}$$

with

$$\bar{\gamma}_{x,y} \quad := \quad \tfrac{\nu}{1+\nu} \min\Bigl\{1, \tfrac{(2+\nu)\langle \nabla^2 f(x)(y-x), y-x\rangle}{2H_\nu \|y-x\|^{2+\nu}}\Bigr\}^{\frac{1}{\nu}}, \quad x \neq y, \quad \nu \in (0,1].$$

Thus, (10) is always *tighter* than (8), employing additional *global second-order information*. The relationship between them is shown on Figure 1. Hence, it seems natural to incorporate the second-order lower bounds into optimization schemes.

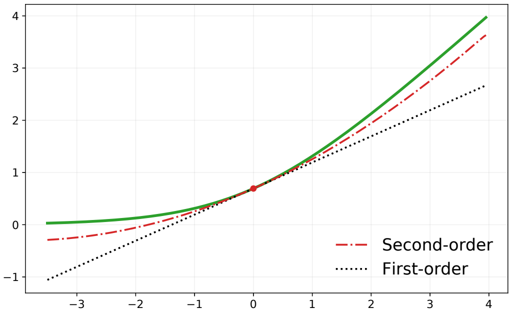

Figure 1: Global lower bounds for logistic regression loss, $f(x) = \log(1 + \exp(x))$.

4 Contracting-Domain Newton Methods

Let us introduce a general scheme of *Contracting-Domain Newton Method*, which is based on global second-order lower bounds. Note, that the right hand side of (10) is nonconvex in y. Hence, it can hardly be used directly in a computational algorithm. To tackle this issue, we use a sequence of contracting coefficients $\{\gamma_k\}_{k \geq 0}$. Each coefficient $\gamma_k \in (0, 1]$ can be seen as an appropriate substitute of $\bar{\gamma}_{x,y}$ in (10). Then, we minimize the corresponding global lower bound augmented by the composite component $\psi(\cdot)$. The next point is taken as a convex combination of the minimizer and the current point. Let us present this method formally, as Algorithm 1.

```
Algorithm 1: Contracting-Domain Newton Method, I
Initialization. Choose $x_0 \in \mathrm{dom}\,\psi$.
Iteration $k \geq 0$.
  1: Pick up $\gamma_k \in (0, 1]$.
  2: Compute
        $v_{k+1} \in \mathop{\mathrm{Argmin}}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{\gamma_k}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \psi(y) \Big\}$.
  3: Set $x_{k+1} := x_k + \gamma_k (v_{k+1} - x_k)$.
```

There is a clear connection of this method with Frank-Wolfe algorithm, [15]. Indeed, instead of the standard first-order approximation (8), we use the lower global quadratic model. Thus, as compared with the gradient methods, every iteration of Algorithm 1 is more expensive. However, this is a standard situation with the second-order schemes (see the below discussion on the iteration complexity). At the same time, our method is *affine-invariant*, since it does not depend on the norms.

It is clear, that for $\gamma_k \equiv 1$ we obtain iterations of the classical Newton method. Its local quadratic convergence for composite optimization problems was established in [26]. However, for the global convergence, we need to adjust the contracting coefficients accordingly. To state the global convergence result, let us introduce the following linear *Estimating Functions* (see [31]):

$$\phi_k(x) \overset{\text{def}}{=} \textstyle\sum_{i=1}^k a_i \big[f(x_i) + \langle \nabla f(x_i), x - x_i \rangle + \psi(x) \big], \qquad \phi_k^* := \min_x \phi_k(x), \tag{11}$$

for the sequence of test points $\{x_k : x_k \in \mathrm{dom}\,\psi\}_{k \geq 1}$ and positive scaling coefficients $\{a_k\}_{k \geq 1}$. We relate them with contracting coefficients, as follows

$$\gamma_k := \frac{a_{k+1}}{A_{k+1}}, \qquad A_k \overset{\text{def}}{=} \textstyle\sum_{i=1}^k a_i. \tag{12}$$

Theorem 1 *Let $A_k := k^3$, and consequently, $\gamma_k := 1 - \big(\frac{k}{k+1}\big)^3 = \mathcal{O}\big(\frac{1}{k}\big)$. Then for the sequence $\{x_k\}_{k \geq 1}$ generated by Algorithm 1, we have*

$$F(x_k) - F^* \leq \ell_k \overset{\text{def}}{=} F(x_k) - \frac{\phi_k^*}{A_k} \leq \mathcal{O}\big(\tfrac{H_\nu D^{2+\nu}}{k^{1+\nu}}\big). \tag{13}$$

For the case $\nu = 1$ (convex functions with Lipschitz continuous Hessian), estimate (13) gives the convergence rate of the order $\mathcal{O}(\frac{1}{k^2})$. This is the same rate, as we can achieve on this functional class by Cubic Regularization of Newton Method [33]. In accordance to (13), in order to obtain ε-accuracy in functional residual, $F(x_K) - F^* \leq \varepsilon$, it is enough to perform

$$K = \mathcal{O}\Big(\big(\tfrac{H_\nu D^{2+\nu}}{\varepsilon} \big)^{1/(1+\nu)} \Big) \tag{14}$$

iterations of Algorithm 1. In [17], there were proposed first *universal* second-order methods (which do not depend on parameters ν and H_ν of the problem class), having complexity guarantees of the same order (14). These methods are based on Cubic regularization and an adaptive search for estimating the regularization parameter at every iteration. It is important that Algorithm 1 is both universal and affine-invariant. Additionally, convergence result (13) provides us with a sequence $\{\ell_k\}_{k \geq 1}$ of computable *accuracy certificates*, which can be used as a stopping criterion of the method.

Now, let us assume that the composite component is *strongly convex* with parameter $\mu > 0$. Thus, for all $x, y \in \mathrm{dom}\,\psi$ and $\psi'(x) \in \partial\psi(x)$, it holds

$$\psi(y) \geq \psi(x) + \langle \psi'(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2. \tag{15}$$

In this situation, we are able to improve convergence estimate (13), as follows.

Theorem 2 *Let $A_k := k^5$, and consequently, $\gamma_k := 1 - \big(\frac{k}{k+1}\big)^5 = \mathcal{O}\big(\frac{1}{k}\big)$. Then for the sequence $\{x_k\}_{k\geq 1}$ generated by Algorithm 1, we have*

$$F(x_k) - F^* \quad \leq \quad \ell_k \quad \leq \quad \mathcal{O}\Big(\tfrac{H_\nu D^\nu}{\mu} \cdot \tfrac{H_\nu D^{2+\nu}}{k^{2+2\nu}}\Big). \tag{16}$$

Moreover, if the second-order <u>*condition number*</u>

$$\omega_\nu \quad \overset{\text{def}}{=} \quad \Big[\tfrac{H_\nu D^\nu}{(1+\nu)\mu}\Big]^{\frac{1}{1+\nu}} \tag{17}$$

is known, then, defining $A_k := (1+\omega_\nu^{-1})^k,\ k \geq 1,\ A_0 := 0,$ and $\gamma_k := \frac{1}{1+\omega_\nu},\ k \geq 1,\ \gamma_0 := 1,$ we obtain the global <u>*linear rate*</u> *of convergence*

$$F(x_k) - F^* \quad \leq \quad \ell_k \quad \leq \quad \exp\big(-\tfrac{k-1}{1+\omega_\nu}\big) \cdot \tfrac{H_\nu D^{2+\nu}}{1+\nu}. \tag{18}$$

According to the estimate (18), in order to get ε-accuracy in function value, it is enough to perform

$$K \quad = \quad \mathcal{O}\big((1+\omega_\nu) \cdot \log \tfrac{F(x_0)-F^*}{\varepsilon}\big)$$

iterations of the method. Hence, condition number ω_ν plays the role of the main complexity factor. This rate corresponds to that one of Cubically Regularized Newton Method (see [11, 12]). At the same time, there exists a second variant of Contracting-Domain Newton Method, where the next point is defined by minimization of the full second-order model for the smooth component augmented by the composite term over the *contracted domain* (this explains the names of our methods).

Algorithm 2: Contracting-Domain Newton Method, II

```
Initialization. Choose $x_0 \in \mathrm{dom}\,\psi$.
Iteration $k \geq 0$.
  1: Pick up $\gamma_k \in (0, 1]$.
  2: Denote
       $S_k(y) := \begin{cases} \psi(y), & y \in \gamma_k \mathrm{dom}\,\psi + (1-\gamma_k)x_k, \\ +\infty, & \text{else.} \end{cases}$
  3: Compute
       $x_{k+1} \in \mathop{\mathrm{Argmin}}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(y-x_k), y-x_k \rangle + S_k(y) \Big\}.$
```

Note, that Algorithm 1 admits similar representation as well. [3] Both methods produce the same sequences of points when $\psi(\cdot)$ is $\{0, +\infty\}$-indicator of a convex set. Otherwise, they are different. Using the same contraction technique, it was shown in [30] that the classical Frank-Wolfe algorithm can be extended onto the case of the composite optimization problems. Additionally, the second-order *Contracting Trust-Region method* was proposed, which has the same form as Algorithm 2. However, its convergence rate was established only at the level $\mathcal{O}(\frac{1}{k})$. Here, we improve its rate as follows.

Theorem 3 *Let $A_k := k^3$ and $\gamma_k := 1 - \big(\frac{k}{k+1}\big)^3 = \mathcal{O}\big(\frac{1}{k}\big)$. Then for the sequence $\{x_k\}_{k\geq 1}$ generated by Algorithm 2, we have*

$$F(x_k) - F^* \quad \leq \quad \ell_k \quad \leq \quad \mathcal{O}\big(\tfrac{H_\nu D^{2+\nu}}{k^{1+\nu}}\big). \tag{19}$$

This result is very similar to Theorem 1. However, the first algorithm can be accelerated on the class of strongly convex functions (see Theorem 2). Thus, it seems that it is more preferable.

Finally, let us consider an example, when the composite component $\psi(\cdot)$ is an ℓ_p-ball, as in (4). Then, iterations of the method can be represented as

$$x_{k+1} \quad \in \quad x_k + \mathop{\mathrm{Argmin}}_h \Big\{ \langle \nabla f(x_k), h \rangle + \tfrac{1}{2}\langle \nabla^2 f(x_k) h, h \rangle \; : \; \|x_k + \tfrac{1}{\gamma_k} h\|_p \leq \tfrac{D}{2} \Big\}. \tag{20}$$

In this form, it looks as a variant of Trust-Region scheme. To solve the subproblem in (20), we can use Interior Point Methods (e.g. Chapter 5 in [31]). See also [9], for techniques, developed for Trust-Region schemes. Usually, complexity of this step can be estimated as $\mathcal{O}(n^3)$ arithmetic operations,

[3] Indeed, it is enough to take $S_k(y) := \gamma_k \psi(x_k + \frac{1}{\gamma_k}(y - x_k))$.

which comes from the cost of computing a suitable factorization for the Hessian matrix. Alternatively, Hessian-free gradient methods can be applied, for computing an inexact step (see [6, 5]).

5 Aggregated second-order models

In this section, we propose more advanced second-order models, based on global lower bound (9). Using the same notation as before, consider a sequence of test points $\{x_k : x_k \in \mathrm{dom}\,\psi\}_{k\geq 0}$ and sequences of coefficients $\{a_k\}_{k\geq 1}$, $\{\gamma_k\}_{k\geq 0}$, satisfying the relations (12). Then, we can introduce the following *Quadratic Estimating Functions* (compare with definition (11)):

$$Q_k(x) \quad \overset{\mathrm{def}}{=} \quad \sum_{i=0}^{k-1} a_{i+1}\Big[f(x_i) + \langle \nabla f(x_i), x - x_i \rangle + \tfrac{\gamma_i}{2}\langle \nabla^2 f(x_i)(x-x_i), x - x_i\rangle + \psi(x) \Big].$$

By (9), we have the main property of Estimating Functions being satisfied. Namely, for all $x \in \mathrm{dom}\,\psi$

$$\begin{array}{rcl} A_k F(x) & \overset{(9)}{\geq} & Q_k(x) - \sum_{i=0}^{k-1} \frac{a_{i+1}\gamma_i^{1+\nu} H_\nu \|x - x_i\|^{2+\nu}}{(1+\nu)(2+\nu)} \\ & \overset{(2)}{\geq} & Q_k(x) - \frac{H_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} \sum_{i=0}^{k-1} a_{i+1}\gamma_i^{1+\nu} \quad =: \quad Q_k(x) - \frac{C_k}{2}. \end{array} \tag{21}$$

Therefore, if we would be able to guarantee for our test points the relation

$$Q_k^* \;:=\; \min_x Q_k(x) \;\geq\; A_k F(x_k) - \tfrac{C_k}{2}, \tag{22}$$

then we could immediately obtain the global convergence in function value. Fortunately, relation (22) can be achieved by simple iterations.

Algorithm 3: Aggregating Newton Method

```
Initialization. Choose $x_0 \in \mathrm{dom}\,\psi$. Set $A_0 := 0$, $Q_0(x) \equiv 0$.
Iteration $k \geq 0$.
  1: Pick up $a_{k+1} > 0$. Set $A_{k+1} := A_k + a_{k+1}$ and $\gamma_k := \frac{a_{k+1}}{A_{k+1}}$.
  2: Update Estimating Function
     $Q_{k+1}(x) \equiv Q_k(x) + a_{k+1}\big[f(x_k) + \langle \nabla f(x_k), x - x_k\rangle + \frac{\gamma_k}{2}\langle \nabla^2 f(x_k)(x-x_k), x - x_k\rangle + \psi(x)\big]$.
  3: Compute $v_{k+1} \in \mathrm{Argmin}_x\, Q_{k+1}(x)$.
  4: Set $x_{k+1} := x_k + \gamma_k(v_{k+1} - x_k)$.
```

Clearly, the most complicated part of this process is Step 3, which is computation of the minimum of Estimating Function. However, the complexity of this step remains the same, as that one for Contracting-Domain Newton Method. We obtain the following convergence result.

Theorem 4 *For the sequence $\{x_k\}_{k\geq 1}$ generated by Algorithm 3, relation* (22) *is satisfied. Consequently, for the choice $A_k := k^3$, we obtain*

$$F(x_k) - F^* \;\overset{(21)}{\leq}\; F(x_k) - \tfrac{Q_k^*}{A_k} + \tfrac{C_k}{2A_k} \;\overset{(22)}{\leq}\; \tfrac{C_k}{A_k} \;\leq\; \mathcal{O}\big(\tfrac{H_\nu D^{2+\nu}}{k^{1+\nu}}\big). \tag{23}$$

Now, for the accuracy certificate we have new expression $\bar{\ell}_k := F(x_k) - \frac{Q_k^*}{A_k} + \frac{C_k}{2A_k}$. The value of Q_k^* is available within the method directly. However, in order to compute $\bar{\ell}_k$ in practice, some estimate for C_k is required. Note, that for the given choice of coefficients $A_k := k^3$, we have $a_k = \mathcal{O}(k^2)$ and $\gamma_k = \mathcal{O}(\frac{1}{k})$. Therefore, new information enters into the model with increasing weights, which seems to be natural.

6 Stochastic finite-sum minimization

In this section, we consider the case when the smooth part f of the objective (1) is represented as a sum of M convex twice-differentiable components,

$$f(x) \quad := \quad \tfrac{1}{M}\sum_{i=1}^{M} f_i(x). \tag{24}$$

This setting appears in many machine learning applications, such as *empirical risk minimization*. Often, the number M is very big. Thus, it becomes expensive to evaluate the whole gradient or the Hessian at every iteration. Hence, *stochastic* or *incremental* methods are the methods of choice in this situation. See [4] for a survey of first-order incremental methods. The Newton-type Incremental Method with superlinear local convergence was proposed in [35]. Local linear rate of stochastic Newton methods was studied in [25]. Global convergence of sub-sampled Newton schemes, based on Damped iterations, and on Cubic regularization, was established in [36, 24, 39].

The basic idea of stochastic algorithms is to substitute the true gradients and Hessians by some random unbiased estimators g_k, and H_k, respectively, with $\mathbb{E}[g_k] = \nabla f(x_k)$ and $\mathbb{E}[H_k] = \nabla^2 f(x_k)$.

First, let us consider the simplest estimation strategy. At iteration k, we sample uniformly and independently two subsets of indices $S_k^g, S_k^H \subseteq \{1, \ldots, M\}$. Their sizes are $m_k^g := |S_k^g|$ and $m_k^H := |S_k^H|$, which are possibly different. Then, in Algorithm 1, we can use the following random estimators:

$$g_k \quad := \quad \tfrac{1}{m_k^g} \textstyle\sum_{i \in S_k^g} \nabla f_i(x_k), \qquad H_k \quad := \quad \tfrac{1}{m_k^H} \textstyle\sum_{i \in S_k^H} \nabla^2 f_i(x_k). \tag{25}$$

Let us present for this process a result on its global convergence. Note that in this section, we use the standard Euclidean norm for vectors and the corresponding induced spectral norm for matrices.

Theorem 5 *Let each component $f_i(\cdot)$ be Lipschitz continuous on $\operatorname{dom}\psi$ with constant L_0, and have Lipschitz continuous gradients and Hessians on $\operatorname{dom}\psi$ with constants L_1 and L_2, respectively. Let $\gamma_k := 1 - \bigl(\frac{k}{k+1}\bigr)^3 = \mathcal{O}\bigl(\frac{1}{k}\bigr)$. Set*

$$m_k^g \quad := \quad 1/\gamma_k^4, \qquad m_k^H \quad := \quad 1/\gamma_k^2. \tag{26}$$

Then, for the iterations $\{x_k\}_{k \geq 1}$ of Algorithm (1)*, based on estimators* (25)*, it holds*

$$\mathbb{E}[F(x_k) - F^*] \quad \leq \quad \mathcal{O}\Bigl(\tfrac{L_2 D^3 \, + \, L_1 D^2 (1 + \log(n)) \, + \, L_0 D}{k^2}\Bigr). \tag{27}$$

Therefore, in order to solve our problem with ε-accuracy in expectation, $\mathbb{E}[F(x_K) - F^*] \leq \varepsilon$, we need to perform $K = \mathcal{O}\bigl(\frac{1}{\varepsilon^{1/2}}\bigr)$ iterations of the method. In this case, the total number of gradient and Hessian samples are $\mathcal{O}\bigl(\frac{1}{\varepsilon^{5/2}}\bigr)$ and $\mathcal{O}\bigl(\frac{1}{\varepsilon^{3/2}}\bigr)$, respectively. It is interesting that we need higher accuracy for estimating the gradients, which results in a bigger batch size.

To improve this result, we incorporate a simple *variance reduction* strategy for the gradients. This is a popular technique in stochastic convex optimization (see [37, 21, 10, 20, 1, 34, 16] and references therein). At some iterations, we recompute the full gradient. However, during the whole optimization process this happens logarithmic number of times in total. Let us denote by $\pi(k)$ the maximal *power of two*, which is less than or equal to k: $\pi(k) := 2^{\lfloor \log_2 k \rfloor}$, for $k > 0$, and define $\pi(0) := 0$. The entire scheme looks as follows.

Algorithm 4: Stochastic Variance-Reduced Contracting-Domain Newton

```
Initialization. Choose $x_0 \in \operatorname{dom}\psi$.
Iteration $k \geq 0$.
    Set anchor point $z_k := x_{\pi(k)}$.
    Sample random batch $S_k \subseteq \{1, \ldots, M\}$ of size $m_k$.
    Compute variance-reduced stochastic gradient
        $g_k := \frac{1}{m_k} \sum_{i \in S_k} \bigl(\nabla f_i(x_k) - \nabla f_i(z_k) + \nabla f(z_k)\bigr)$.
    Compute stochastic Hessian
        $H_k := \frac{1}{m_k} \sum_{i \in S_k} \nabla^2 f_i(x_k)$.
    Pick up $\gamma_k \in (0, 1]$.
    Perform the main step
        $x_{k+1} \in \operatorname*{Argmin}_y \Bigl\{ \langle g_k, y - x_k \rangle + \frac{1}{2} \langle H_k(y - x_k), y - x_k \rangle + \gamma_k \psi(x_k + \frac{1}{\gamma_k}(y - x_k)) \Bigr\}$.
```

Note that this is just Algorithm 1 with random estimators g_k and H_k instead ot the true gradient and Hessian. The following global convergence result holds.

Theorem 6 *Let each component $f_i(\cdot)$ have Lipschitz continuous gradients and Hessians on* $\mathrm{dom}\,\psi$ *with constants L_1 and L_2, respectively. Let $\gamma_k := 1 - \bigl(\frac{k}{k+1}\bigr)^3 = \mathcal{O}(\frac{1}{k})$. Set batch size*

$$m_k \quad := \quad 1/\gamma_k^2. \tag{28}$$

Then, for all iterations $\{x_k\}_{k \geq 1}$ of Algorithm 4, we have

$$\mathbb{E}[F(x_k) - F^*] \quad \leq \quad \mathcal{O}\Bigl(\frac{L_2 D^3 + L_1 D^2(1+\log(n)) + L_1^{1/2} D(F(x_0)-F^*)}{k^2}\Bigr). \tag{29}$$

It is thanks to the variance reduction that we can use the same batch size for both estimators now. To solve the problem with ε-accuracy in expectation, we need $K = \mathcal{O}\bigl(\frac{1}{\varepsilon^{1/2}}\bigr)$ iterations of the method. And the total number of gradient and Hessian samples during these iterations is $\mathcal{O}\bigl(\frac{1}{\varepsilon^{3/2}}\bigr)$.

7 Experiments

Let us demonstrate computational results for the problem of training Logistic Regression model, regularized by ℓ_2-ball constraints. Thus, the smooth part of the objective has the finite-sum representation (24), each component is $f_i(x) := \log(1 + \exp(\langle a_i, x \rangle))$. The composite part is given by (4), with $p = 2$. Diameter D plays the role of regularization parameter, while vectors $\{a_i : a_i \in \mathbb{R}^n\}_{i=1}^M$ are determined by the dataset[4]. First, we compare the performance of Contracting-Domain Newton Method (Algorithm 1) and Aggregating Newton Method (Algorithm 3) with first-order optimization schemes: Frank-Wolfe algorithm [15], the classical Gradient Method, and the Fast Gradient Method [29]. For the latter two we use a line-search at each iteration, to estimate the Lipschitz constant. The results are shown on Figure 2.

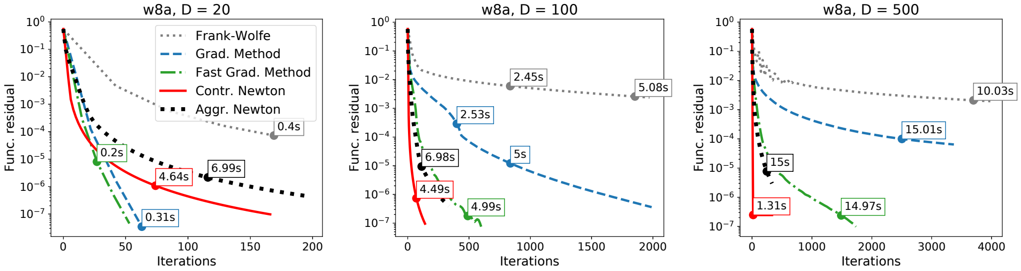

Figure 2: Training logistic regression, *w8a* $(M = 49749, n = 300)$.

We see, that for bigger D, it becomes harder to solve the optimization problem. Second-order methods demonstrate good performance both in terms of the iterations, and the total computational time. [5]

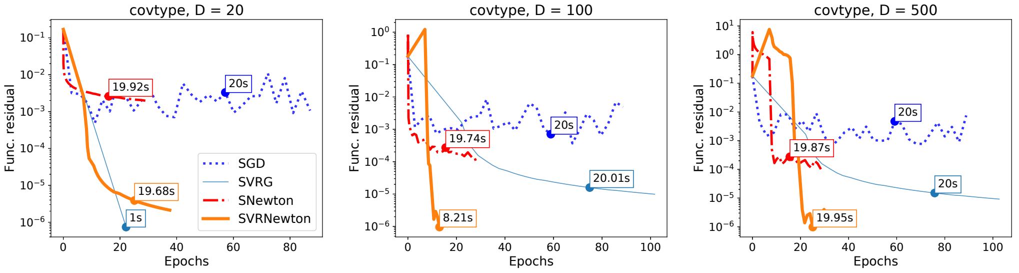

Figure 3: Stochastic methods for training logistic regression, *covtype* $(M = 581012, n = 54)$.

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[5] Clock time was evaluated using the machine with Intel Core i5 CPU, 1.6GHz; 8 GB RAM. All methods were implemented in C++. The source code can be found at https://github.com/doikov/contracting-newton/

In the next set of experiments, we compare the basic stochastic version of our method, using estimators (25) — SNewton, the method with the variance reduction (Algorithm 4) — SVRNewton, and first-order algorithms (with constant step-size, tuned for each problem): SGD and SVRG [21]. We see (Figure 3) that using the variance reduction strategy significantly improve the convergence for both first-order and second-order stochastic optimization methods.

According to these graphs, our second-order algorithms can be more efficient when solving ill-conditioned problems, producing the better solution within a given computational time. See also Section E in the supplementary material for extra experiments.

8 Discussion

Let us discuss complexity estimates, which we established in our work. For the basic versions of our method we have the global convergence in the functional residual of the form

$$F(x_k) - F^* \quad \leq \quad \mathcal{O}\Bigl(\frac{H_\nu D^{2+\nu}}{k^{1+\nu}}\Bigr).$$

Note that the complexity parameter H_ν depends only on the variation of the Hessian (in arbitrary norm). It can be much smaller than the maximal eigenvalue of the Hessian, which typically appears in the rates of first-order methods. It is important that our algorithms are free from using the norms or any other particular parameters of the problem class.

At the same time, the arithmetic complexity of one step of our methods for simple sets can be estimated as the sum of the cost of computing the Hessian, and $\mathcal{O}(n^3)$ additional operations (to compute a suitable factorization of the matrix). For example, the cost of computing the gradient of Logistic Regression is $\mathcal{O}(Mn)$, and the Hessian is $\mathcal{O}(Mn^2)$, where M is the dataset size. Hence, it is preferable to use our algorithms with exact steps in the situation when M is much bigger than n.

Broader Impact

This work does not present any foreseeable societal consequence.

Acknowledgments and Disclosure of Funding

The research results of this paper were obtained in the framework of ERC Advanced Grant 788368.

References

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

[2] Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

[3] Albert A Bennett. Newton's method in general analysis. *Proceedings of the National Academy of Sciences*, 2(10):592–598, 1916.

[4] Dimitri P Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3, 2011.

[5] Alejandro Carderera and Sebastian Pokutta. Second-order conditional gradients. *arXiv preprint arXiv:2002.08907*, 2020.

[6] Yair Carmon and John C Duchi. First-order methods for nonconvex quadratic minimization. *arXiv preprint arXiv:2003.04546*, 2020.

[7] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming*, 127(2):245–295, 2011.

[8] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function-and derivative-evaluation complexity. *Mathematical programming*, 130(2):295–319, 2011.

[9] Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

[10] Aaron Defazio, Francis Bach, and Simon 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, 2014.

[11] Nikita Doikov and Yurii Nesterov. Local convergence of tensor methods. *CORE Discussion Papers 2019/21*, 2019.

[12] Nikita Doikov and Yurii Nesterov. Minimizing uniformly convex functions by cubic regularization of Newton method. *arXiv preprint arXiv:1905.02671*, 2019.

[13] Nikita Doikov and Peter Richtárik. Randomized block cubic Newton method. In *International Conference on Machine Learning*, pages 1289–1297, 2018.

[14] Pavel Dvurechensky and Yurii Nesterov. Global performance guarantees of second-order methods for unconstrained convex minimization. Technical report, CORE Discussion Paper, 2018.

[15] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

[16] Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. *arXiv preprint arXiv:1905.11261*, 2019.

[17] Geovani N Grapiglia and Yurii Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous Hessians. *SIAM Journal on Optimization*, 27(1):478–506, 2017.

[18] Geovani N Grapiglia and Yurii Nesterov. Accelerated regularized Newton methods for minimizing composite convex functions. *SIAM Journal on Optimization*, 29(1):77–99, 2019.

[19] Filip Hanzely, Nikita Doikov, Peter Richtárik, and Yurii Nesterov. Stochastic subspace cubic Newton method. *arXiv preprint arXiv:2002.09526*, 2020.

[20] Elad Hazan and Haipeng Luo. Variance-reduced and projection-free stochastic optimization. In *International Conference on Machine Learning*, pages 1263–1271, 2016.

[21] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[22] Leonid V Kantorovich. Functional analysis and applied mathematics. *Uspekhi Matematicheskikh Nauk*, 3(6):89–185, 1948.

[23] Sai Praneeth Karimireddy, Sebastian U Stich, and Martin Jaggi. Global linear convergence of Newton's method without strong-convexity or Lipschitz gradients. *arXiv preprint arXiv:1806.00413*, 2018.

[24] Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled cubic regularization for non-convex optimization. In *International Conference on Machine Learning*, pages 1895–1904, 2017.

[25] Dmitry Kovalev, Konstantin Mishchenko, and Peter Richtárik. Stochastic Newton and cubic Newton methods with simple local linear-quadratic rates. *arXiv preprint arXiv:1912.01597*, 2019.

[26] Jason D Lee, Yuekai Sun, and Michael A Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

[27] Yurii Nesterov. Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

[28] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

[29] Yurii Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.

[30] Yurii Nesterov. Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 171(1-2):311–330, 2018.

[31] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[32] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

[33] Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton's method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

[34] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621, 2017.

[35] Anton Rodomanov and Dmitry Kropotov. A superlinearly-convergent proximal Newton-type method for the optimization of finite sums. In *International Conference on Machine Learning*, pages 2597–2605, 2016.

[36] Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled newton methods i: globally convergent algorithms. *arXiv preprint arXiv:1601.04737*, 2016.

[37] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

[38] Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for Newton-type methods. *Mathematical Programming*, 178(1-2):145–213, 2019.

[39] Nilesh Tripuraneni, Mitchell Stern, Chi Jin, Jeffrey Regier, and Michael I Jordan. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2899–2908, 2018.

[40] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Supplementary material

A Proof of Lemma 1

First, let us note that inequality (6) follows from the following simple observation, using Newton-Leibniz formula and Hölder continuity of the Hessian, for all $x, y \in \operatorname{dom} \psi$

$$\begin{array}{rcl} \|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\|_* & = & \|\int\limits_0^1 (\nabla^2 f(x + \tau(y-x)) - \nabla^2 f(x))(y-x) d\tau\|_* \\ & \overset{(5)}{\leq} & \frac{H_\nu \|y-x\|^{1+\nu}}{1+\nu}. \end{array}$$

We are ready to prove the lemma.

Lemma 1 *For all $x, y \in \operatorname{dom} \psi$ and $t \in [0,1]$, it holds*

$$f(y) \quad \geq \quad f(x) + \langle \nabla f(x), y-x \rangle + \tfrac{t}{2} \langle \nabla^2 f(x)(y-x), y-x \rangle - \tfrac{t^{1+\nu} H_\nu \|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)}.$$

Proof:

Let us prove the following bound, for all $x, y \in \operatorname{dom} \psi$ and $t \in [0,1]$

$$\langle \nabla f(y) - \nabla f(x), y-x \rangle \quad \geq \quad t \langle \nabla^2 f(x)(y-x), y-x \rangle - \tfrac{t^{1+\nu} H_\nu \|y-x\|^{2+\nu}}{1+\nu}. \tag{30}$$

For $t = 1$ it follows from (6). Therefore, we may assume that $t < 1$. Let us take $z_t := x + t(y-x)$. Then, by convexity of f, we have

$$\begin{array}{rcl} \langle \nabla f(y), y-x \rangle & = & \frac{1}{1-t} \langle \nabla f(y), y - z_t \rangle \\ & \geq & \frac{1}{1-t} \langle \nabla f(z_t), y - z_t \rangle \quad = \quad \langle \nabla f(z_t), y-x \rangle. \end{array}$$

Now, from Hölder continuity of the Hessian, we get

$$\begin{array}{rcl} \langle \nabla f(z_t), y-x \rangle & \overset{(6)}{\geq} & \langle \nabla f(x), y-x \rangle + \langle \nabla^2 f(x)(z_t - x), y-x \rangle - \frac{H_\nu \|z_t - x\|^{1+\nu} \|y-x\|}{1+\nu} \\ & = & \langle \nabla f(x), y-x \rangle + t \langle \nabla^2 f(x)(y-x), y-x \rangle - \frac{t^{1+\nu} H_\nu \|y-x\|^{2+\nu}}{1+\nu}. \end{array}$$

Thus we prove (30). Then, the claim of the lemma can be obtained by simple integration:

$$\begin{array}{rcl} f(y) - f(x) - \langle \nabla f(x), y-x \rangle & = & \int\limits_0^1 \langle \nabla f(z_\tau) - \nabla f(x), y-x \rangle d\tau \\ & \overset{(30)}{\geq} & \int\limits_0^1 t\tau \langle \nabla^2 f(x)(y-x), y-x \rangle - \frac{(t\tau)^{1+\nu} H_\nu \|y-x\|^{2+\nu}}{1+\nu} d\tau \\ & = & \frac{t}{2} \langle \nabla^2 f(x)(y-x), y-x \rangle - \frac{t^{1+\nu} H_\nu \|y-x\|^{2+\nu}}{(1+\nu)(2+\nu)}. \end{array}$$

□

B Convergence of Contracting-Domain Newton Method

In this section, we prove the global convergence of Algorithms 1 and 2. We use the same notation as in the main part. There is a sequence of controlling coefficients $\{a_k\}_{k\geq 1}$ (see relations (12)), and a sequence of linear Estimating Functions $\{\phi_k(x)\}_{k\geq 0}$. We denote by $\mu \geq 0$ the constant of strong convexity of $\psi(\cdot)$. We allow $\mu = 0$ in the following auxiliary lemma, in order to cover both the general convex and the strongly convex cases.

Lemma 2 *For the sequences $\{x_k\}_{k\geq 1}$ and $\{v_k\}_{k\geq 1}$, produced by Algorithm 1, we have*

$$A_k F(x_k) \quad \leq \quad \phi_k(x) \; + \; B_k(x), \qquad x \in \operatorname{dom}\psi, \tag{31}$$

with

$$B_k(x) \quad \equiv \quad \sum_{i=1}^{k}\Big[\tfrac{H_\nu a_i^{2+\nu}\|x-v_i\|\cdot\|x_{i-1}-v_i\|^{1+\nu}}{(1+\nu)A_i^{1+\nu}} - \tfrac{\mu a_i\|x-v_i\|^2}{2} - \tfrac{\mu a_i A_{i-1}\|x_{i-1}-v_i\|^2}{2A_i}\Big]. \tag{32}$$

Proof:

Let us prove (31) by induction. It obviously holds for $k = 0$, since $A_0 := 0$, $\phi_0(x) \equiv 0$, and $B_0(x) \equiv 0$ by definition. Assume that it holds for the current $k \geq 0$, and consider the next iterate. Stationary condition for the method step is

$$\langle \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1}-x_k), x - v_{k+1}\rangle + \psi(x) \quad \geq \quad \psi(v_{k+1}) + \tfrac{\mu}{2}\|x - v_{k+1}\|^2, \tag{33}$$

for all $x \in \operatorname{dom}\psi$. Then, we have

$$\begin{array}{rcl}
\phi_{k+1}(x) & \equiv & a_{k+1}\big[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \psi(x)\big] + \phi_k(x) \\[1em]
& \overset{(31)}{\geq} & a_{k+1}\big[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \psi(x)\big] + A_k F(x_k) - B_k(x) \\[1em]
& \overset{(*)}{\geq} & A_{k+1}\big[f(x_{k+1}) + \langle \nabla f(x_{k+1}), \tfrac{a_{k+1}x + A_k x_k}{A_{k+1}} - x_{k+1}\rangle\big] + a_{k+1}\psi(x) \\[1em]
& & \qquad + \; A_k\psi(x_k) - B_k(x) \\[1em]
& = & A_{k+1}f(x_{k+1}) + a_{k+1}\langle \nabla f(x_{k+1}), x - v_{k+1}\rangle + a_{k+1}\psi(x) \\[1em]
& & \qquad + \; A_k\psi(x_k) - B_k(x) \\[1em]
& = & A_{k+1}f(x_{k+1}) + a_{k+1}\big[\langle \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1}-x_k), x - v_{k+1}\rangle + \psi(x)\big] \\[1em]
& & \qquad + \; a_{k+1}\langle \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1}-x_k), x - v_{k+1}\rangle \\[1em]
& & \qquad + \; A_k\psi(x_k) - B_k(x) \\[1em]
& \overset{(33),(6)}{\geq} & A_{k+1}f(x_{k+1}) + a_{k+1}\big[\psi(v_{k+1}) + \tfrac{\mu}{2}\|x - v_{k+1}\|^2\big] \\[1em]
& & \qquad - \; \tfrac{H_\nu a_{k+1}^{2+\nu}\|x-v_{k+1}\|\cdot\|v_{k+1}-x_k\|^{1+\nu}}{(1+\nu)A_{k+1}^{1+\nu}} + A_k\psi(x_k) - B_k(x) \\[1em]
& \overset{(**)}{\geq} & A_{k+1}F(x_{k+1}) + \tfrac{\mu a_{k+1}\|x-v_{k+1}\|^2}{2} + \tfrac{\mu a_{k+1}A_k}{2A_{k+1}}\|x_k - v_{k+1}\|^2 \\[1em]
& & \qquad - \; \tfrac{H_\nu a_{k+1}^{2+\nu}\|x-v_{k+1}\|\cdot\|v_{k+1}-x_k\|^{1+\nu}}{(1+\nu)A_{k+1}^{1+\nu}} + A_k\psi(x_k) - B_k(x) \\[1em]
& \equiv & A_{k+1}F(x_{k+1}) - B_{k+1}(x),
\end{array}$$

where $(*)$ and $(**)$ stand for convexity of f, and strong convexity of ψ, correspondingly. Thus we have (31) established for all $k \geq 0$. □

B.1 Proof of Theorem 1

Theorem 1 *Let $A_k := k^3$, and consequently, $\gamma_k := 1 - \big(\frac{k}{k+1}\big)^3 = \mathcal{O}\big(\frac{1}{k}\big)$. Then for the sequence $\{x_k\}_{k \geq 1}$ generated by Algorithm 1, we have*

$$F(x_k) - F^* \;\; \leq \;\; \ell_k \;\; \overset{\mathrm{def}}{=} \;\; F(x_k) - \tfrac{\phi_k^*}{A_k} \;\; \leq \;\; \mathcal{O}\big(\tfrac{H_\nu D^{2+\nu}}{k^{1+\nu}}\big).$$

Proof:

First, by convexity of f we have, for all $x \in \mathrm{dom}\,\psi$

$$\phi_k(x) \;\; \leq \;\; A_k F(x).$$

Therefore, for the solution x^* of our problem: $F^* = F(x^*)$, it holds

$$F(x_k) - F^* \;\; \leq \;\; F(x_k) - \tfrac{\phi_k(x^*)}{A_k} \;\; \leq \;\; \ell_k \;\; \overset{\mathrm{def}}{=} \;\; F(x_k) - \tfrac{\phi_k^*}{A_k},$$

and this is the first part of (13).

At the same time, by Lemma 2, and using boundness of the domain, we have

$$\begin{aligned} \phi_k^* \;\; &:= \;\; \min_{x \in \mathrm{dom}\,\psi} \Big\{ \phi_k(x) \Big\} \;\; \overset{(31)}{\geq} \;\; \min_{x \in \mathrm{dom}\,\psi} \Big\{ A_k F(x_k) - B_k(x) \Big\} \\ &\geq \;\; A_k F(x_k) - \tfrac{H_\nu D^{2+\nu}}{1+\nu} \sum_{i=1}^{k} \tfrac{a_i^{2+\nu}}{A_i^{1+\nu}} \end{aligned}$$

Therefore, for the choice $A_k := k^3$, we finally obtain

$$\begin{aligned} \ell_k \;\; &\leq \;\; \tfrac{H_\nu D^{2+\nu}}{(1+\nu)A_k} \sum_{i=1}^{k} \tfrac{a_i^{2+\nu}}{A_i^{1+\nu}} \;\; = \;\; \tfrac{H_\nu D^{2+\nu}}{(1+\nu)k^3} \sum_{i=1}^{k} \tfrac{(i^3-(i-1)^3)^{2+\nu}}{i^{3(1+\nu)}} \\ &\leq \;\; \tfrac{H_\nu D^{2+\nu}}{(1+\nu)k^3} \sum_{i=1}^{k} \tfrac{3^{2+\nu} i^{2(2+\nu)}}{i^{3(1+\nu)}} \;\; = \;\; \tfrac{3^{2+\nu} H_\nu D^{2+\nu}}{(1+\nu)k^3} \sum_{i=1}^{k} i^{1-\nu} \\ &= \;\; \mathcal{O}\big(\tfrac{H_\nu D^{2+\nu}}{k^{1+\nu}}\big). \end{aligned}$$

□

B.2 Proof of Theorem 2

Theorem 2 *Let $A_k := k^5$, and consequently, $\gamma_k := 1 - \big(\frac{k}{k+1}\big)^5 = \mathcal{O}\big(\frac{1}{k}\big)$. Then for the sequence $\{x_k\}_{k \geq 1}$ generated by Algorithm 1, we have*

$$F(x_k) - F^* \;\; \leq \;\; \ell_k \;\; \leq \;\; \mathcal{O}\Big(\tfrac{H_\nu D^\nu}{\mu} \cdot \tfrac{H_\nu D^{2+\nu}}{k^{2+2\nu}}\Big).$$

Moreover, if the second-order condition number

$$\omega_\nu \;\; \overset{\mathrm{def}}{=} \;\; \Big[\tfrac{H_\nu D^\nu}{(1+\nu)\mu}\Big]^{\frac{1}{1+\nu}}$$

is known, then, defining $A_k := (1+\omega_\nu^{-1})^k,\; k \geq 1$, $A_0 := 0$, and $\gamma_k := \frac{1}{1+\omega_\nu},\; k \geq 1$, $\gamma_0 := 1$, we obtain the global linear rate of convergence

$$F(x_k) - F^* \;\; \leq \;\; \ell_k \;\; \leq \;\; \exp\big(-\tfrac{k-1}{1+\omega_\nu}\big) \cdot \tfrac{H_\nu D^{2+\nu}}{1+\nu}.$$

Proof:

Starting from the same reasoning, as in the proof of Theorem 1, we get

$$F(x_k) - F^* \;\; \leq \;\; \ell_k \;\; \overset{\mathrm{def}}{=} \;\; F(x_k) - \tfrac{\phi_k^*}{A_k}.$$

Let us denote by u_k the minimum of the Estimating Function ϕ_k. Thus,

$$\ell_k \;\; = \;\; F(x_k) - \tfrac{\phi_k(u_k)}{A_k} \;\; \overset{(31)}{\leq} \;\; \tfrac{1}{A_k} B_k(u_k) \;\; \equiv \;\; \tfrac{1}{A_k} \sum\limits_{i=1}^{k} B_k^{(i)},$$

with

$$\begin{array}{rcl} B_k^{(i)} & \overset{\mathrm{def}}{=} & a_i \Big[\tfrac{H_\nu a_i^{1+\nu} \|u_k - v_i\| \cdot \|x_{i-1} - v_i\|^{1+\nu}}{(1+\nu) A_i^{1+\nu}} - \tfrac{\mu \|u_k - v_i\|^2}{2} \Big] - \tfrac{\mu a_i A_{i-1} \|x_{i-1} - v_i\|^2}{2 A_i} \\ & \leq & a_i \max\limits_{t \geq 0} \Big\{ \tfrac{H_\nu a_i^{1+\nu} \|x_{i-1} - v_i\|^{1+\nu} t}{(1+\nu) A_i^{1+\nu}} - \tfrac{\mu t^2}{2} \Big\} - \tfrac{\mu a_i A_{i-1} \|x_{i-1} - v_i\|^2}{2 A_i} \\ & = & \tfrac{a_i}{2\mu} \Big(\tfrac{H_\nu a_i^{1+\nu} \|x_{i-1} - v_i\|^{1+\nu}}{(1+\nu) A_i^{1+\nu}} \Big)^2 - \tfrac{\mu a_i A_{i-1} \|x_{i-1} - v_i\|^2}{2 A_i}. \end{array} \qquad (34)$$

Therefore, for the choice $A_k := k^5$, we have

$$\begin{array}{rcl} \ell_k & \leq & \tfrac{1}{A_k} \sum\limits_{i=1}^{k} \tfrac{a_i}{2\mu} \Big(\tfrac{H_\nu a_i^{1+\nu} \|x_{i-1} - v_i\|^{1+\nu}}{(1+\nu) A_i^{1+\nu}} \Big)^2 \;\; \leq \;\; \tfrac{H_\nu^2 D^{2(1+\nu)}}{2\mu(1+\nu)^2 A_k} \sum\limits_{i=1}^{k} \tfrac{a_i^{2(1+\nu)+1}}{A_i^{2(1+\nu)}} \\ & = & \tfrac{H_\nu^2 D^{2(1+\nu)}}{2\mu(1+\nu)^2 k^5} \sum\limits_{i=1}^{k} \tfrac{(i^5 - (i-1)^5)^{2(1+\nu)+1}}{i^{10(1+\nu)}} \;\; \leq \;\; \tfrac{5^{2(1+\nu)+1} H_\nu^2 D^{2(1+\nu)}}{2\mu(1+\nu)^2 k^5} \sum\limits_{i=1}^{k} i^{2-2\nu} \\ & = & \mathcal{O}\big(\tfrac{H_\nu D^\nu}{\mu} \cdot \tfrac{H_\nu D^{2+\nu}}{k^{2+2\nu}} \big). \end{array}$$

Thus we have justified (16). To obtain the linear rate (18), we set

$$A_k \;\; := \;\; (1 + \omega_\nu^{-1})^k, \qquad k \geq 1,$$

and $A_0 := 0$. So, $a_1 = A_1$ and

$$a_i \;\; = \;\; A_i - A_{i-1} \;\; = \;\; \omega_\nu^{-1} A_{i-1}, \qquad i \geq 2.$$

Therefore, for the values $\{B_k^{(i)}\}_{i=1}^k$, we have

$$B_k^{(1)} \;\; \leq \;\; a_1 \tfrac{H_\nu D^{2+\nu}}{1+\nu} \;\; = \;\; A_1 \tfrac{H_\nu D^{2+\nu}}{1+\nu},$$

and

$$\begin{array}{rcl} B_k^{(i)} & \overset{(34)}{\leq} & \tfrac{H_\nu^2 D^{2\nu} \|x_{i-1} - v_i\|^2 a_i^{3+2\nu}}{2\mu(1+\nu)^2 A_i^{2+2\nu}} - \tfrac{\mu a_i A_{i-1} \|x_{i-1} - v_i\|^2}{2 A_i} \\ & = & \tfrac{\mu a_i A_{i-1} \|x_{i-1} - v_i\|^2}{2 A_i} \Big(\Big[\tfrac{H_\nu D^\nu}{(1+\nu)\mu} \Big]^2 \tfrac{a_i^{2+2\nu}}{A_i^{1+2\nu} A_{i-1}} - 1 \Big) \\ & \leq & \tfrac{\mu a_i A_{i-1} \|x_{i-1} - v_i\|^2}{2 A_i} \Big(\Big[\tfrac{H_\nu D^\nu}{(1+\nu)\mu} \Big]^2 \Big[\tfrac{a_i}{A_{i-1}} \Big]^{2(1+\nu)} - 1 \Big) \\ & = & 0, \qquad 2 \leq i \leq k, \end{array}$$

since by our choice

$$\tfrac{a_i}{A_{i-1}} \;\; = \;\; \omega_\nu^{-1} \;\; \overset{(17)}{=} \;\; \Big[\tfrac{(1+\nu)\mu}{H_\nu D^\nu} \Big]^{\frac{1}{1+\nu}}.$$

Finally, we obtain

$$\begin{array}{rcl} \ell_k & \leq & \tfrac{1}{A_k} B_k^{(1)} \;\; \leq \;\; \tfrac{A_1}{A_k} \cdot \tfrac{H_\nu D^{2+\nu}}{1+\nu} \;\; = \;\; \tfrac{1}{(1+\omega_\nu^{-1})^{k-1}} \cdot \tfrac{H_\nu D^{2+\nu}}{1+\nu} \\ & \leq & \exp\big(-\tfrac{k-1}{1+\omega_\nu} \big) \cdot \tfrac{H_\nu D^{2+\nu}}{1+\nu}. \end{array}$$

□

B.3 Proof of Theorem 3

Theorem 3 *Let $A_k := k^3$ and $\gamma_k := 1 - \bigl(\frac{k}{k+1}\bigr)^3 = \mathcal{O}\bigl(\frac{1}{k}\bigr)$. Then for the sequence $\{x_k\}_{k \geq 1}$ generated by Algorithm 2, we have*

$$F(x_k) - F^* \quad \leq \quad \ell_k \quad \leq \quad \mathcal{O}\bigl(\tfrac{H_\nu D^{2+\nu}}{k^{1+\nu}}\bigr).$$

Proof:

The proof is very similar to that one for Algorithm 1. First, stationary condition for one iteration of Algorithm 2 is

$$\begin{aligned} & \langle \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k), x - v_{k+1} \rangle + \tfrac{1}{\gamma_k}\psi\bigl(\gamma_k x + (1-\gamma_k)x_k\bigr) \\ \geq \quad & \tfrac{1}{\gamma_k}\psi(x_{k+1}), \end{aligned} \tag{35}$$

for all $x \in \mathrm{dom}\,\psi$ and $k \geq 0$ (compare with (33)), where

$$v_{k+1} \quad := \quad x_k + \tfrac{1}{\gamma_k}(x_{k+1} - x_k) \quad \in \quad \mathrm{dom}\,\psi.$$

Now, let us prove by induction the following bound

$$\phi_k(x) \quad \geq \quad A_k F(x_k) - B_k, \qquad x \in \mathrm{dom}\,\psi, \tag{36}$$

with $B_k := \frac{H_\nu D^{2+\nu}}{1+\nu} \sum_{i=1}^k \frac{a_i^{2+\nu}}{A_i^{1+\nu}}$. It obviously holds for $k = 0$, since both sides are zero. Assume that it holds for the current $k \geq 0$. Then, we have for the next iterate

$$\begin{aligned} \phi_{k+1}(x) \quad & \equiv \quad a_{k+1}\bigl[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \psi(x)\bigr] + \phi_k(x) \\ & \overset{(36)}{\geq} \quad a_{k+1}\bigl[f(x_{k+1}) + \langle \nabla f(x_{k+1}), x - x_{k+1}\rangle + \psi(x)\bigr] + A_k F(x_k) - B_k \\ & \overset{(*)}{\geq} \quad A_{k+1}\bigl[f(x_{k+1}) + \langle \nabla f(x_{k+1}), \tfrac{a_{k+1}x + A_k x_k}{A_{k+1}} - x_{k+1}\rangle\bigr] + a_{k+1}\psi(x) + A_k\psi(x_k) \\ & \qquad - B_k \\ & \overset{(**)}{\geq} \quad A_{k+1}\bigl[f(x_{k+1}) + \langle \nabla f(x_{k+1}), \tfrac{a_{k+1}x + A_k x_k}{A_{k+1}} - x_{k+1}\rangle + \psi\bigl(\tfrac{a_{k+1}x + A_k x_k}{A_{k+1}}\bigr)\bigr] - B_k, \end{aligned}$$

where $(*)$ and $(**)$ stand for convexity of f and ψ, correspondingly. Using both stationary condition and smoothness, we obtain, for all $x \in \mathrm{dom}\,\psi$

$$\begin{aligned} & \langle \nabla f(x_{k+1}), \tfrac{a_{k+1}x + A_k x_k}{A_{k+1}} - x_{k+1}\rangle + \psi\bigl(\tfrac{a_{k+1}x + A_k x_k}{A_{k+1}}\bigr) \\ = \quad & \gamma_k \langle \nabla f(x_{k+1}), x - v_{k+1}\rangle + \psi\bigl(\gamma_k x + (1-\gamma_k)x_k\bigr) \\ = \quad & \gamma_k \langle \nabla f(x_k) + \nabla^2 f(x_k)(x_{k+1} - x_k), x - v_{k+1}\rangle + \psi\bigl(\gamma_k x + (1-\gamma_k)x_k\bigr) \\ & \qquad + \gamma_k \langle \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k), x - v_{k+1}\rangle \\ \overset{(35),(6)}{\geq} \quad & \psi(x_{k+1}) - \tfrac{\gamma_k H_\nu \|x_{k+1} - x_k\|^{1+\nu}\|x - v_{k+1}\|}{1+\nu} \quad = \quad \psi(x_{k+1}) - \tfrac{\gamma_k^{2+\nu} H_\nu \|v_{k+1} - x_k\|^{1+\nu}\|x - v_{k+1}\|}{1+\nu} \\ \geq \quad & \psi(x_{k+1}) - \tfrac{\gamma_k^{2+\nu} H_\nu D^{2+\nu}}{1+\nu}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} \phi_{k+1}(x) \quad & \geq \quad A_{k+1}\bigl[f(x_{k+1}) + \psi(x_{k+1}) - \tfrac{\gamma_k^{2+\nu} H_\nu D^{2+\nu}}{1+\nu}\bigr] - B_k \\ & = \quad A_{k+1} F(x_{k+1}) - B_{k+1}, \end{aligned}$$

and (36) is justified for all $k \geq 0$. Finally, by convexity of f, we get

$$\begin{aligned} F(x_k) - F^* \quad & \leq \quad \ell_k \overset{\mathrm{def}}{=} F(x_k) - \tfrac{\phi_k^*}{A_k} \\ & \overset{(36)}{\leq} \quad \tfrac{B_k}{A_k} \; = \; \tfrac{H_\nu D^{2+\nu}}{(1+\nu)A_k} \sum\limits_{i=1}^{k} \tfrac{a_i^{2+\nu}}{A_i^{1+\nu}} \\ & = \quad \mathcal{O}\big(\tfrac{H_\nu D^{2+\nu}}{k^{1+\nu}}\big), \end{aligned}$$

where the last equation holds from the choice $A_k := k^3$ (see the end of the proof of Theorem 1). □

C Convergence of Aggregating Newton Method

In this section, we establish the convergence result for Algorithm 3.

C.1 Proof of Theorem 4

Theorem 4 *For the sequence $\{x_k\}_{k\geq 1}$ generated by Algorithm 3, relation* (22) *is satisfied.*

Proof:

Let us establish the relation (22) by induction. It obviously holds for $k = 0$. Assume that it is proven for the current iterate $k \geq 0$, and consider the next step:

$$\begin{aligned} & Q_{k+1}(v_{k+1}) \\ \equiv \quad & a_{k+1}\big[f(x_k) + \langle \nabla f(x_k), v_{k+1} - x_k \rangle + \tfrac{\gamma_k}{2}\langle \nabla^2 f(x_k)(v_{k+1} - x_k), v_{k+1} - x_k \rangle \\ & \qquad + \psi(v_{k+1})\big] \; + \; Q_k(v_{k+1}) \\ \overset{(22)}{\geq} \quad & a_{k+1}\big[f(x_k) + \langle \nabla f(x_k), v_{k+1} - x_k \rangle + \tfrac{\gamma_k}{2}\langle \nabla^2 f(x_k)(v_{k+1} - x_k), v_{k+1} - x_k \rangle \\ & \qquad + \psi(v_{k+1})\big] \; + \; A_k F(x_k) - \tfrac{C_k}{2} \\ = \quad & A_{k+1}\big[f(x_k) + \gamma_k \langle \nabla f(x_k), v_{k+1} - x_k \rangle + \tfrac{\gamma_k^2}{2}\langle \nabla^2 f(x_k)(v_{k+1} - x_k), v_{k+1} - x_k \rangle\big] \\ & \qquad + a_{k+1}\psi(v_{k+1}) + A_k \psi(x_k) - \tfrac{C_k}{2} \\ = \quad & A_{k+1}\big[f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \tfrac{1}{2}\langle \nabla^2 f(x_k)(x_{k+1} - x_k), x_{k+1} - x_k \rangle\big] \\ & \qquad + a_{k+1}\psi(v_{k+1}) + A_k \psi(x_k) - \tfrac{C_k}{2} \\ \overset{(7)}{\geq} \quad & A_{k+1}\big[f(x_{k+1}) - \tfrac{H_\nu \|x_{k+1} - x_k\|^{2+\nu}}{(1+\nu)(2+\nu)}\big] + a_{k+1}\psi(v_{k+1}) + A_k \psi(x_k) - \tfrac{C_k}{2} \\ = \quad & A_{k+1} f(x_{k+1}) - \tfrac{A_{k+1}\gamma_k^{2+\nu} H_\nu \|v_{k+1} - x_k\|^{2+\nu}}{(1+\nu)(2+\nu)} + a_{k+1}\psi(v_{k+1}) + A_k \psi(x_k) - \tfrac{C_k}{2} \\ \geq \quad & A_{k+1} f(x_{k+1}) - \tfrac{a_{k+1}\gamma_k^{1+\nu} \mathcal{H}_\nu D^{2+\nu}}{(1+\nu)(2+\nu)} + A_{k+1}\psi(x_{k+1}) - \tfrac{C_k}{2} \\ = \quad & A_{k+1} F(x_{k+1}) - \tfrac{C_{k+1}}{2}. \end{aligned}$$

Thus, we have (22) justified for all $k \geq 0$. □

D Convergence of stochastic methods

Let us consider the following general iterations, for solving optimization problem (1):

$$x_{k+1} \quad \in \quad \operatorname*{Argmin}_{y} \Bigl\{ \langle g_k, y - x_k \rangle + \tfrac{1}{2} \langle H_k (y - x_k), y - x_k \rangle + S_k(y) \Bigr\}, \quad k \geq 0 \tag{37}$$

with $S_k(y) := \gamma_k \psi(x_k + \frac{1}{\gamma_k}(y - x_k))$. This is Algorithm 1 with substituted vector g_k and matrix H_k instead of the true gradient and the Hessian. First, we need to study the convergence of this process. For simplicity, let us study the case $\nu = 1$ only (convex functions with Lipschitz continuous Hessian, we denote the corresponding Lipschitz constant by L_2). Recall, that in this section we use the standard Euclidean norm for vectors and induced spectral norm for matrices.

As before, we use the sequence of positive numbers $\{a_k\}_{k \geq 1}$, and set

$$\gamma_k \quad := \quad \tfrac{a_{k+1}}{A_{k+1}}, \qquad A_k \quad \overset{\text{def}}{=} \quad \sum_{i=1}^{k} a_i.$$

Lemma 3 *For iterations* (37)*, we have for all* $k \geq 1$

$$F(x_k) - F^* \quad \leq \quad \tfrac{B_k}{A_k}, \tag{38}$$

with

$$B_k \quad := \quad \tfrac{L_2 D^3}{2} \sum_{i=0}^{k-1} \tfrac{a_{i+1}^3}{A_{i+1}^2} + D \sum_{i=0}^{k-1} a_{i+1} \|\nabla f(x_i) - g_i\| + D^2 \sum_{i=0}^{k-1} \tfrac{a_{i+1}^2}{A_{i+1}} \|\nabla^2 f(x_i) - H_i\|.$$

Proof:

Let us prove by induction the following inequality

$$A_k F(x) \quad \geq \quad A_k F(x_k) - B_k, \qquad x \in \operatorname{dom} \psi. \tag{39}$$

It obviously holds for $k = 0$, and for $k \geq 1$ it is equivalent to (38).

Assume that (39) is satisfied for some $k \geq 0$, and consider the next step:

$$\begin{array}{rcl} A_{k+1} F(x) & = & a_{k+1} F(x) + A_k F(x) \\ & \overset{(39)}{\geq} & a_{k+1} F(x) + A_k F(x_k) - B_k \\ & \overset{(*)}{\geq} & A_{k+1} f\bigl(\frac{a_{k+1} x + A_k x_k}{A_{k+1}} \bigr) + a_{k+1} \psi(x) + A_k \psi(x_k) - B_k \\ & \overset{(*)}{\geq} & A_{k+1} \bigl[f(x_{k+1}) + \langle \nabla f(x_{k+1}), \frac{a_{k+1} x + A_k x_k}{A_{k+1}} - x_{k+1} \rangle \bigr] + a_{k+1} \psi(x) \\ & & \quad + A_k \psi(x_k) - B_k, \end{array} \tag{40}$$

where $(*)$ stands for convexity of f. Now, let us denote the point

$$v_{k+1} \quad := \quad x_k + \tfrac{1}{\gamma_k}(x_{k+1} - x_k) \quad \in \quad \operatorname{dom} \psi.$$

Then, stationary condition for the method step (37) can be written as

$$\langle g_k + H_k (x_{k+1} - x_k), x - v_{k+1} \rangle + \psi(x) \quad \geq \quad \psi(v_{k+1}), \tag{41}$$

for all $x \in \operatorname{dom} \psi$. Therefore,

$$\begin{aligned}
& A_{k+1} \langle \nabla f(x_{k+1}), \tfrac{a_{k+1}x + A_k x_k}{A_{k+1}} - x_{k+1} \rangle + a_{k+1}\psi(x) \\
& = \quad a_{k+1}\big[\langle \nabla f(x_{k+1}), x - v_{k+1} \rangle + \psi(x) \big] \\
& = \quad a_{k+1}\big[\langle g_k + H_k(x_{k+1} - x_k), x - v_{k+1} \rangle + \psi(x) \\
& \qquad\qquad + \langle \nabla f(x_k) - g_k, x - v_{k+1} \rangle \\
& \qquad\qquad + \langle (\nabla^2 f(x_k) - H_k)(x_{k+1} - x_k), x - v_{k+1} \rangle \\
& \qquad\qquad + \langle \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k), x - v_{k+1} \rangle \big] \qquad (42) \\
& \overset{(41),(6)}{\geq} \quad a_{k+1}\big[\psi(v_{k+1}) - \|\nabla f(x_k) - g_k\| \cdot \|x - v_{k+1}\| \\
& \qquad\qquad - \gamma_k \|\nabla^2 f(x_k) - H_k\| \cdot \|v_{k+1} - x_k\| \cdot \|x - v_{k+1}\| \\
& \qquad\qquad - \tfrac{L_2 \gamma_k^2 \|v_{k+1} - x_k\|^2 \cdot \|x - v_{k+1}\|}{2} \big] \\
& \geq \quad a_{k+1}\psi(v_{k+1}) - a_{k+1} D \|\nabla f(x_k) - g_k\|_* - \tfrac{a_{k+1}^2 D^2 \|\nabla^2 f(x_k) - H_k\|}{A_{k+1}} - \tfrac{a_{k+1}^3 L_2 D^3}{A_{k+1}^2}.
\end{aligned}$$

Thus, combining all together, and using convexity of ψ, we obtain

$$\begin{aligned}
A_{k+1}F(x) \quad & \overset{(40),(42)}{\geq} \quad A_{k+1} f(x_{k+1}) + a_{k+1}\psi(v_{k+1}) + A_k \psi(x_k) - B_k \\
& \qquad\quad - a_{k+1} D \|\nabla f(x_k) - g_k\| - \tfrac{a_{k+1}^2 D^2 \|\nabla^2 f(x_k) - H_k\|}{A_{k+1}} - \tfrac{a_{k+1}^3 L_2 D^3}{A_{k+1}^2} \\
& \geq \quad A_{k+1} F(x_{k+1}) - B_{k+1}.
\end{aligned}$$

So, we have (39) justified for all $k \geq 0$. □

Now, we are ready to prove convergence results for the process (37) with the basic variant of stochastic estimators (25), and with the variance reduction strategy for the gradients, incorporated into Algorithm 4.

D.1 Proof of Theorem 5

Theorem 5 *Let each component $f_i(\cdot)$ be Lipschitz continuous on $\operatorname{dom}\psi$ with constant L_0, and have Lipschitz continuous gradients and Hessians on $\operatorname{dom}\psi$ with constants L_1 and L_2, respectively. Let $\gamma_k := 1 - \big(\frac{k}{k+1}\big)^3 = \mathcal{O}\big(\frac{1}{k}\big)$. Set*

$$m_k^g \;:=\; 1/\gamma_k^4, \qquad m_k^H \;:=\; 1/\gamma_k^2.$$

Then, for the iterations $\{x_k\}_{k \geq 1}$ of Algorithm (1), based on estimators (25), it holds

$$\mathbb{E}[F(x_k) - F^*] \;\leq\; \mathcal{O}\Big(\tfrac{L_2 D^3 + L_1 D^2 (1 + \log(n)) + L_0 D}{k^2} \Big).$$

Proof:

Let us fix iteration $k \geq 0$. For one uniform random sample $i \in \{1, \ldots, M\}$, we have

$$\mathbb{E}\|\nabla f(x_k) - \nabla f_i(x_k)\|^2 \;=\; \mathbb{E}\|\nabla f_i(x_k)\|^2 - \|\nabla f(x_k)\|^2 \;\leq\; L_0^2. \qquad (43)$$

Therefore, for the random batch of size m_k^g, we obtain

$$\begin{array}{rcl} \mathbb{E}\|\nabla f(x_k) - g_k\| & \leq & \sqrt{\mathbb{E}\|\nabla f(x_k) - g_k\|^2} \\ & = & \sqrt{\frac{1}{(m_k^g)^2}\mathbb{E}\| \sum_{i \in S_k^g} (\nabla f(x_k) - \nabla f_i(x_k))\|^2} \\ & = & \sqrt{\frac{1}{(m_k^g)^2} \sum_{i \in S_k^g} \mathbb{E}\|\nabla f(x_k) - \nabla f_i(x_k)\|^2} \\ & \overset{(43)}{\leq} & \frac{L_0}{\sqrt{m_k^g}}. \end{array} \tag{44}$$

More advanced reasoning for matrices (Matrix Bernstein Inequality; see Chapter 6 in [40]) gives

$$\begin{array}{rcl} \mathbb{E}\|\nabla^2 f(x_k) - H_k\| & \leq & L_1\Big(\sqrt{\frac{2\log(2n)}{m_k^H}} + \frac{2\log(2n)}{3m_k^H}\Big) \\ & \leq & \frac{L_1(3\sqrt{2\log(2n)} + 2\log(2n))}{3\sqrt{m_k^H}} \;\leq\; \frac{L_1(6+7\log(2n))}{6\sqrt{m_k^H}}. \end{array} \tag{45}$$

So, using these estimates together, we have, for every $k \geq 1$

$$\begin{array}{rcl} \mathbb{E}[F(x_k) - F^*] & \overset{(38)}{\leq} & \frac{1}{A_k}\Big(\frac{L_2 D^3}{2}\sum\limits_{i=0}^{k-1}\frac{a_{i+1}^3}{A_{i+1}^2} + D\sum\limits_{i=0}^{k-1} a_{i+1}\mathbb{E}\|\nabla f(x_i) - g_i\| \\ & & \qquad + D^2 \sum\limits_{i=0}^{k-1}\frac{a_{i+1}^2}{A_{i+1}}\mathbb{E}\|\nabla^2 f(x_i) - H_i\|\Big) \\ & \overset{(44),(45)}{\leq} & \frac{1}{A_k}\Big(\frac{L_2 D^3}{2}\sum\limits_{i=0}^{k-1}\frac{a_{i+1}^3}{A_{i+1}^2} + L_0 D \sum\limits_{i=0}^{k-1}\frac{a_{i+1}}{\sqrt{m_i^g}} \\ & & \qquad + \frac{L_1 D^2(6+7\log(2n))}{6}\sum\limits_{i=0}^{k-1}\frac{a_{i+1}^2}{A_{i+1}\sqrt{m_i^H}}\Big) \\ & \overset{(26)}{=} & \frac{1}{A_k}\Big(\frac{L_2 D^3}{2} + L_0 D + \frac{L_1 D^2 (6+7\log(2n))}{6}\Big)\sum\limits_{i=0}^{k-1}\frac{a_{i+1}^3}{A_{i+1}^2}. \end{array}$$

Thus, for the choice $A_k := k^3$, we get

$$\mathbb{E}[F(x_k) - F^*] \quad \leq \quad \mathcal{O}\Big(\frac{L_2 D^3 + L_1 D^2(1+\log(n)) + L_0 D}{k^2}\Big).$$

□

D.2 Proof of Theorem 6

Theorem 6 *Let each component $f_i(\cdot)$ have Lipschitz continuous gradients and Hessians on $\operatorname{dom}\psi$ with constants L_1 and L_2, respectively. Let $\gamma_k := 1 - \big(\frac{k}{k+1}\big)^3 = \mathcal{O}(\frac{1}{k})$. Set batch size*

$$m_k \quad := \quad 1/\gamma_k^2.$$

Then, for all iterations $\{x_k\}_{k \geq 1}$ of Algorithm 4, we have

$$\mathbb{E}[F(x_k) - F^*] \quad \leq \quad \mathcal{O}\Big(\frac{L_2 D^3 + L_1 D^2(1+\log(n)) + L_1^{1/2} D (F(x_0) - F^*)}{k^2}\Big).$$

Proof:

Let us consider the following stochastic estimate

$$g_k^i \quad := \quad \nabla f_i(x_k) - \nabla f_i(z_k) + \nabla f(z_k),$$

for a uniform random sample $i \in \{1, \ldots, M\}$, and a current iterate $k \geq 0$. We denote by x^* the solution of our problem: $F^* = F(x^*)$, stationary condition for which is

$$\langle \nabla f(x^*), x - x^* \rangle + \psi(x) \quad \geq \quad \psi(x^*), \qquad x \in \operatorname{dom} \psi. \tag{46}$$

Then, it holds

$$\begin{array}{rcl}
\mathbb{E}\|\nabla f(x_k) - g_k^i\|^2 & = & \mathbb{E}\|(\nabla f(x_k) - \nabla f(x^*)) \\
& + & (\nabla f_i(z_k) - \nabla f_i(x^*) - \nabla f(z_k) + \nabla f(x^*)) \\
& + & (\nabla f_i(x^*) - \nabla f_i(x_k))\|^2 \\
& \leq & 3\mathbb{E}\|\nabla f(x_k) - \nabla f(x^*)\|^2 \\
& + & 3\mathbb{E}\|(\nabla f_i(z_k) - \nabla f_i(x^*)) - (\nabla f(z_k) - \nabla f(x^*))\|^2 \\
& + & 3\mathbb{E}\|\nabla f_i(x_k) - \nabla f_i(x^*)\|^2 \\
& \leq & 3\Big(\mathbb{E}\|\nabla f(x_k) - \nabla f(x^*)\|^2 + \mathbb{E}\|\nabla f_i(z_k) - \nabla f_i(x^*)\|^2 \\
& + & \mathbb{E}\|\nabla f_i(x_k) - \nabla f_i(x^*)\|^2\Big),
\end{array}$$

where we used the following simple bounds:

$$\begin{array}{rcl}
\|a + b + c\|^2 & \leq & 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2, \\
\mathbb{E}\|\xi - \mathbb{E}\xi\|^2 & \leq & \mathbb{E}\|\xi\|^2,
\end{array}$$

which are valid for any $a, b, c \in \mathbb{R}^n$ and arbitrary random vector $\xi \in \mathbb{R}^n$.

Now, by Lipschitz continuity of the gradients, we have (see Theorem 2.1.5 in [31])

$$\begin{array}{rcl}
\|\nabla f(x_k) - \nabla f(x^*)\|^2 & \leq & 2L_1\big(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle\big) \\
& \overset{(46)}{\leq} & 2L_1\big(F(x_k) - F^*\big).
\end{array}$$

The same holds for the random sample i, for arbitrary fixed $x \in \operatorname{dom} \psi$

$$\begin{array}{rcl}
\mathbb{E}_i\|\nabla f_i(x) - \nabla f_i(x^*)\|^2 & \leq & 2L_1\mathbb{E}_i\big[f_i(x) - f_i(x^*) - \langle \nabla f_i(x^*), x - x^* \rangle\big] \\
& = & 2L_1\big(f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle\big) \\
& \overset{(46)}{\leq} & 2L_1\big(F(x) - F^*\big).
\end{array}$$

Thus, we obtain

$$\mathbb{E}\|\nabla f(x_k) - g_k^i\|^2 \quad \leq \quad 12L_1\mathbb{E}[F(x_k) - F^*] + 6L_1\mathbb{E}[F(z_k) - F^*]. \tag{47}$$

Consequently, for the random batch

$$g_k \quad := \quad \tfrac{1}{m_k}\textstyle\sum_{i \in S_k} g_k^i,$$

we have (compare with (44))

$$\begin{array}{rcl}
\mathbb{E}\|\nabla f(x_k) - g_k\| & \leq & \sqrt{\tfrac{1}{(m_k)^2}\sum_{i \in S_k}\mathbb{E}\|\nabla f(x_k) - g_k^i\|^2} \\
& \overset{(47)}{\leq} & \sqrt{\tfrac{6L_1}{m_k}\big(2\mathbb{E}[F(x_k) - F^*] + \mathbb{E}[F(z_k) - F^*]\big)} \\
& \leq & \sqrt{\tfrac{12L_1}{m_k}\mathbb{E}[F(x_k) - F^*]} + \sqrt{\tfrac{6L_1}{m_k}\mathbb{E}[F(z_k) - F^*]}.
\end{array} \tag{48}$$

So, using the variance reduction for the gradients, and the basic estimate for the Hessians, we have, for every $k \geq 1$

$$\begin{aligned} \mathbb{E}[F(x_k) - F^*] \quad & \overset{(38),(48),(45)}{\leq} \quad \tfrac{1}{A_k}\Big(\tfrac{L_2 D^3}{2} \sum\limits_{i=0}^{k-1} \tfrac{a_{i+1}^3}{A_{i+1}^2} \\ & \qquad + D\sqrt{6L_1} \sum\limits_{i=0}^{k-1} \tfrac{a_{i+1}}{\sqrt{m_i}} \big(\sqrt{2\mathbb{E}[F(x_i) - F^*]} + \sqrt{\mathbb{E}[F(z_i) - F^*]} \big) \\ & \qquad + \tfrac{L_1 D^2 (6+7\log(2n))}{6} \sum\limits_{i=0}^{k-1} \tfrac{a_{i+1}^2}{A_{i+1}\sqrt{m_i}} \Big) \\ & \overset{(28)}{=} \quad \tfrac{1}{A_k}\Big(\Big[\tfrac{3L_2 D^3 + L_1 D^2(6+7\log(2n))}{6} \Big] \sum\limits_{i=0}^{k-1} \tfrac{a_{i+1}^3}{A_{i+1}^2} \\ & \qquad + D\sqrt{6L_1} \sum\limits_{i=0}^{k-1} \tfrac{a_{i+1}^2}{A_{i+1}} \big(\sqrt{2\mathbb{E}[F(x_i) - F^*]} + \sqrt{\mathbb{E}[F(z_i) - F^*]} \big) \Big). \end{aligned}$$

Now, let us set $A_{i+1} := (i+1)^3$, and thus $a_{i+1} := (i+1)^3 - i^3 \leq 3(i+1)^2$, so we have

$$\begin{aligned} \mathbb{E}[F(x_k) - F^*] \quad & \leq \quad \tfrac{\alpha + \beta(\sqrt{2}+1)(F(x_0) - F^*)}{k^2} \\ & \qquad + \tfrac{\beta}{k^3} \sum\limits_{i=1}^{k-1} \Big((i+1)\big(\sqrt{2\mathbb{E}[F(x_i) - F^*]} + \sqrt{\mathbb{E}[F(z_i) - F^*]} \big) \Big), \end{aligned} \tag{49}$$

where

$$\alpha \quad := \quad 27 \cdot \Big[\tfrac{3L_2 D^3 + L_1 D^2 (6+7\log(2n))}{6} \Big], \qquad \beta \quad := \quad 9 \cdot D\sqrt{6L_1}.$$

We are going to prove by induction, for every $k \geq 1$

$$\mathbb{E}[F(x_k) - F^*] \quad \leq \quad \tfrac{c}{k^2}, \tag{50}$$

with

$$\begin{aligned} c \quad & := \quad \big(4\beta + \sqrt{\alpha + 3\beta(F(x_0) - F^*) + 16\beta^2} \big)^2 \quad \leq \quad 74\beta^2 + 2\alpha + 6\beta(F(x_0) - F^*) \\ & = \quad \mathcal{O}\big(L_2 D^3 + L_1 D^2 (1 + \log(n)) + L_1^{1/2} D (F(x_0) - F^*) \big). \end{aligned} \tag{51}$$

Hence, if (50) is true, then we essentially obtain the claim of the theorem. For $k = 1$, (50) follows directly from (49). Assume that (50) holds for all $1 \leq i \leq k$, and consider iteration $k+1$:

$$\begin{aligned} \mathbb{E}[F(x_{k+1}) - F^*] \quad & \overset{(49),(50)}{\leq} \quad \tfrac{\alpha + \beta(\sqrt{2}+1)(F(x_0) - F^*)}{k^2} + \tfrac{\beta}{k^3} \sum\limits_{i=1}^{k} \Big((i+1)\Big(\tfrac{\sqrt{2c}}{i} + \tfrac{\sqrt{c}}{\pi(i)} \Big) \Big) \\ & \overset{(*)}{\leq} \quad \tfrac{\alpha + \beta(\sqrt{2}+1)(F(x_0) - F^*)}{k^2} + \tfrac{\beta\sqrt{c}}{k^3} \sum\limits_{i=1}^{k} \Big((i+1)\Big(\tfrac{2\sqrt{2}+4}{i+1} \Big) \Big) \\ & = \quad \tfrac{\alpha + (\sqrt{2}+1)\beta(F(x_0) - F^*) + (2\sqrt{2}+4)\beta\sqrt{c}}{k^2} \\ & \leq \quad \tfrac{\alpha + 3\beta(F(x_0) - F^*) + 8\beta\sqrt{c}}{k^2} \quad \overset{(51)}{=} \quad \tfrac{c}{k^2}, \end{aligned}$$

where in $(*)$ we have used two simple bounds: $i \leq 2\pi(i)$, and $i + 1 \leq 2i$, valid for all $i \geq 1$. □

E Extra experiments

In this section, we provide additional experimental results for the problem of training Logistic Regression model, regularized by ℓ_2-ball constraints: Figure 4 for the exact methods, and Figure 6 for the stochastic algorithms.

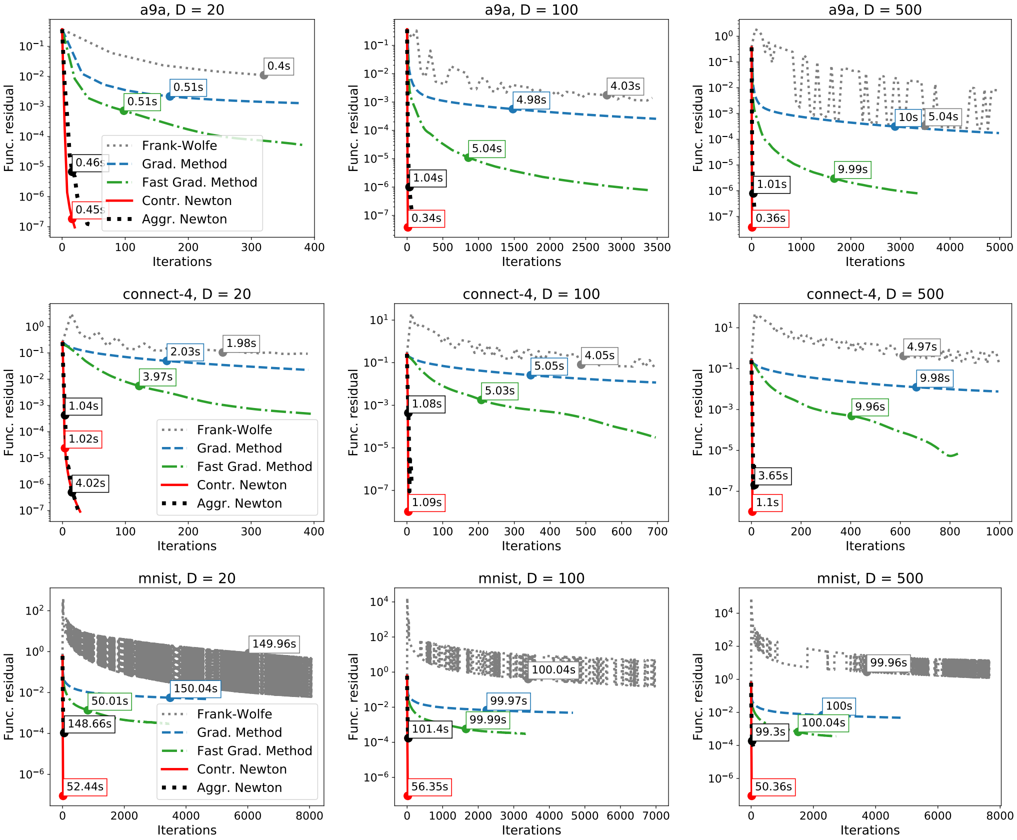

Figure 4: Training logistic regression, datasets: *a9a* $(M = 32561, n = 123)$, *connect-4* $(M = 67557, n = 126)$, *mnist* $(M = 60000, n = 780)$.

We see, that the second-order schemes usually outperforms first-order methods, in terms of the number of iterations, and the number of epochs. Despite the fact, that the Newton step is more expensive, in many situations we see superiority of the second-order schemes in terms of the total computational time as well.

Comparing Contracting-Domain Newton Method (Algorithm 1), and Aggregating Newton Method (Algorithm 3), we conclude that both of the algorithms show reasonably good performance in practice. The latter one works a bit slower. However, the aggregation of the Hessians helps to improve numerical stability. On Figure 5, we demonstrate influence of the parameter of inner accuracy (EPS), which we use in our subsolver, on the convergence of the algorithms. We see much more robust behaviour for Aggregating Newton Method, while the first algorithm can potentially stop, or even start to diverge, if the parameter is chosen in a wrong way.

To compute one step of our second-order methods for this task, we need to solve subproblem (20) for $p = 2$. This is minimization of quadratic function over the standard Euclidean ball. First, we compute *tridiagonal* decomposition of the Hessian (it requires $\mathcal{O}(n^3)$ arithmetical operations). Then, we solve the dual to our subproblem (which is maximization of one-dimensional concave function) by classical Newton iterations (the cost of each iteration is $\mathcal{O}(n)$). For more details, see Chapter 7 in [9].

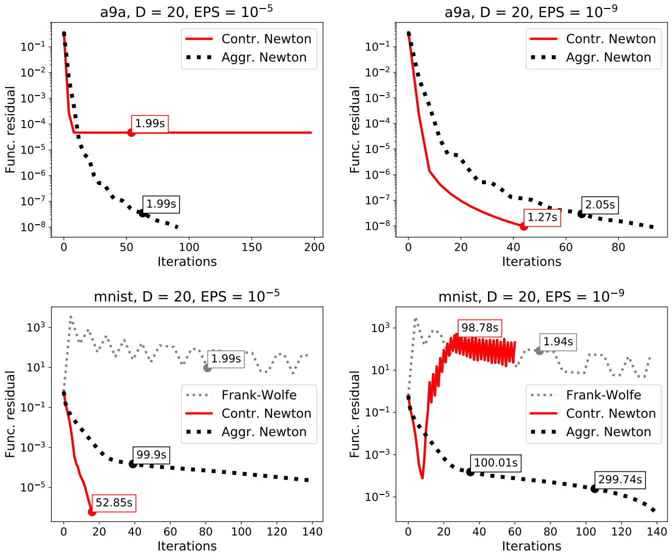

Figure 5: Influence of the parameter of inner accuracy.

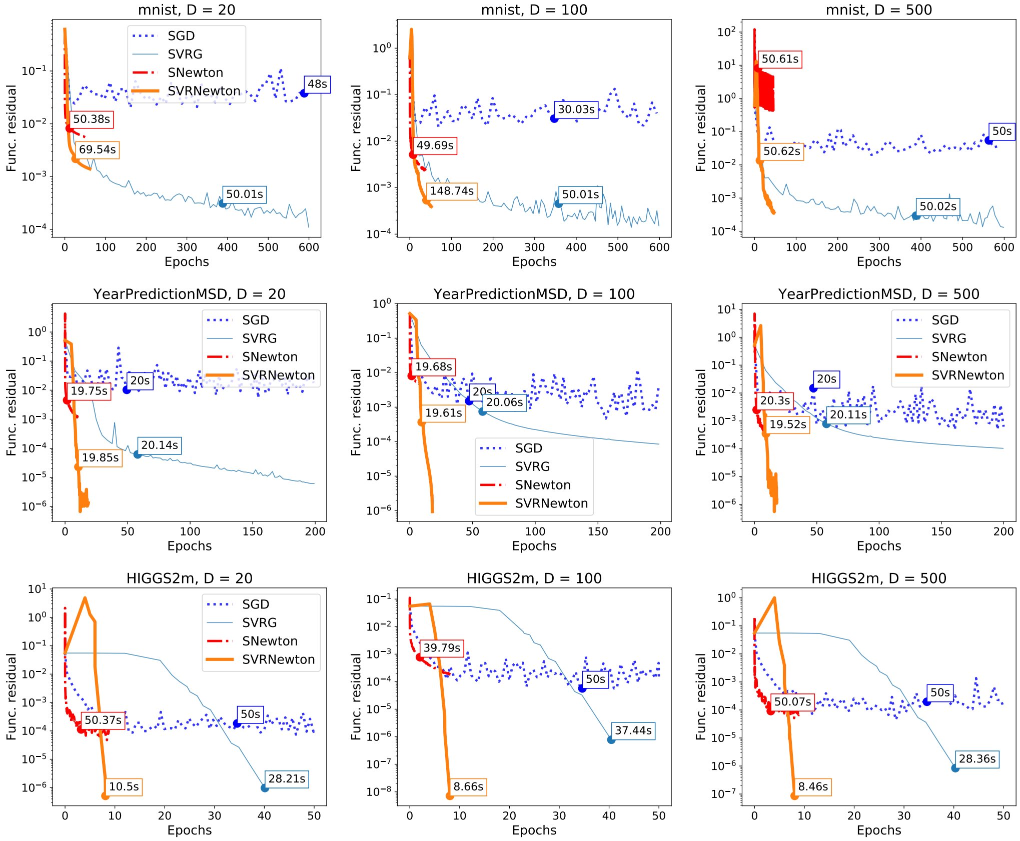

Figure 6: Stochastic methods for training logistic regression, datasets: *mnist* $(M = 60000, n = 780)$, *YearPredictionMSD* $(M = 463715, n = 90)$, *HIGGS2m* $(M = 2 \cdot 10^6, n = 28)$.