

# A STOCHASTIC SEMISMOOTH NEWTON METHOD FOR NONSMOOTH NONCONVEX OPTIMIZATION

ANDRE MILZAREK\*, XIANTAO XIAO<sup>†</sup>, SHICONG CEN<sup>‡</sup>, ZAIWEN WEN<sup>§</sup>, AND MICHAEL ULBRICH<sup>¶</sup>

**Abstract.** In this work, we present a globalized stochastic semismooth Newton method for solving stochastic optimization problems involving smooth nonconvex and nonsmooth convex terms in the objective function. We assume that only noisy gradient and Hessian information of the smooth part of the objective function is available via calling stochastic first and second order oracles. The proposed method can be seen as a hybrid approach combining stochastic semismooth Newton steps and stochastic proximal gradient steps. Two inexact growth conditions are incorporated to monitor the convergence and the acceptance of the semismooth Newton steps and it is shown that the algorithm converges globally to stationary points in expectation. Moreover, under standard assumptions and utilizing random matrix concentration inequalities, we prove that the proposed approach locally turns into a pure stochastic semismooth Newton method and converges  $r$ -superlinearly with high probability. We present numerical results and comparisons on  $\ell_1$ -regularized logistic regression and nonconvex binary classification that demonstrate the efficiency of our algorithm.

**Key words.** nonsmooth stochastic optimization, stochastic approximation, semismooth Newton method, stochastic second order information, global and local convergence.

**AMS subject classifications.** 49M15, 65C60, 65K05, 90C06

**1. Introduction.** In this paper, we propose and analyze a stochastic semismooth Newton framework for solving general nonsmooth, nonconvex optimization problems of the form

$$(1.1) \quad \min_{x \in \mathbb{R}^n} \psi(x) := f(x) + r(x),$$

where  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is a (twice) continuously differentiable but possibly nonconvex function and  $r : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is a convex, lower semicontinuous, and proper mapping. Although the function  $f$  is smooth, we assume that a full evaluation of  $f$  and an exact computation of the gradient and Hessian values  $\nabla f(x)$  and  $\nabla^2 f(x)$  is either not completely possible or too expensive in practice. Instead, we suppose that only noisy gradient and Hessian information is available which can be accessed via calls to *stochastic first (SFO)* and *second order oracles (SSO)*. Composite problems of the type (1.1) arise frequently in statistics and in large-scale statistical learning, see, e.g., [37, 48, 8, 76, 15], and in many other applications. In these examples and problems, the smooth mapping  $f$  is typically of the form

$$(1.2) \quad f(x) := \mathbb{E}[F(x, \xi)] = \int_{\Omega} F(x, \xi(\omega)) \, d\mathbb{P}(\omega), \quad \text{or} \quad f(x) := \frac{1}{N} \sum_{i=1}^N f_i(x),$$

where  $\xi : \Omega \rightarrow W$  is a random variable defined on a given probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ ,  $W$  is a measurable space, and  $F : \mathbb{R}^n \times W \rightarrow \mathbb{R}$  and the component functions  $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $i = 1, \dots, N$ , correspond to certain loss models. More specifically, in the latter case, when the nonsmooth term  $r \equiv 0$  vanishes, the problem (1.1) reduces to the so-called and well-studied

---

\*Beijing International Center for Mathematical Research, BICMR, Peking University, Beijing, China (andremilzarek@bicmr.pku.edu.cn). Research supported by the Boya Postdoctoral Fellowship.

<sup>†</sup>School of Mathematical Sciences, Dalian University of Technology, Dalian, China (txtiao@dlut.edu.cn).

<sup>‡</sup>School of Mathematical Sciences, Peking University, Beijing, China (tsen9731@pku.edu.cn).

<sup>§</sup>Beijing International Center for Mathematical Research, BICMR, Peking University, Beijing, China (wenzw@pku.edu.cn). Research supported in part by the NSFC grant 11421101 and by the National Basic Research Project under the grant 2015CB856002.

<sup>¶</sup>Chair of Mathematical Optimization, Department of Mathematics, Technical University Munich, Garching b. München, Germany (mulbrich@ma.tum.de).

*empirical risk minimization problem*

$$(1.3) \quad \min_{x \in \mathbb{R}^n} f(x), \quad f(x) := \frac{1}{N} \sum_{i=1}^N f_i(x).$$

Since the distribution  $\mathbb{P}$  in (1.2) might not be fully known and the number of components  $N$  in (1.3) can be extremely large, stochastic approximation techniques, such as the mentioned stochastic oracles, have become an increasingly important tool in the design of efficient and computationally tractable numerical algorithms for the problems (1.1) and (1.3), [58, 29, 76, 30, 31, 86, 85]. Moreover, in various interesting problems such as deep learning, dictionary learning, training of neural networks, and classification tasks with nonconvex activation functions, [49, 48, 8, 23, 42, 73, 33], the loss function  $f$  is nonconvex, which represents another major challenge for stochastic optimization approaches. For further applications and additional connections to simulation-based optimization, we refer to [29, 31].

**1.1. Related Work.** The pioneering idea of utilizing stochastic approximations and the development of the associated, classical stochastic gradient descent method (SGD) for problem (1.3) and other stochastic programs can be traced back to the seminal work of Robbins and Monro [69]. Since then, a plethora of stochastic optimization methods, strategies, and extensions has been studied and proposed for different problem formulations and under different basic assumptions. In the following, we give a brief overview of related research directions and related work.

*First order methods.* Advances in the research of stochastic first order methods for the smooth empirical risk problem (1.3) are numerous and we will only name a few recent directions here. Lately, based on the popularity and flexible applicability of the basic SGD method, a strong focus has been on the development and analysis of more sophisticated stochastic first order oracles to reduce the variance induced by gradient sampling and to improve the overall performance of the underlying SGD method. Examples of algorithms that utilize such variance reduction techniques include SVRG [40], SDCA [77], SAG [74], and SAGA [22]. Moreover, Friedlander and Schmidt [27] analyze the convergence of a mini-batch stochastic gradient method for strongly convex  $f$ , in which the sampling rates are increased progressively. Incorporating different acceleration strategies, the first order algorithms Catalyst [46] and Katyusha [3] further improve the iteration complexity of the (proximal) SGD method.

Several of the mentioned algorithms can also be extended to the nonsmooth setting  $r \neq 0$  by using the proximity operator of  $r$  and associated stochastic proximal gradient steps, see, e.g., the perturbed proximal gradient method [6] studied by Atchadé et al., prox-SVRG [90], prox-SAGA [22], and prox-SDCA [78]. AdaGrad [24] is another extension of the classical SGD method that utilizes special adaptive step size strategies. Under the assumption that  $r$  is block separable, Richtárik and Takáč [68] develop a randomized block-coordinate descent method for (1.1). An accelerated variant of this approach is investigated by Lin et al. [47].

The methods discussed so far either require convexity of  $f$  or of each of the component functions  $f_i$  or even stronger assumptions. Ghadimi and Lan [29, 30] generalize the basic and accelerated SGD method to solve nonconvex and smooth minimization problems. Allen-Zhu and Hazan [5] and Reddi et al. [66] analyze stochastic variance reduction techniques for the nonconvex version of problem (1.3). Moreover, Reddi et al. [67] and Allen-Zhu [4] further extend existing stochastic first order methods to find approximate stationary points of the general nonconvex, nonsmooth model (1.1). In [31], Ghadimi et al. discuss complexity and convergence results for a mini-batch stochastic projected gradient algorithm for problem (1.1). Xu and Yin [95] present and analyze a block stochastic gradient method for convex, nonconvex, and nonsmooth variants of the problem (1.1).

*Quasi-Newton and second order methods.* Recently, in order to accelerate and robustify the convergence of first order algorithms, stochastic second order methods have gained much attention. So far, the majority of stochastic second order methods is designed for the smooth problem (1.3) and is based on variants of the sub-sampled Newton method in which approximations of the gradient and Hessian of  $f$  are generated by selecting only a sub-sample or mini-batch of the components  $\nabla f_i$  and  $\nabla^2 f_i$ ,  $i = 1, \dots, N$ . In [16, 17], assuming positive definiteness of the sub-sampled Hessians, the authors analyze the convergence of a sub-sampled Newton-CG method and discuss strategies for selecting the sample sets. Erdogdu and Montanari [25] derive convergence rates of a projected sub-sampled Newton method with rank thresholding. In [71, 72], Roosta-Khorasani and Mahoney establish non-asymptotic, probabilistic global and local convergence rates for sub-sampled Newton methods by applying matrix concentration inequalities. Xu et al. [94] present convergence and complexity results for a sub-sampled Newton-type approach with non-uniform sampling. Bollapragada et al. [13] consider a sub-sampled Newton method for problems with the more general loss function given in (1.2) and derive  $r$ -superlinear convergence rates in expectation using a “bounded moment” condition to overcome the nonequivalence of norms in infinite dimensions. In [85], Wang and Zhang propose an algorithm that combines the advantages of variance reduction techniques and sub-sampled Newton methods. Convergence properties are studied under the assumption that  $f$  is strongly convex and the Hessians  $\nabla^2 f_i$  are Lipschitz continuous (with a uniform constant). Based on the existence of a suitable square-root decomposition of the Hessian, Pilanci and Wainwright [62] propose a Newton sketch method for general convex, smooth programs. In [10], the numerical performance of the Newton sketch method and different sub-sampled Newton approaches is compared. Furthermore, based on unbiased estimators of the inverse Hessian, a stochastic method called LiSSA is studied in [1]. A recent discussion of different stochastic second order algorithms can also be found in [97].

Stochastic quasi-Newton methods represent another large and important class of stochastic numerical algorithms for problem (1.3). Typically, these methods combine specific sub-sampling schemes for  $\nabla f$  with randomized BFGS or BFGS-type updates to approximate the Hessian  $\nabla^2 f$ . In [75], the authors propose a basic stochastic quasi-Newton algorithm for quadratic loss functions. Bordes et al. [14] present a quasi-Newton approach that is based on diagonal curvature estimation. Mokhtari and Ribeiro [54] investigate a regularized stochastic BFGS method for solving strongly convex problems. In [18], Byrd et al. consider a stochastic limited-memory BFGS (L-BFGS) algorithm that incorporates exact Hessian information of the functions  $f_i$  to build the BFGS-type updates. The stochastic L-BFGS method discussed in [57] uses variance reduction techniques to improve its convergence and performance. Moreover, Gower et al. [34] establish linear convergence of a stochastic block L-BFGS method if the functions  $f_i$  are strongly convex.

In contrast, the number of stochastic second order algorithms for smooth but nonconvex problems seems to be still quite limited. Based on a damping strategy for BFGS-type updates introduced by Powell [63] and using general stochastic first order oracles, Wang et al. [86] propose a stochastic L-BFGS method for smooth, nonconvex problems. Under the assumption that the full gradient of the objective function is available, Xu et al. [92] derive worst-case optimal iteration complexity results for an adaptive cubic regularization method with inexact or sub-sampled Hessian information. Generalizations and further aspects of this approach have been considered very recently in [93, 96].

Finally, in [85], Wang and Zhang mention an extension of their hybrid method to the non-smooth setting. A similar and related idea has also been presented in [79]. In particular, these approaches can be interpreted as stochastic variants of the proximal Newton method [44] for the general problem (1.1). Nevertheless, strong and uniform convexity assumptions are still required to guarantee convergence and well-definedness of the inner steps and subproblems.

Let us note that there is also a vast literature on incremental methods for (1.1) and (1.3), see, e.g., [11, 35], which is beyond the scope of this paper.

**1.2. Contents and Contributions.** In this paper, we develop a stochastic second order framework for the general optimization problem (1.1). Our basic idea is to apply a semismooth Newton method to approximately solve the nonsmooth fixed point-type equation

$$(1.4) \quad F^\Lambda(x) := x - \text{prox}_r^\Lambda(x - \Lambda^{-1} \nabla f(x)) = 0, \quad \Lambda \in \mathbb{S}_{++}^n,$$

which represents a reformulation of the associated first order optimality conditions of problem (1.1). Specifically, we will consider stochastic variants of the nonsmooth residual (1.4) and of the semismooth Newton method, in which the gradient and Hessian of  $f$  are substituted by stochastic oracles. Motivated by deterministic Newton-type approaches [52, 51, 91], our proposed method combines stochastic semismooth Newton steps, stochastic proximal gradient steps, and a globalization strategy that is based on controlling the acceptance of the Newton steps via growth conditions. In this way, the resulting stochastic algorithm can be guaranteed to converge globally in expectation and almost surely, i.e., for a generated sequence of iterates  $(x^k)_k$ , we have

$$\mathbb{E}[\|F^\Lambda(x^k)\|^2] \rightarrow 0 \quad \text{and} \quad F^\Lambda(x^k) \rightarrow 0 \quad \text{almost surely,} \quad k \rightarrow \infty.$$

Furthermore, inspired by [71, 72] and using matrix concentration inequalities [81], we prove that transition to fast local  $r$ -linear or  $r$ -superlinear convergence can be established with high probability if the sample sizes and sampling rates are chosen appropriately and increase sufficiently fast. To the best of our knowledge, rigorous extensions of existing stochastic second order methods to the nonsmooth, nonconvex setting considered in this work do not seem to be available. We now briefly summarize some of the main challenges and contributions.

- We provide a unified convergence theory for the proposed stochastic Newton method covering global and local aspects and transition to fast local convergence. In contrast to many other works, convexity of the smooth function  $f$  or of the objective function  $\psi$  is not required in our analysis.
- In order to ensure global convergence and based on an acceptance test, the algorithm is allowed to switch between Newton and proximal gradient steps. Hence, a priori, it is not clear whether the generated iterates correspond to measurable random variables or to a stochastic process. This structural mechanism is significantly different from other existing stochastic approaches and will be discussed in detail in section 3.
- Our algorithmic approach and theoretical results are applicable for general stochastic oracles. Consequently, a large variety of approximation schemes, such as basic subsampling strategies or more elaborate variance reduction techniques, [40, 90, 67, 86], can be used within our framework. In particular, in our numerical experiments, we investigate a variance reduced version of our method. Similar to [86], the numerical results indicate that the combination of second order information and variance reduction techniques is also very effective in the nonsmooth setting. We note that the proposed method (using different stochastic oracles) performs quite well in comparison with other state-of-the-art algorithms in general.

**1.3. Organization.** This paper is organized as follows. Our specific stochastic setup, a derivation of the equation (1.4), and the main algorithm are stated in section 2. The global and local convergence results are presented in section 3 and section 4, respectively. Finally, in section 5, we report and discuss our numerical comparisons and experiments in detail.

**1.4. Notation.** For any  $n \in \mathbb{N}$ , we set  $[n] := \{1, \dots, n\}$ . By  $\langle \cdot, \cdot \rangle$  and  $\| \cdot \| := \| \cdot \|_2$  we denote the standard Euclidean inner product and norm. The set of symmetric and positive definite  $n \times n$  matrices is denoted by  $\mathbb{S}_{++}^n$ . For a given matrix  $\Lambda \in \mathbb{S}_{++}^n$ , we define the inner product  $\langle x, y \rangle_\Lambda := \langle x, \Lambda y \rangle = \langle \Lambda x, y \rangle$  and  $\|x\|_\Lambda := \sqrt{\langle x, x \rangle_\Lambda}$ . The set  $\text{lev}_\alpha f := \{x : f(x) \leq \alpha\}$  denotes the lower level set of a function  $f$  at level  $\alpha \in \mathbb{R}$ . For a given set  $S \subset \mathbb{R}^n$ , the set  $\text{cl } S$  denotes the closure of  $S$  and  $\mathbb{1}_S : \mathbb{R}^n \rightarrow \{0, 1\}$  is the associated characteristic function of  $S$ . For  $p \in (0, \infty)$  the space  $\ell_+^p$  consists of all sequences  $(x_n)_{n \geq 0}$  satisfying  $x_n \geq 0$ ,  $n \geq 0$ , and  $\sum x_n^p < \infty$ . Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a given probability space. The space  $L^p(\Omega) := L^p(\Omega, \mathbb{P})$ ,  $p \in [1, \infty]$ , denotes the standard  $L^p$  space on  $\Omega$ . We write  $\xi \in \mathcal{F}$  for “ $\xi$  is  $\mathcal{F}$ -measurable”. Moreover, we use  $\sigma(\xi^1, \dots, \xi^k)$  to denote the  $\sigma$ -algebra generated by the family of random variables  $\xi^1, \dots, \xi^k$ . For a random variable  $\xi \in L^1(\Omega)$  and a sub- $\sigma$ -algebra  $\mathcal{H} \subseteq \mathcal{F}$ , the conditional expectation of  $\xi$  given  $\mathcal{H}$  is denoted by  $\mathbb{E}[\xi | \mathcal{H}]$ . The conditional probability of  $S \in \mathcal{F}$  given  $\mathcal{H}$  is defined as  $\mathbb{P}(S | \mathcal{H}) := \mathbb{E}[\mathbb{1}_S | \mathcal{H}]$ . We use the abbreviations “a.e.” and “a.s.” for “almost everywhere” and “almost surely”, respectively.

## 2. A Stochastic Semismooth Newton Method.

**2.1. Probabilistic Setting and Preliminaries.** In this section, we introduce several basic definitions and preparatory results. We start with an overview of the stochastic setting and the sampling strategy.

**2.1.1. Stochastic Setup.** Although the function  $f$  is smooth, we assume that an exact or full evaluation of the gradient  $\nabla f$  and Hessian  $\nabla^2 f$  is not possible or is simply too expensive. Hence, we will work with *stochastic first* ( $SFO$ ) and *second order oracles* ( $SSO$ ),

$$\mathcal{G} : \mathbb{R}^n \times \Xi \rightarrow \mathbb{R}^n, \quad \mathcal{H} : \mathbb{R}^n \times \Xi \rightarrow \mathbb{S}^n$$

to approximate gradient and Hessian information. Specifically, given an underlying probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and a measurable space  $(\Xi, \mathcal{X})$ , we generate two mini-batches of random samples

$$s^k := \{s_1^k, \dots, s_{\mathbf{n}_k^g}^k\} \quad \text{and} \quad t^k := \{t_1^k, \dots, t_{\mathbf{n}_k^h}^k\}$$

and calculate the stochastic approximations  $\mathcal{G}(x, s_i^k) \approx \nabla f(x)$  and  $\mathcal{H}(x, t_j^k) \approx \nabla^2 f(x)$  in each iteration. Here, we assume that the space  $\Omega$  is sufficiently rich allowing us to model and describe the (possibly independent) sample batches  $s^k, t^k$  and other associated stochastic processes in a unified way. Moreover, each of the samples  $s_i^k, t_j^k : \Omega \rightarrow \Xi$ ,  $i \in [\mathbf{n}_k^g]$ ,  $j \in [\mathbf{n}_k^h]$ , corresponds to an  $(\mathcal{F}, \mathcal{X})$ -measurable, random mapping and  $\mathbf{n}_k^g$  and  $\mathbf{n}_k^h$  denote the chosen sample rates or sample sizes of  $s^k$  and  $t^k$ , respectively. Similar to [21, 95, 30, 31, 86], we then construct a mini-batch-type, stochastic gradient  $G_{s^k}(x)$  and Hessian  $H_{t^k}(x)$  as follows

$$(2.1) \quad G_{s^k}(x) := \frac{1}{\mathbf{n}_k^g} \sum_{i=1}^{\mathbf{n}_k^g} \mathcal{G}(x, s_i^k), \quad H_{t^k}(x) := \frac{1}{\mathbf{n}_k^h} \sum_{i=1}^{\mathbf{n}_k^h} \mathcal{H}(x, t_i^k).$$

Throughout this work, we assume that the stochastic oracles  $\mathcal{G}$  and  $\mathcal{H}$  are *Carathéodory functions*<sup>1</sup>. Further assumptions on the stochastic setting will be introduced later in [subsection 3.1](#) and [subsection 4.1](#).

We will also sometimes drop the index  $k$  from the mini-batches  $s^k, t^k$  and sample sizes  $\mathbf{n}_k^g, \mathbf{n}_k^h$  when we consider a general pair of batches  $s$  and  $t$ .

<sup>1</sup>A mapping  $F : \mathbb{R}^n \times \Xi \rightarrow \mathbb{R}$  is called *Carathéodory function* if  $F(\cdot, z) : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuous for all  $z \in \Xi$  and if  $F(x, \cdot) : \Xi \rightarrow \mathbb{R}$  is measurable for all  $x \in \mathbb{R}^n$ .

**2.1.2. Definitions and First Order Optimality.** In the following, we derive first order optimality conditions for the composite problem (1.1). Suppose that  $x^* \in \text{dom } r$  is a local solution of problem (1.1). Then,  $x^*$  satisfies the mixed-type variational inequality

$$(2.2) \quad \langle \nabla f(x^*), x - x^* \rangle + r(x) - r(x^*) \geq 0, \quad \forall x \in \mathbb{R}^n.$$

By definition, the latter condition is equivalent to  $-\nabla f(x^*) \in \partial r(x^*)$ , where  $\partial r$  denotes the convex subdifferential of  $r$ . We now introduce the well-known *proximal mapping*  $\text{prox}_r^\Lambda : \mathbb{R}^n \rightarrow \mathbb{R}^n$  of  $r$ . For an arbitrary parameter matrix  $\Lambda \in \mathbb{S}_{++}^n$ , the proximity operator  $\text{prox}_r^\Lambda(x)$  of  $r$  at  $x$  is defined as

$$(2.3) \quad \text{prox}_r^\Lambda(x) := \arg \min_{y \in \mathbb{R}^n} r(y) + \frac{1}{2} \|x - y\|_\Lambda^2.$$

The proximity operator is a  $\Lambda$ -firmly nonexpansive mapping, i.e., it satisfies

$$\|\text{prox}_r^\Lambda(x) - \text{prox}_r^\Lambda(y)\|_\Lambda^2 \leq \langle \text{prox}_r^\Lambda(x) - \text{prox}_r^\Lambda(y), x - y \rangle_\Lambda, \quad \forall x, y \in \mathbb{R}^n.$$

Consequently,  $\text{prox}_r^\Lambda$  is Lipschitz continuous with modulus 1 with respect to the norm  $\|\cdot\|_\Lambda$ . We refer to [56, 20, 8, 9] for more details and (computational) properties. Let us further note that the proximity operator can also be uniquely characterized by the optimality conditions of the underlying optimization problem (2.3), i.e.,

$$(2.4) \quad \text{prox}_r^\Lambda(x) \in x - \Lambda^{-1} \cdot \partial r(\text{prox}_r^\Lambda(x)).$$

Using this characterization, condition (2.2) can be equivalently rewritten as follows:

$$(2.5) \quad F^\Lambda(x^*) = 0, \quad \text{where} \quad F^\Lambda(x) := x - \text{prox}_r^\Lambda(x - \Lambda^{-1} \nabla f(x)).$$

We call  $x \in \mathbb{R}^n$  a *stationary point* of problem (1.1), if it is a solution of the nonsmooth equation (2.5). If the problem is convex, e.g., if  $f$  is a convex function, then every stationary point is automatically a local and global solution of (1.1). The fixed point-type equation (2.5) forms the basis of the proximal gradient method, [53, 28, 20, 60], which has been studied intensively during the last decades.

For an arbitrary sample  $s$ , the corresponding stochastic residual is given by

$$F_s^\Lambda(x) := x - \text{prox}_r^\Lambda(x - \Lambda^{-1} G_s(x)).$$

We will also use  $u_s^\Lambda(x) := x - \Lambda^{-1} G_s(x)$  and  $p_s^\Lambda(x) := \text{prox}_r^\Lambda(u_s^\Lambda(x))$  to denote the stochastic (proximal) gradient steps.

**2.2. Algorithmic Framework.** In this section, we describe our algorithmic approach in detail. The overall idea is to use a stochastic semismooth Newton method to calculate an approximate solution of the optimality system

$$F^\Lambda(x) = 0.$$

The associated Newton step  $d^k$  at iteration  $k$  is then given by the linear system of equations

$$(2.6) \quad M_k d^k = -F_{s^k}^{\Lambda_k}(x^k), \quad M_k \in \mathcal{M}_{s^k, t^k}^{\Lambda_k}(x^k).$$

Here, we consider the following set of generalized derivatives

$$(2.7) \quad \mathcal{M}_{s, t}^\Lambda(x) := \{M \in \mathbb{R}^{n \times n} : M = (I - D) + D\Lambda^{-1}H_t(x), \quad D \in \partial \text{prox}_r^\Lambda(u_s^\Lambda(x))\},$$



**Algorithm 1:** A Stochastic Semismooth Newton Method

- 
- 1 Initialization: Choose an initial point  $x^0 \in \text{dom } r$ ,  $\theta_0 \in \mathbb{R}_+$ , and mini-batches  $s^0, t^0$ .  
 Select sample sizes  $(\mathbf{n}_k^g)_k, (\mathbf{n}_k^h)_k$ , parameter matrices  $(\Lambda_k)_k \subset \mathbb{S}_{++}^n$ , and step sizes  $(\alpha_k)_k$ . Choose  $\eta, p \in (0, 1)$ ,  $\beta > 0$ , and  $(\nu_k)_k, (\varepsilon_k^1)_k, (\varepsilon_k^2)_k$ . Set iteration  $k := 0$ .
  - 2 **while** *did not converge* **do**
  - 3   Compute  $F_{s^k}^{\Lambda_k}(x^k)$  and choose  $M_k \in \mathcal{M}_{s^k, t^k}^{\Lambda_k}(x^k)$ . For all  $i = 1, \dots, \mathbf{n}_{k+1}^g$  and  $j = 1, \dots, \mathbf{n}_{k+1}^h$  select new samples  $s_i^{k+1}, t_j^{k+1}$ .
  - 4   Compute the Newton step  $d^k$  by solving
 
$$M_k d^k = -F_{s^k}^{\Lambda_k}(x^k).$$
  - 5   Set  $z_n^k = x^k + d^k$ . If the conditions  $z_n^k \in \text{dom } r$ , (2.8), and (2.9) are satisfied, skip step 6 and set  $x^{k+1} = z_n^k$ ,  $\theta_{k+1} = \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\|$ . Otherwise go to step 6.
  - 6   Set  $v^k = -F_{s^k}^{\Lambda_k}(x^k)$ ,  $x^{k+1} = x^k + \alpha_k v^k$ , and  $\theta_{k+1} = \theta_k$ .
  - 7   Set  $k \leftarrow k + 1$ .
- 

where  $\partial \text{prox}_r^\Lambda(u_s^\Lambda(x))$  denotes the Clarke subdifferential of  $\text{prox}_r^\Lambda$  at the point  $u_s^\Lambda(x)$ . The set  $\mathcal{M}_{s,t}^\Lambda(x)$  depends on the stochastic gradient and on the stochastic Hessian defined in (2.1). Moreover, the samples  $s^k, t^k$  and the matrix  $\Lambda_k$  used in (2.6) may change in each iteration, see also Remark 3.6. We further note that, in practice, the system (2.6) can be solved inexactly via iterative approaches such as the conjugate gradient or other Krylov subspace methods.

In the deterministic setting, the set  $\mathcal{M}_{s,t}^\Lambda(x)$  reduces to  $\mathcal{M}^\Lambda(x) := \{M = (I - D) + D\Lambda^{-1}\nabla^2 f(x), D \in \partial \text{prox}_r^\Lambda(u^\Lambda(x))\}$  with  $u^\Lambda(x) = x - \Lambda^{-1}\nabla f(x)$ . In general,  $\mathcal{M}^\Lambda(x)$  does not coincide with Clarke's subdifferential  $\partial F^\Lambda(x)$ . As shown in [19], we can only guarantee  $\partial F^\Lambda(x)h \subseteq \text{co}(\mathcal{M}^\Lambda(x)h)$  for  $h \in \mathbb{R}^n$ . However, the set-valued mapping  $\mathcal{M}^\Lambda : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$  defines a so-called (*strong*) *linear Newton approximation* at  $x$  if the proximity operator  $\text{prox}_r^\Lambda$  is (strongly) semismooth at  $u^\Lambda(x)$ . In particular,  $\mathcal{M}^\Lambda$  is upper semicontinuous and compact-valued. We refer to [26, Chapter 7] and [61] for more details. We also note that the chain rule for semismooth functions implies that  $F^\Lambda(x)$  is semismooth at  $x$  with respect to  $\mathcal{M}^\Lambda(x)$  if  $\text{prox}_r^\Lambda$  is semismooth at  $u^\Lambda(x)$ . Furthermore, in various important examples including, e.g.,  $\ell_1$ - or nuclear norm-regularized optimization, group sparse problems or semidefinite programming, the associated proximal mapping  $\text{prox}_r^\Lambda$  can be shown to be (strongly) semismooth and there exist explicit and computationally tractable representations of the generalized derivatives  $D \in \partial \text{prox}_r^\Lambda(\cdot)$ , see [61, 51, 91] for a detailed discussion.

In order to control the acceptance of the Newton steps and to achieve global convergence of our algorithm, we introduce the following growth conditions for the trial step  $z_n^k = x^k + d^k$ :

$$(2.8) \quad \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\| \leq (\eta + \nu_k) \cdot \theta_k + \varepsilon_k^1,$$

$$(2.9) \quad \psi(z_n^k) \leq \psi(x^k) + \beta \cdot \theta_k^{1-p} \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\|^p + \varepsilon_k^2.$$

If the trial point  $z_n^k$  satisfies both conditions and is feasible, i.e., if  $z_n^k \in \text{dom } r$ , we accept it and compute the new iterate  $x^{k+1}$  via  $x^{k+1} = z_n^k$ . The parameter sequences  $(\nu_k)_k, (\varepsilon_k^1)_k$ , and  $(\varepsilon_k^2)_k$  are supposed to be nonnegative and summable and can be chosen during the initialization or during the iteration process. Furthermore, the parameter  $\theta_k$  keeps track of the norm of the residual  $F_{s^i}^{\Lambda_i}(x^i)$  of the last *accepted* Newton iterate  $x^i$ ,  $i < k$ , and is updated after a successful Newton step. The parameters  $\beta > 0, \eta, p \in (0, 1)$  are given constants. If the trial point  $z_n^k$  does not satisfy the conditions (2.8) and (2.9), we reject it and perform an

alternative proximal gradient step using the stochastic residual  $F_{s^k}^{\Lambda_k}(x^k)$  as an approximate descent direction. We also introduce a step size  $\alpha_k$  to damp the proximal gradient step and to guarantee sufficient decrease in the objective function  $\psi$ . A precise bound for the step sizes  $\alpha_k$  is derived in [Lemma 3.8](#). The details of the method are summarized in [Algorithm 1](#).

Our method can be seen as a hybrid of the semismooth Newton method and the standard proximal gradient method generalizing the deterministic Newton approaches presented in [\[52, 51\]](#) to the stochastic setting. Our globalization technique is inspired by [\[52\]](#), where a filter globalization strategy was proposed to control the acceptance of the Newton steps. Similar to [\[52, 51\]](#), we add condition [\(2.8\)](#) to monitor the behavior and convergence of the Newton steps. The second condition [\(2.9\)](#) (together with the feasibility condition  $z_n^k \in \text{dom } r$ ) is required to bound the possible  $\psi$ -ascent of intermediate Newton steps. In contrast to smooth optimization problems, descent-based damping techniques or step size selections, as used in, e.g., [\[16, 13, 18, 71, 72, 86\]](#), can not always guarantee sufficient  $\psi$ -descent of the semismooth Newton steps due to the nonsmooth nature of problem [\(1.1\)](#). This complicates the analysis and globalization of semismooth Newton methods in general. In practice, the second growth condition [\(2.9\)](#) can be restrictive since an evaluation of the full objective function is required. However, similar descent conditions also appeared in other globalization strategies for smooth problems, [\[71, 92, 93, 96\]](#). In the next section, we verify that [Algorithm 1](#) using the proposed growth conditions [\(2.8\)–\(2.9\)](#) converges globally in expectation. Moreover, in [Theorem 3.12](#), we establish global convergence of [Algorithm 1](#) without condition [\(2.9\)](#) in a strongly convex setting. Under standard assumptions and if the sample sizes  $n^g$  and  $n^h$  are chosen sufficiently large, we can further show that the conditions [\(2.8\)](#) and [\(2.9\)](#) are satisfied locally in a neighborhood of a stationary point with high probability. This enables us to derive fast local convergence results in probability. Let us note that the growth conditions [\(2.8\)–\(2.9\)](#) are checked using a new sample mini-batch  $s^{k+1}$ . Thus, only one gradient evaluation is required per iteration if the Newton step is accepted. We note that the feasibility condition  $z_n^k \in \text{dom } r$  can be circumvented by setting  $x^{k+1} = \mathcal{P}_{\text{dom } r}(z_n^k)$ , where  $\mathcal{P}_{\text{dom } r}$  denotes the projection onto the set  $\text{dom } r$ . We refer to [Remark 4.7](#) for a related discussion.

Let us mention that an alternative globalization is analyzed in [\[61, 80\]](#) where the authors propose the so-called forward-backward envelope (FBE) as a smooth merit function for problem [\(1.1\)](#). Since this framework requires an additional proximal gradient step (and thus, an additional gradient evaluation) after each iteration, we do not consider this approach here.

**3. Global Convergence.** In this section, we analyze the global convergence behavior of [Algorithm 1](#). We first present and summarize our main assumptions.

**3.1. Assumptions.** Throughout this paper, we assume that  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is continuously differentiable on  $\mathbb{R}^n$  and  $r : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is convex, lower semicontinuous, and proper. As already mentioned, we also assume that the oracles  $\mathcal{G}, \mathcal{H} : \mathbb{R}^n \times \Xi \rightarrow \mathbb{R}$  are Carathéodory functions. In the following, we further specify the assumptions on the functions  $f$  and  $r$ .

ASSUMPTION 3.1. *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be given as in [\(1.1\)](#). We assume:*

- (A.1) *The gradient mapping  $\nabla f$  is Lipschitz continuous on  $\mathbb{R}^n$  with modulus  $L > 0$ .*
- (A.2) *The objective function  $\psi$  is bounded from below on  $\text{dom } r$ .*
- (A.3) *There exist parameters  $\mu_f \in \mathbb{R}$  and  $\mu_r \geq 0$  with  $\bar{\mu} := \mu_f + \mu_r > 0$  such that the shifted functions  $f - \frac{\mu_f}{2} \|\cdot\|^2$  and  $r - \frac{\mu_r}{2} \|\cdot\|^2$  are convex.*
- (A.4) *There exists a constant  $\bar{g}_r > 0$  such that for all  $x \in \text{dom } r$  there exists  $\lambda \in \partial r(x)$  with  $\|\lambda\| \leq \bar{g}_r$ .*

Assumption (A.3) implies that the function  $\psi$  is strongly convex with convexity parameter  $\bar{\mu}$ . Furthermore, if both assumption (A.1) and (A.3) are satisfied, then the parameter  $\mu_f$  is



bounded by the Lipschitz constant  $L$ , i.e., we have  $|\mu_f| \leq L$ . The assumptions (A.3)–(A.4) are only required for a variant of Algorithm 1 that uses a modified globalization strategy, see Theorem 3.12. A concrete example for  $r$  that satisfies (A.4) is given in Remark 3.13. We continue with the assumptions on the parameters used within our algorithmic framework.

**ASSUMPTION 3.2.** *Let  $(\Lambda_k)_k \subset \mathbb{S}_{++}^n$  be a family of symmetric, positive definite parameter matrices and let  $(\nu_k)_k$ ,  $(\varepsilon_k^1)_k$ , and  $(\varepsilon_k^2)_k$  be given sequences. Then, for some given parameter  $p \in (0, 1)$  we assume:*

(B.1) *There exist  $0 < \lambda_m \leq \lambda_M < \infty$  such that  $\lambda_M I \succeq \Lambda_k \succeq \lambda_m I$  for all  $k \in \mathbb{N}$ .*

(B.2) *It holds  $(\nu_k)_k, (\varepsilon_k^2)_k \in \ell_+^1$ , and  $(\varepsilon_k^1)_k \in \ell_+^p$ .*

In the following sections, we study the convergence properties of the stochastic process  $(x^k)_k$  generated by Algorithm 1 with respect to the filtrations

$$\mathcal{F}_k := \sigma(s^0, \dots, s^k, t^0, \dots, t^k), \quad \text{and} \quad \hat{\mathcal{F}}_k := \sigma(s^0, \dots, s^k, s^{k+1}, t^0, \dots, t^k).$$

The filtration  $\mathcal{F}_k$  represents the information that is collected up to iteration  $k$  and that is used to compute the trial point  $z_n^k$  or a proximal gradient step

$$z_p^k := x^k + \alpha_k v^k = x^k - \alpha_k F_{s^k}^{\Lambda_k}(x^k).$$

The filtration  $\hat{\mathcal{F}}_k$  has a similar interpretation, but it also contains the information produced by deciding whether the Newton step  $z_n^k$  should be accepted or rejected, i.e., it holds  $\hat{\mathcal{F}}_k = \sigma(\mathcal{F}_k \cup \sigma(s^{k+1}))$ . The filtrations  $\{\mathcal{F}_k, \hat{\mathcal{F}}_k\}$  naturally describe the aggregation of information generated by Algorithm 1. We will work with the following stochastic conditions.

**ASSUMPTION 3.3.** *We assume:*

(C.1) *For all  $k \in \mathbb{N}_0$ , the generalized derivative  $M_k$ , chosen in step 3 of Algorithm 1, is an  $\mathcal{F}_k$ -measurable mapping, i.e., the function  $M_k : \Omega \rightarrow \mathbb{R}^{n \times n}$  is an  $\mathcal{F}_k$ -measurable selection of the multifunction  $\mathcal{M}_k : \Omega \rightrightarrows \mathbb{R}^{n \times n}$ ,  $\mathcal{M}_k(\omega) := \mathcal{M}_{s^k(\omega), t^k(\omega)}^{\Lambda_k}(x^k(\omega))$ .*

(C.2) *The variance of the individual stochastic gradients is bounded, i.e., for all  $k \in \mathbb{N}$  there exists  $\sigma_k \geq 0$  such that*

$$\mathbb{E}[\|\nabla f(x^k) - G_{s^k}(x^k)\|^2] \leq \sigma_k^2.$$

The second condition is common in stochastic programming, see, e.g., [29, 95, 13, 18, 31, 86]. Since the generalized derivative  $M_k$  is generated iteratively and depends on the random process  $(x^k)_k$  and on the mini-batches  $(s^k)_k, (t^k)_k$ , condition (C.1) is required to guarantee that the selected matrices  $M_k$  actually define  $\mathcal{F}_k$ -measurable random operators. A similar assumption was also used in [86]. Furthermore, applying the techniques and theoretical results presented in [83, 84] for infinite-dimensional nonsmooth operator equations, we can ensure that the multifunction  $\mathcal{M}_k$  admits at least one measurable selection  $M_k : \Omega \rightarrow \mathbb{R}^{n \times n}$ . We discuss this important observation together with a proof of Fact 3.4 in Appendix A.1. Let us note that it is also possible to generalize the assumptions and allow  $\mathcal{F}_k$ -measurable parameter matrices  $\Lambda_k$ . However, in order to simplify our analysis, we focus on a deterministic choice of  $(\Lambda_k)_k$  and do not consider this extension here.

As a consequence of condition (C.1) and of the assumptions on  $f$ , we can infer that the random processes  $(z_n^k)_k, (z_p^k)_k$ , and  $(x^k)_k$  are adapted to the filtrations  $\mathcal{F}_k$  and  $\hat{\mathcal{F}}_k$ .

**FACT 3.4.** *Under assumption (C.1), it holds  $z_n^k, z_p^k \in \mathcal{F}_k$  and  $x^{k+1} \in \hat{\mathcal{F}}_k$  for all  $k \in \mathbb{N}_0$ .*

Since the choice of the iterate  $x^{k+1}$  depends on the various criteria, the properties stated in Fact 3.4 are not immediately obvious. In particular, we need to verify that the decision of accepting or rejecting the stochastic semismooth Newton step  $z_n^k$  is an  $\hat{\mathcal{F}}_k$ -measurable action. A proof of Fact 3.4 is presented in Appendix A.1.

**3.2. Properties of  $F^\Lambda$ .** In this subsection, we discuss several useful properties of the nonsmooth function  $F^\Lambda$  and of its stochastic version  $F_s^\Lambda$ . The next statement shows that  $\|F_s^\Lambda(x)\|$  does not grow too much when the parameter matrix  $\Lambda$  changes. This result was first established by Tseng and Yun in [82].

LEMMA 3.5. *Let  $\Lambda_1, \Lambda_2 \in \mathbb{S}_{++}^n$  be two arbitrary matrices. Then, for all  $x \in \mathbb{R}^n$ , for all samples  $s$ , and for  $W := \Lambda_2^{-\frac{1}{2}} \Lambda_1 \Lambda_2^{-\frac{1}{2}}$ , it follows*

$$\|F_s^{\Lambda_1}(x)\| \leq \frac{1 + \lambda_{\max}(W) + \sqrt{1 - 2\lambda_{\min}(W) + \lambda_{\max}(W)^2}}{2} \frac{\lambda_{\max}(\Lambda_2)}{\lambda_{\min}(\Lambda_1)} \|F_s^{\Lambda_2}(x)\|.$$

*Proof.* The proof is identical to the proof of [82, Lemma 3] and will be omitted here.  $\square$

REMARK 3.6. *Let  $\Lambda \in \mathbb{S}_{++}^n$  be given and let  $(\Lambda_k)_k \subset \mathbb{S}_{++}^n$  be a family of symmetric, positive definite matrices satisfying assumption (B.1). Then, it easily follows*

$$\frac{\lambda_{\max}(\Lambda)}{\lambda_m} I \succeq \Lambda_k^{-\frac{1}{2}} \Lambda \Lambda_k^{-\frac{1}{2}} \succeq \frac{\lambda_{\min}(\Lambda)}{\lambda_M} I \quad \text{and} \quad \frac{\lambda_M}{\lambda_{\min}(\Lambda)} I \succeq \Lambda^{-\frac{1}{2}} \Lambda_k \Lambda^{-\frac{1}{2}} \succeq \frac{\lambda_m}{\lambda_{\max}(\Lambda)} I,$$

for all  $k \in \mathbb{N}$ , and, due to Lemma 3.5, we obtain the following bounds

$$(3.1) \quad \underline{\lambda} \cdot \|F_s^\Lambda(x)\| \leq \|F_s^{\Lambda_k}(x)\| \leq \bar{\lambda} \cdot \|F_s^\Lambda(x)\|, \quad \forall k \in \mathbb{N},$$

and for all mini-batches  $s$ ,  $x \in \mathbb{R}^n$ . The constants  $\underline{\lambda}, \bar{\lambda} > 0$  do not depend on  $k$ ,  $\Lambda_k$ , or  $s$ . Thus, the latter inequalities imply:

$$F^\Lambda(x^k) \rightarrow 0 \quad \Longleftrightarrow \quad F^{\Lambda_k}(x^k) \rightarrow 0, \quad k \rightarrow \infty.$$

As indicated in the last section, this can be used in the design of our algorithm. In particular, adaptive schemes or other techniques can be applied to update  $\Lambda$ .

The following result is a simple extension of [82, Theorem 4]; see also [90, Lemma 3.7] and [95] for comparison.

LEMMA 3.7. *Suppose that the assumptions (A.1), (A.3) are satisfied and let  $\Lambda \in \mathbb{S}_{++}^n$  be given with  $\lambda_M I \succeq \Lambda \succeq \lambda_m I$ . Furthermore, let  $x^*$  denote the unique solution of the problem  $\min_x \psi(x)$  and for any  $\tau > 0$  let us set*

$$b_1 := L - 2\lambda_m - \mu_r, \quad b_2 := \frac{(\lambda_M + \mu_r)^2}{\bar{\mu}}, \quad B_1(\tau) := \frac{1 + \tau}{\bar{\mu}} (\sqrt{b_1 + b_2 + \tau} + \sqrt{b_2})^2$$

Then, there exists some positive constant  $B_2(\tau)$  that only depends on  $\tau$  such that

$$(3.2) \quad \|x - x^*\|^2 \leq B_1(\tau) \cdot \|F_s^\Lambda(x)\|^2 + B_2(\tau) \cdot \|\nabla f(x) - G_s(x)\|^2,$$

for all  $x \in \mathbb{R}^n$  and for every sample  $s$ . If the full gradient is used, the term  $\|\nabla f(x) - G_s(x)\|$  vanishes for all  $x$  and (3.2) holds with  $B_1(\tau) \equiv B_1(0)$ .

*Proof.* The proof of Lemma 3.7 and an explicit derivation of the constant  $B_2(\tau)$  are presented in Appendix A.2.  $\square$

**3.3. Convergence Analysis.** In the following, we first verify that a stochastic proximal gradient step yields approximate  $\psi$ -descent whenever the step size  $\alpha_k$  in step 6 of Algorithm 1 is chosen sufficiently small. We also give a bound for the step sizes  $\alpha_k$ . Let us note that similar results were shown in [95, 31, 30] and that the proof of Lemma 3.8 mainly relies on the well-known descent lemma

$$(3.3) \quad f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n,$$

which is a direct consequence of assumption (A.1).

LEMMA 3.8. *Let  $x \in \text{dom } r$  and  $\Lambda \in \mathbb{S}_{++}^n$  be arbitrary and suppose that the conditions (A.1) and (B.1) (for  $\Lambda$ ) are satisfied. Moreover, let  $\gamma \in (0, 1)$ ,  $\rho \in (1, \gamma^{-1})$ , and the mini-batch  $s$  be given and set  $\bar{\alpha} := 2(1 - \gamma\rho)\lambda_m L^{-1}$ . Then, for all  $\alpha \in [0, \min\{1, \bar{\alpha}\}]$  it holds*

$$(3.4) \quad \psi(x + \alpha v) - \psi(x) \leq -\alpha\gamma\|v\|_\Lambda^2 + \frac{\alpha}{4\gamma(\rho - 1)\lambda_m} \|\nabla f(x) - G_s(x)\|^2,$$

where  $v := -F_s^\Lambda(x)$ .

*Proof.* We first define  $\Delta := \langle G_s(x), v \rangle + r(x + v) - r(x)$ . Then, applying the optimality condition of the proximity operator (2.4), it follows

$$\Delta \leq \langle G_s(x), v \rangle + \langle -\Lambda v - G_s(x), v \rangle = -\|v\|_\Lambda^2.$$

Using the descent lemma (3.3), the convexity of  $r$ , and Young's inequality, we now obtain

$$\begin{aligned} \psi(x + \alpha v) - \psi(x) + \alpha\gamma\|v\|_\Lambda^2 &\leq \alpha(\langle \nabla f(x), v \rangle + r(x + v) - r(x)) + \frac{L\alpha^2}{2} \|v\|^2 + \alpha\gamma\|v\|_\Lambda^2 \\ &\leq \frac{L\alpha^2}{2} \|v\|^2 - \alpha(1 - \gamma)\|v\|_\Lambda^2 + \alpha\langle \nabla f(x) - G_s(x), v \rangle \\ &\leq \alpha \left( \frac{1}{2} L\alpha - (1 - \gamma\rho)\lambda_m \right) \|v\|^2 + \frac{\alpha}{4\gamma(\rho - 1)\lambda_m} \|\nabla f(x) - G_s(x)\|^2. \end{aligned}$$

Since the first term is nonpositive for all  $\alpha \leq \bar{\alpha}$ , this establishes (3.4).  $\square$

In the special case  $\lambda_m \geq L$  and  $\rho \leq (2\gamma)^{-1}$ , Lemma 3.8 implies that the approximate descent condition (3.4) holds for all  $\alpha \in [0, 1]$ . The next lemma is one of our key tools to analyze the stochastic behavior of the Newton iterates and to bound the associated residual terms  $\|F_{s_{k+1}}^{\Lambda_{k+1}}(z_n^k)\|$ .

LEMMA 3.9. *Let  $(Y_k)_k$  be an arbitrary binary sequence in  $\{0, 1\}$  and let  $a_0 \geq 0$ ,  $\eta \in (0, 1)$ ,  $p \in (0, 1]$ , and  $(\nu_k)_k \in \ell_+^1$ ,  $(\varepsilon_k)_k \in \ell_+^p$  be given. Let the sequence  $(a_k)_k$  be defined by*

$$a_{k+1} := (\eta + \nu_k)^{Y_k} a_k + Y_k \varepsilon_k, \quad \forall k \in \mathbb{N}_0.$$

*Then, for all  $R \geq 1$ ,  $k \geq 0$ , and all  $q \in [p, 1]$ , it holds*

$$a_{k+1} \leq C_\nu \left[ a_0 + \sum_{k=0}^{\infty} \varepsilon_k \right] \quad \text{and} \quad \sum_{k=0}^{R-1} Y_k a_{k+1}^q \leq \frac{C_\nu^q}{1 - \eta^q} \left[ (\eta a_0)^q + \sum_{k=0}^{\infty} \varepsilon_k^q \right],$$

where  $C_\nu := \exp(\eta^{-1} \sum_{i=0}^{\infty} \nu_i)$ .

*Proof.* Using an induction, we can derive an explicit representation for  $a_{k+1}$

$$(3.5) \quad a_{k+1} = \left\{ \prod_{i=0}^k (\eta + \nu_i)^{Y_i} \right\} a_0 + \sum_{j=0}^{k-1} \left\{ \prod_{i=j+1}^k (\eta + \nu_i)^{Y_i} \right\} Y_j \varepsilon_j + Y_k \varepsilon_k, \quad \forall k \geq 0.$$

Next, using  $Y_i \in \{0, 1\}$ ,  $i \in \mathbb{N}$ , and  $\log(1 + \nu_i \eta^{-1}) \leq \nu_i \eta^{-1}$ , we obtain the estimate

$$(3.6) \quad \prod_{i=\ell}^k (\eta + \nu_i)^{Y_i} \leq \left\{ \prod_{i=\ell}^k \eta^{Y_i} \right\} \cdot \exp \left( \eta^{-1} \sum_{i=\ell}^k Y_i \nu_i \right) \leq C_\nu \cdot \eta^{\sum_{i=\ell}^k Y_i}, \quad \ell \geq 0.$$

The bound on  $a_{k+1}$  now follows from (3.5), (3.6), and  $\eta \leq 1$ . Let us now define the set  $K_\ell := \{i \in \{\ell, \dots, R-1\} : Y_i = 1\}$ . Then, it holds

$$\sum_{k=\ell}^{R-1} Y_k \eta^{\sum_{i=\ell}^k q Y_i} = \sum_{k \in K_\ell} \eta^{\sum_{i \in K_\ell, i \leq k} q} = \sum_{j=1}^{|K_\ell|} \eta^{jq} \leq \sum_{k=\ell}^{R-1} \eta^{(k-\ell+1)q}, \quad \ell \in \{0, \dots, R-1\}.$$

Combining the last results and using the subadditivity of  $x \mapsto x^q$ ,  $q \in [p, 1]$ , we have

$$\begin{aligned} \sum_{k=0}^{R-1} Y_k a_{k+1}^q &\leq C_\nu^q \sum_{k=0}^{R-1} Y_k \eta^{\sum_{i=0}^k q Y_i} a_0^q + C_\nu^q \sum_{k=0}^{R-1} \sum_{j=0}^{k-1} \eta^{\sum_{i=j+1}^k q Y_i} Y_k Y_j \varepsilon_j^q + \sum_{k=0}^{R-1} Y_k \varepsilon_k^q \\ &\leq C_\nu^q \sum_{k=0}^{R-1} \eta^{q(k+1)} a_0^q + C_\nu^q \sum_{j=0}^{R-1} \left\{ \sum_{k=j}^{R-1} \eta^{\sum_{i=j}^k q Y_i} Y_k \right\} \eta^{-q Y_j} Y_j \varepsilon_j^q \\ &\leq \frac{C_\nu^q}{1 - \eta^q} \cdot (\eta a_0)^q + C_\nu^q \sum_{j=0}^{R-1} \left\{ \sum_{k=j}^{R-1} \eta^{q(k-j)} \right\} \varepsilon_j^q \leq \frac{C_\nu^q}{1 - \eta^q} \left[ (\eta a_0)^q + \sum_{j=0}^{\infty} \varepsilon_j^q \right] \end{aligned}$$

as desired. Let us also note that the inclusion  $\ell_+^q \subset \ell_+^p$  is used in the last step.  $\square$

We are now in the position to establish global convergence of Algorithm 1 in the sense that the expectation  $\mathbb{E}[\|F^\Lambda(x^k)\|^2]$  converges to zero as  $k \rightarrow \infty$ . We first show convergence of Algorithm 1 under the conditions (C.1)–(C.2) and under the additional assumptions that the step sizes are diminishing and that the scaled stochastic error terms  $\alpha_k \sigma_k^2$ ,  $k \in \mathbb{N}$ , are summable which is a common requirement in the analysis of stochastic methods for nonsmooth, nonconvex optimization, see, e.g., [95, 31, 30].

Our basic idea is to show that both the proximal gradient and the semismooth Newton step yield approximate  $\psi$ -descent and that the error induced by gradient and Hessian sampling can be controlled in expectation. For a proximal gradient step this basically follows from Lemma 3.8. For a Newton step, we combine the growth conditions (2.8)–(2.9) and Lemma 3.9 to establish an estimate similar to (3.4). An analogous strategy was also used in [51, 52]. In our situation, however, a more careful discussion of the possible effects of the semismooth Newton steps is needed to cope with the stochastic situation. More specifically, since our convergence result is stated in expectation, all possible realizations of the random mini-batches  $s^k$  and  $t^k$ ,  $k \in \mathbb{N}_0$ , and their influence on the conditions (2.8)–(2.9) have to be considered. In order to apply Lemma 3.9, we now set up some preparatory definitions.

Let  $k \in \mathbb{N}_0$  be given and let us define  $Q_k : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $Q_k(a) := \|a\| - (\eta + \nu_k)\theta_k$  and  $P_k : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $P_k(a, b) := \psi(a) - \beta \theta_k^{1-p} \|b\|^p - \psi(x^k)$ , and

$$S_k := [\text{dom } r \times \mathbb{R}^n] \cap [\mathbb{R}^n \times \text{lev}_{\varepsilon_k^1} Q_k] \cap \text{lev}_{\varepsilon_k^2} P_k.$$

Then, setting  $Y_{k+1} := \mathbb{1}_{S_k}(z_n^k, F_{s_{k+1}}^{\Lambda_{k+1}}(z_n^k))$ , it holds

$$Y_{k+1} = \begin{cases} 1 & \text{if } z_n^k = x^k + d^k \text{ is feasible and satisfies the conditions (2.8) and (2.9),} \\ 0 & \text{otherwise} \end{cases}$$

and consequently, each iterate  $x^{k+1}$  can be calculated as follows

$$\begin{aligned} (3.7) \quad x^{k+1} &= (1 - Y_{k+1})z_p^k + Y_{k+1}z_n^k \\ &= (1 - Y_{k+1})[x^k - \alpha_k F_{s^k}^{\Lambda_k}(x^k)] + Y_{k+1}[x^k - M_k^+ F_{s^k}^{\Lambda_k}(x^k)]. \end{aligned}$$

Here, the matrix  $M_k^+$  denotes the *Moore-Penrose inverse* of the generalized derivative  $M_k$ . Let us note that this compact representation of our iterative scheme turns out to be particularly useful in the proof of [Fact 3.4](#), see [Appendix A.1](#). We also introduce the parameters  $Z_k$ ,  $k \in \mathbb{N}_0$ , which are defined recursively via

$$Z_0 := \theta_0 \in \mathbb{R}_+, \quad Z_{k+1} := (\eta + \nu_k)^{Y_{k+1}} Z_k + Y_{k+1} \varepsilon_k^1.$$

By construction of [Algorithm 1](#) and by induction, we have  $\theta_k \leq Z_k$  and thus,

$$(3.8) \quad Y_{k+1} \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\| \leq Y_{k+1} Z_{k+1}, \quad \forall k \in \mathbb{N}_0.$$

Moreover, by identifying  $Y_k \equiv Y_{k+1}$  and  $a_k \equiv Z_k$ , [Lemma 3.9](#) yields the following sample-independent and uniform bounds

$$(3.9) \quad \sum_{k=0}^{R-1} Y_{k+1} Z_{k+1}^q \leq \frac{C_\nu^q}{1 - \eta^q} \left[ (\eta \theta_0)^q + \sum_{k=0}^{\infty} (\varepsilon_k^1)^q \right] =: C_z(q) < \infty$$

and  $Z_{k+1} \leq C_\nu [\theta_0 + \sum_{k=0}^{\infty} \varepsilon_k^1] =: C_z$  for all  $k \geq 0$ ,  $R \in \mathbb{N}$ , and  $q \in [p, 1]$ . We now state our main result of this section.

**THEOREM 3.10.** *Let the sequence  $(x^k)_k$  be generated by [Algorithm 1](#). Suppose that the assumptions (A.1)–(A.2), (B.1)–(B.2), and (C.1)–(C.2) are satisfied. Furthermore, suppose that the step sizes  $\alpha_k \in [0, 1]$ ,  $k \in \mathbb{N}$ , are chosen such that the approximate descent condition (3.4) holds for some given  $\gamma$  and  $\rho$ . Then, under the additional assumptions*

$$(\alpha_k)_k \text{ is monotonically decreasing,} \quad \sum \alpha_k = \infty, \quad \sum \alpha_k \sigma_k^2 < \infty,$$

*it holds  $\liminf_{k \rightarrow \infty} \mathbb{E}[\|F^\Lambda(x^k)\|^2] = 0$  and  $\liminf_{k \rightarrow \infty} F^\Lambda(x^k) = 0$  a.s. for any  $\Lambda \in \mathbb{S}_{++}^n$ .*

*Proof.* Assumption (A.1) implies that the gradient mapping  $\nabla f(x)$  is Lipschitz continuous on  $\mathbb{R}^n$  with Lipschitz constant  $L$ . Thus, for any matrix  $\Gamma \in \mathbb{S}_{++}^n$  with  $\lambda_M I \succeq \Gamma \succeq \lambda_m I$ , we obtain the Lipschitz constant  $1 + L\lambda_m^{-1}$  for  $u^\Gamma(x)$ . Since the proximity operator  $\text{prox}_r^\Gamma$  is  $\Gamma$ -nonexpansive, we now have

$$\begin{aligned} \|F^\Gamma(x) - F^\Gamma(y)\| &\leq \|x - y\| + \lambda_m^{-\frac{1}{2}} \|\text{prox}_r^\Gamma(u^\Gamma(x)) - \text{prox}_r^\Gamma(u^\Gamma(y))\|_\Gamma \\ &\leq \|x - y\| + (\lambda_m^{-1} \lambda_M)^{\frac{1}{2}} \|u^\Gamma(x) - u^\Gamma(y)\| \\ &\leq (1 + (\lambda_m^{-1} \lambda_M)^{\frac{1}{2}} + L(\lambda_m^{-3} \lambda_M)^{\frac{1}{2}}) \|x - y\| \end{aligned}$$

for all  $x, y \in \mathbb{R}^n$ . Hence, by assumption (B.1), the functions  $x \mapsto F^{\Lambda_k}(x)$ ,  $k \in \mathbb{N}$ , are all Lipschitz continuous on  $\mathbb{R}^n$  with modulus  $L_F := 1 + (\lambda_m^{-1} \lambda_M)^{\frac{1}{2}} + L(\lambda_m^{-3} \lambda_M)^{\frac{1}{2}}$ .

We first consider the case where  $x^{k+1} = z_p^k = x^k + \alpha_k v^k$  is generated by the proximal gradient method in step 6. Then, due to (B.1) and [Remark 3.6](#), there exists a constant  $\underline{\lambda} = \underline{\lambda}(\lambda_m, \lambda_M)$  such that

$$\begin{aligned} \|F^{\Lambda_k}(x^{k+1})\| &\leq \underline{\lambda}^{-1} (L_F \alpha_k \|v^k\| + \lambda_m^{-\frac{1}{2}} \|F^{\Lambda_k}(x^k) - F_{s^k}^{\Lambda_k}(x^k)\|_{\Lambda_k} + \lambda_m^{-\frac{1}{2}} \|F_{s^k}^{\Lambda_k}(x^k)\|_{\Lambda_k}) \\ &\leq \underline{\lambda}^{-1} \lambda_m^{-\frac{1}{2}} (L_F + 1) \|F_{s^k}^{\Lambda_k}(x^k)\|_{\Lambda_k} + (\underline{\lambda} \lambda_m)^{-1} \|\nabla f(x^k) - G_{s^k}(x^k)\|. \end{aligned}$$

Here, we again used the  $\Lambda_k$ -nonexpansiveness of the proximity operator  $\text{prox}_r^{\Lambda_k}$ . Thus, applying [Lemma 3.8](#), using the estimate  $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ , for  $a, b \in \mathbb{R}^n$ , and setting

$\mathcal{E}_k := \|\nabla f(x_k) - G_{s^k}(x^k)\|$ , we obtain

$$\begin{aligned} & \psi(x^k) - \psi(x^{k+1}) \\ & \geq \underbrace{\frac{\gamma \underline{\lambda}^2 \lambda_m}{2(L_F + 1)^2}}_{=: c_1} \cdot \alpha_k \|F^\Lambda(x^{k+1})\|^2 - \underbrace{\frac{1}{\lambda_m} \left( \frac{\gamma}{(L_F + 1)^2} + \frac{1}{4\gamma(\rho - 1)} \right)}_{=: c_2} \cdot \alpha_k \mathcal{E}_k^2. \end{aligned}$$

Next, we derive a similar estimate for a Newton step  $x^{k+1} = z_n^k = x^k + d^k$ . As before and due to assumption (B.1) and [Remark 3.6](#), we have

$$(3.10) \quad \|F^\Lambda(x^{k+1})\|^2 \leq 2\underline{\lambda}^{-2} \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\|^2 + 2(\underline{\lambda} \lambda_m)^{-2} \|\nabla f(x^{k+1}) - G_{s^{k+1}}(x^{k+1})\|^2.$$

Combining the growth condition [\(2.9\)](#), [\(3.10\)](#), and the bound  $\alpha_{k+1} \leq 1$ , it holds

$$\begin{aligned} \psi(x^k) - \psi(x^{k+1}) & \geq -\beta \theta_k^{1-p} \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\|^p - \varepsilon_k^2 \\ & \geq c_1 \cdot \alpha_{k+1} \|F^\Lambda(x^{k+1})\|^2 - \varepsilon_k^2 - 2c_1(\underline{\lambda} \lambda_m)^{-2} \cdot \alpha_{k+1} \mathcal{E}_{k+1}^2 \\ & \quad - \underbrace{\left( 2c_1 \underline{\lambda}^{-2} \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\|^{2-p} + \beta \theta_k^{1-p} \right)}_{=: \mathbb{T}_k} \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\|^p. \end{aligned}$$

Furthermore, using  $\|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\| = \theta_{k+1} \leq C_z$  and  $\theta_k \leq C_z$ , it can be easily shown that the term  $\mathbb{T}_k$  is bounded by a constant  $\bar{\mathbb{T}}$  that does not depend on any of the random mini-batches  $s^j, t^j, j \in \mathbb{N}_0$ .

Now, let  $R \in \mathbb{N}$  be arbitrary. Then, the monotonicity of  $(\alpha_k)_k$ , [\(3.8\)–\(3.9\)](#), and our last results imply

$$\begin{aligned} & \psi(x^0) - \psi(x^{R+1}) \\ & \geq \sum_{k=0}^R c_1 \min\{\alpha_k, \alpha_{k+1}\} \|F^\Lambda(x^{k+1})\|^2 - \sum_{k=0}^R (1 - \Upsilon_{k+1}) c_2 \cdot \alpha_k \mathcal{E}_k^2 \\ & \quad - \sum_{k=0}^R \Upsilon_{k+1} \left[ 2c_1(\underline{\lambda} \lambda_m)^{-2} \cdot \alpha_{k+1} \mathcal{E}_{k+1}^2 + \varepsilon_k^2 + \bar{\mathbb{T}} \cdot \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\|^p \right] \\ & \geq \sum_{k=0}^R [c_1 \alpha_{k+1} \|F^\Lambda(x^{k+1})\|^2 - c_2 \alpha_k \mathcal{E}_k^2 - 2c_1(\underline{\lambda} \lambda_m)^{-2} \alpha_{k+1} \mathcal{E}_{k+1}^2] - \bar{\mathbb{T}} C_z(p) - \sum_{k=0}^R \varepsilon_k^2. \end{aligned}$$

Thus, taking expectation and setting  $c_3 := c_2 + 2c_1(\underline{\lambda} \lambda_m)^{-2}$ , we obtain

$$\sum_{k=0}^R c_1 \alpha_{k+1} \mathbb{E}[\|F^\Lambda(x^{k+1})\|^2] \leq \psi(x^0) - \mathbb{E}[\psi(x^{R+1})] + \bar{\mathbb{T}} C_z(p) + \sum_{k=0}^R \varepsilon_k^2 + c_3 \sum_{k=0}^{R+1} \alpha_k \sigma_k^2.$$

Since the objective function  $\psi$  is bounded from below and the sequences  $(\alpha_k \sigma_k^2)_k$  and  $(\varepsilon_k^2)_k$  are summable, this obviously yields  $\sum \alpha_k \mathbb{E}[\|F^\Lambda(x^k)\|^2] < \infty$ . Consequently, our first claim follows from the assumption  $\sum \alpha_k = \infty$ . On the other hand, Fatou's lemma implies

$$\mathbb{E} \left[ \sum_{k=0}^{\infty} \alpha_k \|F^\Lambda(x^k)\|^2 \right] \leq \liminf_{R \rightarrow \infty} \mathbb{E} \left[ \sum_{k=0}^R \alpha_k \|F^\Lambda(x^k)\|^2 \right] < \infty$$

and hence, we have  $\sum \alpha_k \|F^\Lambda(x^k)\|^2 < \infty$  with probability 1. As before we can now infer  $\liminf_{k \rightarrow \infty} F^\Lambda(x^k) = 0$  with probability 1 which completes our proof.  $\square$



REMARK 3.11. In the case  $\sum \sigma_k^2 < \infty$  and if the step sizes  $\alpha_k$  are fixed or bounded, our results in [Theorem 3.10](#) can be strengthened to  $\lim_{k \rightarrow \infty} \mathbb{E}[\|F^\Lambda(x^k)\|^2] = 0$  and we have  $\lim_{k \rightarrow \infty} F^\Lambda(x^k) = 0$  almost surely. Let us now assume that the samples  $s_i^k$ ,  $i \in [\mathbf{n}_k^g]$ ,  $k \in \mathbb{N}$ , are chosen independently of each other and that the conditions

$$(3.11) \quad \mathbb{E}[\mathcal{G}(x, s_i^k)] = \nabla f(x), \quad \mathbb{E}[\|\nabla f(x) - \mathcal{G}(x, s_i^k)\|^2] \leq \bar{\sigma}^2,$$

hold uniformly for all  $i \in [\mathbf{n}_k^g]$ ,  $k \in \mathbb{N}_0$ , and  $x \in \mathbb{R}^n$  and for some  $\bar{\sigma} > 0$ . Then, as shown in [\[31, 39\]](#), it follows  $\mathbb{E}[\|\nabla f(x) - G_{s^k}(x)\|^2] \leq \bar{\sigma}^2[\mathbf{n}_k^g]^{-1}$  for all  $x \in \mathbb{R}^n$  and consequently, due to  $Y_k \in \{0, 1\}$  and  $z_n^{k-1}, z_p^{k-1} \in \mathcal{F}_{k-1}$ , we have

$$\begin{aligned} \mathbb{E}[\|\nabla f(x^k) - G_{s^k}(x^k)\|^2] &\leq 2\mathbb{E}[(1 - Y_k)\|\nabla f(z_p^{k-1}) - G_{s^k}(z_p^{k-1})\|^2] \\ &\quad + 2\mathbb{E}[Y_k\|\nabla f(z_n^{k-1}) - G_{s^k}(z_n^{k-1})\|^2] \leq 4\bar{\sigma}^2[\mathbf{n}_k^g]^{-1}. \end{aligned}$$

Hence, one way to guarantee summability of the error terms  $\sigma_k^2$  is to asymptotically increase the sample size  $\mathbf{n}_k^g$  and set  $\mathbf{n}_k^g = \mathcal{O}(k^{1+\varpi})$  for some  $\varpi > 0$ . This observation is similar to the results in [\[95, 31\]](#). We will discuss the conditions (3.11) in more detail in the next section.

In the following, we present a situation where the approximate  $\psi$ -descent condition (2.9) is not needed in order to guarantee global convergence of the method. In applications, this can be quite important, since calculating the full objective function  $\psi$  may be similarly expensive as evaluating the full gradient  $\nabla f$ . The following variant of [Theorem 3.10](#) is mainly based on the strong convexity assumption (A.3) and on the boundedness assumption (A.4).

THEOREM 3.12. Let the sequence  $(x^k)_k$  be generated by Algorithm 1 without checking the growth condition (2.9). Suppose that the assumptions (A.1), (A.3)–(A.4), (B.1)–(B.2), and (C.1)–(C.2) are satisfied. Furthermore, suppose that the step sizes  $(\alpha_k)_k$  are chosen via  $\alpha_k \in [\underline{\alpha}, \min\{1, \bar{\alpha}\}]$  for some  $\underline{\alpha} > 0$  and all  $k \in \mathbb{N}$ . Then, under the additional assumption

$$\sum \sigma_k < \infty,$$

it holds  $\lim_{k \rightarrow \infty} \mathbb{E}[\|F^\Lambda(x^k)\|^2] = 0$  and  $\lim_{k \rightarrow \infty} F^\Lambda(x^k) = 0$  a.s. for any  $\Lambda \in \mathbb{S}_{++}^n$ .

*Proof.* As in the proof of [Theorem 3.10](#), we want to derive suitable lower bounds for the  $\psi$ -descent  $\psi(x^k) - \psi(x^{k+1})$ . We first consider the case where  $x^{k+1} = z_p^k$  is generated by the proximal gradient method. Then, as shown in the proof of [Theorem 3.10](#) and using the bound on  $\alpha_k$ , we have

$$\psi(x^k) - \psi(x^{k+1}) \geq c_1 \underline{\alpha} \|F^\Lambda(x^{k+1})\|^2 - c_2 \cdot \mathcal{E}_k^2.$$

Next, we discuss the second case  $x^{k+1} = z_n^k$ . By [Lemma 3.7](#) and reusing the estimate  $\|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\| \leq C_z$  (see again (3.8)), it holds

$$\|z_n^k - x^*\|^2 \leq B_1(\tau) \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\|^2 + B_2(\tau) \mathcal{E}_{k+1}^2 \leq B_1(\tau) C_z \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\| + B_2(\tau) \mathcal{E}_{k+1}^2$$

for some  $\tau > 0$ . By assumption (A.3), the functions  $f - \frac{\mu_f}{2} \|\cdot\|^2$  and  $r - \frac{\mu_r}{2} \|\cdot\|^2$  are convex (and directionally differentiable) and hence, we have

$$(3.12) \quad \psi(y) - \psi(x) \geq r'(x; y - x) + \langle \nabla f(x), y - x \rangle + \frac{\bar{\mu}}{2} \|y - x\|^2, \quad \forall x, y \in \text{dom } r.$$

Now, applying the optimality of  $x^*$ , (3.12) with  $x \equiv z_n^k$  and  $y \equiv x^*$ ,  $z_n^k \in \text{dom } r$ , the Lipschitz continuity of  $\nabla f$  and  $F^\Lambda$ , the subadditivity of the square root, and defining

$$d_1 := \sqrt{B_1(\tau)(\bar{g}_r + \|\nabla f(x^*)\|)} + B_1(\tau)C_zL, \quad d_2 := \sqrt{B_2(\tau)(\bar{g}_r + \|\nabla f(x^*)\|)},$$

we obtain

$$\begin{aligned}
\psi(x^k) - \psi(x^{k+1}) &= \psi(x^k) - \psi(x^*) + \psi(x^*) - \psi(x^{k+1}) \\
&\geq r'(z_n^k; x^* - z_n^k) + \langle \nabla f(z_n^k), x^* - z_n^k \rangle + \frac{\bar{\mu}}{2} \|z_n^k - x^*\|^2 \\
&\geq -(\bar{g}_r + \|\nabla f(x^*)\|) \cdot \|z_n^k - x^*\| + \left(\frac{\bar{\mu}}{2} - L\right) \|z_n^k - x^*\|^2 \\
&\geq -d_1 \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\| - d_2 \mathcal{E}_{k+1} - B_2(\tau) L \mathcal{E}_{k+1}^2 + \frac{\bar{\mu} \underline{\lambda}^2}{2L_F^2} \|F^\Lambda(x^{k+1})\|^2.
\end{aligned}$$

Combining the last inequalities, setting  $d_3 := \min\{c_1 \underline{\lambda}, (2L_F^2)^{-1} \bar{\mu} \underline{\lambda}^2\}$  and using again (3.9) with  $q = 1$ , it holds

$$\psi(x^0) - \psi(x^{R+1}) \geq \sum_{k=1}^{R+1} [d_3 \|F^\Lambda(x^k)\|^2 - d_2 \mathcal{E}_k] - (c_2 + B_2(\tau) L) \sum_{k=0}^{R+1} \mathcal{E}_k^2 - d_1 C_z(1),$$

for all  $R \in \mathbb{N}$ . Taking expectation, our first claim now follows from (C.2), Jensen's inequality,  $\sum \sigma_k < \infty$ , and from the lower boundedness of  $\psi(x^{R+1})$ . The probabilistic convergence of the sequence  $(F^\Lambda(x^k))_k$  can then be inferred as in the proof of [Theorem 3.10](#).  $\square$

**REMARK 3.13.** *Let us note that assumption (A.4) is required to derive a suitable lower bound for the difference terms  $\psi(x^k) - \psi(z_n^k)$  which allows us to apply [Lemma 3.9](#). Furthermore, condition (A.4) is always satisfied in the following situation. Suppose that the mapping  $r$  has the special form  $r = \iota_{\mathcal{C}} + \varphi$ , where  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$  is a real-valued, convex function and  $\iota_{\mathcal{C}} : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is the indicator function of a nonempty, convex, and closed set  $\mathcal{C} \subset \mathbb{R}^n$ . Then, assumption (A.4) holds if the set  $\mathcal{C}$  is either compact or if  $\varphi$  is positively homogeneous. In particular, condition (A.4) is satisfied if  $r$  is a norm.*

*Proof.* For every feasible point  $x \in \text{dom } r = \mathcal{C}$ , we have  $0 \in \partial \iota_{\mathcal{C}}(x)$  and thus,  $\partial \varphi(x) \subset \partial r(x)$ . By [9, Proposition 16.17], the set  $\bigcup_{x \in \mathcal{C}} \partial \varphi(x)$  is bounded if  $\mathcal{C}$  is bounded. On the other hand, if  $\varphi$  is positively homogeneous, then it follows  $\partial \varphi(x) = \{\lambda \in \partial \varphi(0) : \langle \lambda, x \rangle = \varphi(x)\} \subset \partial \varphi(0)$ , see, e.g., [9, Proposition 16.18] and [51, Example 2.5.17]. Since  $\partial \varphi(0)$  is again a compact set, this proves our claim.  $\square$

**REMARK 3.14.** *The result in [Theorem 3.12](#) can be further improved by additionally damping the semismooth Newton step and setting  $z_n^k := x^k + \alpha_k d^k$ . Then, due to the convexity of  $\psi$ , we have  $\psi(x^k) - \psi(z_n^k) \geq \alpha_k (\psi(x^k) - \psi(x^k + d^k))$  and we can use the weaker conditions*

$$(\alpha_k)_k \text{ is monotonically decreasing, } \sum \alpha_k = \infty, \quad \sum \alpha_k \sigma_k < \infty$$

*to guarantee  $\liminf_{k \rightarrow \infty} \mathbb{E}[\|F^\Lambda(x^k)\|^2] = 0$ . Similar to [27, 95, 13] it is also possible to derive global convergence rates in terms of the expected distance to optimality  $\mathbb{E}[\psi(x^k) - \psi(x^*)]$ . However, in our case these rates will depend on the occurrence and total number of accepted Newton steps which are of stochastic nature in general.*

Finally, let us emphasize that our global results do not explicitly depend on the sampling strategy or on any (uniform) invertibility properties of the stochastic second order oracle  $H_t$  or of the chosen generalized derivatives  $M_k$ . Moreover, our results still hold if a different type of direction  $d^k$  is used instead of the semismooth Newton direction  $d^k = -M_k^+ F_{s^k}^{\Lambda_k}(x^k)$ . (In our proofs, we only require  $\mathcal{F}_k$ -measurability of  $d^k$ ).

**4. Local convergence.** In this part of the paper, we analyze the local convergence properties of our proposed method in detail. We will focus on a probabilistic setting, i.e., we consider a single trajectory of the stochastic process  $(x^k)_k$  and show that transition to fast local convergence and a fast rate of convergence can be achieved with high probability if the sample sizes  $\mathbf{n}_k^g$  and  $\mathbf{n}_k^h$  are chosen appropriately. With a slight abuse of notation, we will use  $(x^k)_k$  to denote either the underlying stochastic process or a corresponding trajectory generated by a single run of Algorithm 1 which should be clear from the context.

Our analysis heavily relies on different second order properties of the proximity operator  $\text{prox}_r^\Lambda$  and on concentration inequalities for vector- and matrix-valued martingales. In particular, these inequalities will allow us to quantify and control the errors induced by the stochastic oracles and by approximating the gradient and Hessian of  $f$ . A similar strategy was also used in [71, 72, 92, 97] for the analysis of pure, sub-sampled Newton methods for smooth optimization problems. In the next subsection, we present our local assumptions and the mentioned concentration results.

**4.1. Assumptions and Conditional Concentration Inequalities.** We will mainly work with the following set of local assumptions.

ASSUMPTION 4.1. *Let the trajectory  $(x^k)_k$  and the sequence  $(\Lambda_k)_k$  be generated by Algorithm 1 and suppose that  $x^*$  and  $\Lambda_*$  are accumulation points of  $(x^k)_k$  and  $(\Lambda_k)_k$ , respectively. We assume that the following conditions are satisfied.*

- (D.1) *There exists  $\bar{k} \in \mathbb{N}$  such that  $\Lambda_k = \Lambda_*$  for all  $k \geq \bar{k}$ .*
- (D.2) *The function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is twice continuously differentiable on  $\mathbb{R}^n$ .*
- (D.3) *The proximity operator  $\text{prox}_r^{\Lambda_*}$  is semismooth at  $u^{\Lambda_*}(x^*)$ .*
- (D.4) *The function  $\psi$  is Lipschitz continuous in a neighborhood of  $x^*$  with constant  $L_\psi$ .*
- (D.5) *There exists  $C > 0$  such that every generalized derivative  $M \in \mathcal{M}^{\Lambda_*}(x^*)$  is non-singular with  $\|M^{-1}\| \leq C$ .*

If, in addition,  $x^*$  is a stationary point of (1.1), then we assume:

- (D.6) *The accumulation point  $x^*$  is a local minimum of the problem (1.1).*

Let us briefly discuss the conditions in Assumption 4.1. Assumption (D.5) can be interpreted as a *BD- or CD-regularity condition* which is a common condition in the local analysis of nonsmooth optimization methods, see, e.g., [65, 64, 59]. Let us also mention that, by [9, Corollary 8.30], the Lipschitz condition (D.4) is equivalent to  $x^* \in \text{int dom } r$ . Hence, in a suitable neighborhood of  $x^*$ , any point  $x$  will be feasible with  $x \in \text{dom } r$ . See Remark 4.7 for further comments. Finally, as shown in [51, section 5.4], the assumptions (D.5) and (D.6) are both satisfied if  $x^*$  is a stationary point and  $\nabla^2 f(x^*)$  is positive definite.

In the following, we introduce three additional conditions that are connected to the variance of the error terms  $\mathcal{E}_k^g$  and  $\mathcal{E}_k^h$ ,

$$\mathcal{E}_k^g(x) := \|G_{s^k}(x) - \nabla f(x)\|, \quad \mathcal{E}_k^h(x) := \|H_{t^k}(x) - \nabla^2 f(x)\|,$$

and that extend assumption (C.2).

ASSUMPTION 4.2. *We consider the conditions:*

- (E.1) *The random mappings  $s_i^k, t_j^k$  are mutually independent to each other for all  $i \in [\mathbf{n}_k^g]$ ,  $j \in [\mathbf{n}_k^h]$ , and  $k \in \mathbb{N}_0$ . Furthermore, the stochastic oracles  $\mathcal{SFO}$  and  $\mathcal{SSO}$  generate unbiased estimators of the gradient and Hessian of  $f$ , i.e., for all  $x \in \mathbb{R}^n$  it holds*

$$\mathbb{E}[\mathcal{G}(x, s_i^k)] = \nabla f(x), \quad \mathbb{E}[\mathcal{H}(x, t_j^k)] = \nabla^2 f(x), \quad \forall i \in [\mathbf{n}_k^g], \forall j \in [\mathbf{n}_k^h], \forall k \in \mathbb{N}.$$

- (E.2) *We have  $\mathbb{E}[(\mathcal{E}_k^g(z_n^{k-1}))^2] < \infty$ ,  $\mathbb{E}[(\mathcal{E}_k^h(z_p^{k-1}))^2] < \infty$ ,  $\mathbb{E}[(\mathcal{E}_k^h(x^k))^2] < \infty$  for all  $k \in \mathbb{N}$  and there exists  $\bar{\sigma}, \bar{\rho} \geq 0$  such that for all  $k \in \mathbb{N}_0$ ,  $i \in [\mathbf{n}_k^g]$ ,  $j \in [\mathbf{n}_k^h]$  it holds*

$$(4.1) \quad \mathbb{E}[\|\mathcal{G}(x, s_i^k) - \nabla f(x)\|^2] \leq \bar{\sigma}^2, \quad \mathbb{E}[\|\mathcal{H}(x, t_j^k) - \nabla^2 f(x)\|^2] \leq \bar{\rho}^2, \quad \forall x \in \mathbb{R}^n.$$

(E.3) *There exists  $\bar{\sigma}, \bar{\rho} > 0$  such that  $\mathbb{E}[\exp([\mathcal{E}_k^g(z_n^{k-1})]^2/\bar{\sigma}^2)] < \infty$ ,*

$$\mathbb{E}[\exp([\mathcal{E}_k^g(z_p^{k-1})]^2/\bar{\sigma}^2)] < \infty, \quad \mathbb{E}[\exp([\mathcal{E}_k^h(x^k)]^2/\bar{\rho}^2)] < \infty, \quad \forall k \in \mathbb{N},$$

*and for all  $x \in \mathbb{R}^n$  and all  $k \in \mathbb{N}_0$ ,  $i \in [\mathbf{n}_k^g]$ ,  $j \in [\mathbf{n}_k^h]$ , we have*

$$(4.2) \quad \mathbb{E}[\exp(\|\mathcal{G}(x, s_i^k) - \nabla f(x)\|^2/\bar{\sigma}^2)] \leq e, \quad \mathbb{E}[\exp(\|\mathcal{H}(x, t_j^k) - \nabla^2 f(x)\|^2/\bar{\rho}^2)] \leq e.$$

The inequalities and properties stated in (E.1)–(E.2) have already been discussed in [Remark 3.11](#). As we have seen and as we will verify in [Lemma 4.3](#) and in [subsection 4.3](#) in more detail, these conditions allow us to bound and control the error terms  $\mathcal{E}_k^g$  and  $\mathcal{E}_k^h$  by means of the sample sizes  $\mathbf{n}_k^g$  and  $\mathbf{n}_k^h$ . We note that the conditions given in [Assumption 4.2](#) are commonly used in the complexity and convergence analysis of stochastic optimization methods, see, e.g., [\[29, 13, 31\]](#). Next, we summarize several conditional, large deviation bounds for vector- and matrix-valued martingales. For more information on tail bounds and additional matrix concentration inequalities, we refer to the papers [\[41, 81\]](#). To the best of our knowledge, the “light tail” result for symmetric random matrices presented in [Lemma 4.3](#) (ii) seems to be new.

**LEMMA 4.3.** *Let  $(\mathcal{U}_k)_{k=0}^m$  be a given filtration of the  $\sigma$ -algebra  $\mathcal{F}$  and let  $\bar{\sigma} \in \mathbb{R}^m$  be a given vector with  $\bar{\sigma}_k \neq 0$  for all  $k$ . It holds:*

(i) *Let  $(X_k)_{k=1}^m$ ,  $X_k : \Omega \rightarrow \mathbb{R}^n$ , be a family of random vectors, satisfying  $X_k \in \mathcal{U}_k$ ,  $\mathbb{E}[X_k | \mathcal{U}_{k-1}] = 0$ , and  $\mathbb{E}[\|X_k\|^2 | \mathcal{U}_{k-1}] \leq \bar{\sigma}_k^2$  a.e. for all  $k \in [m]$ . Then, we have*

$$\mathbb{E}[\|\sum_{k=1}^m X_k\|^2 | \mathcal{U}_0] \leq \|\bar{\sigma}\|^2 \quad \text{and} \quad \mathbb{P}(\|\sum_{k=1}^m X_k\| \geq \tau \|\bar{\sigma}\| | \mathcal{U}_0) \leq \tau^{-2}, \quad \forall \tau > 0$$

*almost everywhere. In addition, if it holds  $\mathbb{E}[\exp(\|X_k\|^2/\bar{\sigma}_k^2) | \mathcal{U}_{k-1}] \leq \exp(1)$  a.e. and for all  $k \in [m]$ , then with probability 1 it follows*

$$\mathbb{P}(\|\sum_{k=1}^m X_k\| \geq (1 + \tau) \|\bar{\sigma}\| | \mathcal{U}_0) \leq \exp(-\tau^2/3), \quad \forall \tau > 0.$$

(ii) *Let  $(X_k)_{k=1}^m$  be a sequence of symmetric random  $n \times n$  matrices satisfying  $X_k \in \mathcal{U}_k$ ,  $\mathbb{E}[X_k | \mathcal{U}_{k-1}] = 0$ , and  $\mathbb{E}[\|X_k\|^2 | \mathcal{U}_{k-1}] \leq \bar{\sigma}_k^2$  a.e. for all  $k \in [m]$ . Then, it holds*

$$\mathbb{P}(\|\sum_{k=1}^m X_k\| \geq \tau \|\bar{\sigma}\| | \mathcal{U}_0) \leq \kappa_n \cdot \tau^{-2}, \quad \forall \tau > 0$$

*a.e. with  $\kappa_n := (2 \log(n+2) - 1)e$ . Additionally, if we have  $\mathbb{E}[\exp(\|X_k\|^2/\bar{\sigma}_k^2) | \mathcal{U}_{k-1}] \leq \exp(1)$  a.e. and for all  $k \in [m]$ , then with probability 1 it follows*

$$(4.3) \quad \mathbb{P}(\|\sum_{k=1}^m X_k\| \geq \tau \|\bar{\sigma}\| | \mathcal{U}_0) \leq 2n \cdot \exp(-\tau^2/3), \quad \forall \tau > 0.$$

*Proof.* The first result in part (i) is well-known, see [\[41, 31, 39\]](#). The associated probability bound directly follows from the conditional Markov inequality. Since the Euclidean norm is 1-smooth, the second result in part (i) follows from [\[41, Theorem 4.1\]](#). In [\[41\]](#), Juditsky and Nemirovski also verified that the spectral norm is  $\kappa_n$ -regular which implies

$$\mathbb{E}[\|\sum_{k=1}^m X_k\|^2 | \mathcal{U}_0] \leq \kappa_n \|\bar{\sigma}\|^2 \quad \text{a.e.}$$

and establishes the first bound in part (ii). The remaining result can be shown by combining the techniques presented in [\[81\]](#) and [\[41, Proposition 4.2\]](#). For the sake of completeness, an explicit proof is given in [Appendix A.3](#).  $\square$

Let us now suppose that the assumptions (E.1)–(E.2) are satisfied. Then, using the integrability condition  $\mathbb{E}[\|\mathcal{E}_k^g(z_n^{k-1})\|^2] < \infty$ , [Fact 3.4](#), and since  $\mathcal{G}$  is a Carathéodory function and the  $\sigma$ -algebras  $\mathcal{F}_{k-1}$  and  $\sigma(s_i^k)$  are independent, it follows

- $\mathbb{E}[\mathcal{G}(z_n^{k-1}, s_i^k) - \nabla f(z_n^{k-1}) \mid \mathcal{F}_{k-1}] = \mathbb{E}[\mathcal{G}(\cdot, s_i^k)](z_n^{k-1}) - \nabla f(z_n^{k-1}) = 0$ ,
- $\mathbb{E}[\|\mathcal{G}(z_n^{k-1}, s_i^k) - \nabla f(z_n^{k-1})\|^2 \mid \mathcal{F}_{k-1}] = \mathbb{E}[\|\mathcal{G}(\cdot, s_i^k) - \nabla f(\cdot)\|^2](z_n^{k-1}) \leq \bar{\sigma}^2$ ,

a.e. and for all  $i = 1, \dots, \mathbf{n}_k^g$  and  $k \in \mathbb{N}$ , see [\[12, Theorem 2.10\]](#). In a similar fashion (and since  $\sigma(t_j^k)$  and  $\hat{\mathcal{F}}_{k-1}$  are independent), we can derive almost sure bounds for the proximal gradient step  $z_p^{k-1}$  and for the Hessian error terms  $\mathcal{H}(x^k, t_j^k) - \nabla^2 f(x^k)$ ,  $j \in [\mathbf{n}_k^h]$ . Hence, the results and bounds in [Lemma 4.3](#) are applicable in this situation. We also want to point out that the integrability conditions in [Assumption 4.2](#) are only required for the well-definedness of the conditional expectations and for the associated calculus, [\[88, 12\]](#).

**4.2. Second Order Properties and Bounded Invertibility.** In this subsection, we derive a general invertibility result that can be applied to guarantee bounded invertibility of the generalized derivatives used in our stochastic semismooth Newton framework.

We start with the presentation of several second order properties of the proximity operator that are essential for our analysis and that are motivated by the results in [\[38, 50\]](#).

Since the proximity operator is a Lipschitz continuous function, Rademacher's theorem implies that the proximal mapping  $\text{prox}_r^\Lambda$  is Fréchet differentiable almost everywhere. Let  $\Omega_r^\Lambda \subset \mathbb{R}^n$  denote the set of all points at which the proximity operator  $\text{prox}_r^\Lambda$  is differentiable. Then, as shown in [\[51, Section 3.3\]](#), the following statements are valid:

- For all  $x \in \Omega_r^\Lambda$  the matrix  $\Lambda D\text{prox}_r^\Lambda(x)$  is symmetric and positive semidefinite.
- For all  $x \in \Omega_r^\Lambda$  the matrix  $\Lambda(I - D\text{prox}_r^\Lambda(x))$  is symmetric and positive semidefinite.

A continuity argument shows that the last properties are also satisfied for every generalized derivative  $D \in \partial_B \text{prox}_r^\Lambda(x)$ . In the following lemma, we summarize our observations and state an analogue result for the Clarke subdifferential of the proximity operator  $\text{prox}_r^\Lambda$ . Let us mention that Meng et al. [\[50\]](#) established a similar result for metric projections onto convex, nonempty, and closed sets. Our result also extends Theorem 3.2 in [\[61\]](#).

**LEMMA 4.4.** *Let  $\Lambda \in \mathbb{S}_{++}^n$  and  $x \in \mathbb{R}^n$  be arbitrary. Then, for every  $D \in \partial \text{prox}_r^\Lambda(x)$ , the following statements are true:*

- The matrices  $\Lambda D$  and  $\Lambda(I - D)$  are symmetric and positive semidefinite.*
- It holds  $\langle Dh, \Lambda(I - D)h \rangle \geq 0$  for all  $h \in \mathbb{R}^n$ .*

*Proof.* The first part is a consequence of  $\partial \text{prox}_r^\Lambda(x) = \text{conv}(\partial_B \text{prox}_r^\Lambda(x))$  and of the invertibility of  $\Lambda$ . The proof of the second part is identical to the proof of [\[50, Proposition 1\]](#) and therefore will be omitted.  $\square$

Next, we present the promised, local invertibility result.

**LEMMA 4.5.** *Suppose that the conditions (A.1), (D.2), and (D.5) are satisfied and let  $s$  and  $t$  be arbitrary sample mini-batches. Furthermore, let  $x^*$  and  $\Lambda_*$  be given as in [Assumption 4.1](#) with  $\lambda_M I \succeq \Lambda_* \succeq \lambda_m I$ . Then, for all  $\gamma_c \in (0, \beta_c/C)$ ,  $\beta_c \in (0, 1)$ , there exists  $\varepsilon_c > 0$  (that does not depend on  $s$  or  $t$ ) such that under the additional conditions*

$$(4.4) \quad x \in B_{\varepsilon_c}(x^*), \quad \|\nabla f(x) - G_s(x)\| \leq \varepsilon_c, \quad \|\nabla^2 f(x) - H_t(x)\| \leq 0.5\lambda_m\gamma_c,$$

*the matrices  $M \in \mathcal{M}_{s,t}^{\Lambda_*}(x)$  are all boundedly invertible with  $\|M^{-1}\| \leq C/(1 - \beta_c)$ .*

*Proof.* Let us set  $K_* := \|\nabla^2 f(x^*)\|$ . Since the multifunction  $\partial \text{prox}_r^{\Lambda_*} : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$

is upper semicontinuous, there exists  $\tilde{\varepsilon} > 0$  such that

$$(4.5) \quad \partial \text{prox}_r^{\Lambda^*}(y) \subset \partial \text{prox}_r^{\Lambda^*}(u^{\Lambda^*}(x^*)) + B_{\tilde{\delta}}(0), \quad \forall y \in B_{\tilde{\varepsilon}}(u^{\Lambda^*}(x^*)),$$

where  $\tilde{\delta} := \lambda_m \gamma_c / (4(\lambda_m + K_*))$ , see, e.g., [19, Proposition 2.6.2]. Moreover, by the continuity of the Hessian  $\nabla^2 f$ , we also have  $\|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leq 0.25 \lambda_m \gamma_c$  for all  $x \in B_{\tilde{\varepsilon}}(x^*)$  without loss of generality. Let us now set  $\varepsilon_c := \min\{(L + \lambda_m)^{-1}, 1\} \frac{\tilde{\varepsilon} \lambda_m}{2}$  and let us consider an arbitrary matrix  $M \in \mathcal{M}_{s,t}^{\Lambda^*}(x)$  with

$$M = I - D + D\Lambda_*^{-1}H_t(x) \quad \text{and} \quad D \in \partial \text{prox}_r^{\Lambda^*}(u_s^{\Lambda^*}(x)).$$

Then, due to

$$(4.6) \quad \begin{aligned} \|u_s^{\Lambda^*}(x) - u^{\Lambda^*}(x^*)\| &\leq \|u_s^{\Lambda^*}(x) - u^{\Lambda^*}(x)\| + \|u^{\Lambda^*}(x) - u^{\Lambda^*}(x^*)\| \\ &\leq \lambda_m^{-1} \|\nabla f(x) - G_s(x)\| + (1 + L\lambda_m^{-1})\|x - x^*\| \leq \tilde{\varepsilon} \end{aligned}$$

and (4.5), there exists  $D_* \in \partial \text{prox}_r^{\Lambda^*}(u^{\Lambda^*}(x^*))$  such that  $\|D - D_*\| \leq \tilde{\delta}$ . Using Lemma 4.4 (i), we have  $\|D\Lambda_*^{-1}\| = \|\Lambda_*^{-\frac{1}{2}}[\Lambda_*^{\frac{1}{2}}D\Lambda_*^{-\frac{1}{2}}]\Lambda_*^{-\frac{1}{2}}\| \leq \|\Lambda_*^{-1}\|$ . Thus, defining  $M_* := I - D_* + D_*\Lambda_*^{-1}\nabla^2 f(x^*) \in \mathcal{M}^{\Lambda^*}(x^*)$ , it follows

$$\begin{aligned} \|M - M_*\| &= \|(D_* - D)(I - \Lambda_*^{-1}\nabla^2 f(x^*)) + D\Lambda_*^{-1}(H_t(x) - \nabla^2 f(x^*))\| \\ &\leq (1 + \lambda_m^{-1}K_*)\tilde{\delta} + \lambda_m^{-1}(\|\nabla^2 f(x) - H_t(x)\| + \|\nabla^2 f(x) - \nabla^2 f(x^*)\|) \leq \gamma_c \end{aligned}$$

Due to (D.5) and  $\gamma_c \leq \beta_c/C$ , it now holds  $\|M_*^{-1}(M - M_*)\| \leq \beta_c < 1$ . Consequently, by the Banach perturbation lemma,  $M$  is invertible with

$$\|M^{-1}\| = \|(M_* + M - M_*)^{-1}\| \leq \frac{\|M_*^{-1}\|}{1 - \|M_*^{-1}(M - M_*)\|} \leq \frac{C}{1 - \beta_c}.$$

This finishes the proof of Lemma 4.5.  $\square$

**4.3. Transition to Fast Local Convergence and Convergence Rates.** We now present our local convergence theory. As mentioned, our analysis and results rely on the observation that the stochastic Newton step  $z_n^k = x^k + d^k$  is always accepted as a new iterate with high probability if  $x^k$  is close to a local solution  $x^*$  and if the sample sizes  $\mathbf{n}_k^g$  and  $\mathbf{n}_k^h$  are sufficiently large. This will be discussed in detail in the next theorem. Throughout this subsection, we will work with the following functions and constants  $\mu(x) := \min\{(2L_FC)^{-1}, 1\}x$ ,  $\mu_p(x) := \min\{x, \min\{x^{p-1}, x^{(1-p)^{-1}}\}\}$ ,

$$\Upsilon_k := \min\{\mu(\varepsilon_k^1), \mu_p(\varepsilon_k^2)\}, \quad \text{and} \quad \Gamma_k := \min\{\Upsilon_{k-1}, \Upsilon_k\},$$

where  $L_F$  denotes the Lipschitz constant of  $F^{\Lambda^*}$ . (See the proof of Theorem 3.10). Furthermore, let us set  $\beta_1 := (2^p L_\psi C)^{-1} \min\{\beta, \frac{1}{2}\}$  and

$$(4.7) \quad \beta_2 := \min\left\{\frac{6\eta}{4L_FC + 3\eta}, \beta_1^{\frac{1}{1-p}}\right\}, \quad \gamma_f := \frac{1}{2 \max\{C, 1\}} \min\left\{\frac{1}{2}, \beta_2\right\}.$$

**THEOREM 4.6.** *Assume that the conditions (A.1)–(A.2), (B.1)–(B.2), (C.1), and (E.1)–(E.2) are satisfied and let the sequences  $(x^k)_k$ ,  $(\Lambda_k)_k$ , and  $(\alpha_k)_k$  be generated by Algorithm 1. Let  $x^*$  and  $\Lambda_*$  be accumulation points of  $(x^k)_k$  and  $(\Lambda_k)_k$  fulfilling the conditions (D.1)–(D.6) and let  $\gamma_f \in (0, 1)$  be given as in (4.7). Moreover, suppose that the selected step sizes  $(\alpha_k)_k$  are bounded via  $\alpha_k \in [\underline{\alpha}, \min\{1, \bar{\alpha}\}]$  for some  $\underline{\alpha} > 0$  and all  $k \in \mathbb{N}$ . Then, for a given sequence  $(\delta_k)_k \subset (0, 1)$ , the following statements are true:*



(i) Suppose that there exists  $\bar{\ell} \in \mathbb{N}$  such that

$$(4.8) \quad \mathbf{n}_k^g \geq \frac{1}{\delta_k} \left[ \frac{2\bar{\sigma}}{\lambda_m \Gamma_k} \right]^2, \quad \mathbf{n}_k^h \geq \frac{\kappa_n}{\delta_k} \left[ \frac{2\bar{\rho}}{\lambda_m \gamma_f} \right]^2, \quad \forall k \geq \bar{\ell}.$$

Then, with probability  $\delta_* := \prod_{k=\bar{\ell}}^\infty (1 - \delta_k)(1 - 2\delta_k)$ , the point  $x^*$  is a stationary point of (1.1), there exists  $\ell_* \in \mathbb{N}$  such that  $x^k$  results from a stochastic semismooth Newton step for all  $k \geq \ell_*$ , and the whole sequence  $(x^k)_k$  converges to  $x^*$ .

(ii) Let  $\gamma_\eta \in (0, 1)$  and  $\bar{\ell} \in \mathbb{N}$  be given constants and let us set  $\Gamma_k^\circ := \min\{\Upsilon_{k-1}^\circ, \Upsilon_k^\circ\}$  and  $\Upsilon_k^\circ := \min\{\mu(\min\{\varepsilon_k^1, \gamma_\eta^{k-\bar{\ell}}\}), \mu_p(\varepsilon_k^2)\}$ . Suppose that the bounds

$$(4.9) \quad \mathbf{n}_k^g \geq \frac{1}{\delta_k} \left[ \frac{2\bar{\sigma}}{\lambda_m \Gamma_k^\circ} \right]^2, \quad \mathbf{n}_k^h \geq \frac{\kappa_n}{\delta_k} \left[ \frac{2\bar{\rho}}{\lambda_m \gamma_f} \right]^2$$

hold for all  $k \geq \bar{\ell}$ . Then, with probability  $\delta_*$ , the statements in part (i) are satisfied and  $(x^k)_k$  converges  $r$ -linearly to  $x^*$  with rate  $\max\{\gamma_\eta, \frac{1}{2}\}$ .

(iii) Let  $(\gamma_k)_k \subset (0, \infty)$  be a non-increasing sequence with  $\gamma_k \rightarrow 0$  and let  $(\rho_k)_k \subset (0, \infty)$  with  $\rho_k \rightarrow 0$  and  $\bar{\ell} \in \mathbb{N}$  be given. Let us define  $\Gamma_k^\circ := \min\{\Upsilon_{k-1}^\circ, \Upsilon_k^\circ\}$  and  $\Upsilon_k^\circ := \min\{\mu(\min\{\varepsilon_k^1, \gamma_k^{k-\bar{\ell}}\}), \mu_p(\varepsilon_k^2)\}$  and assume that the sample sizes fulfill

$$(4.10) \quad \mathbf{n}_k^g \geq \frac{1}{\delta_k} \left[ \frac{2\bar{\sigma}}{\lambda_m \Gamma_k^\circ} \right]^2, \quad \mathbf{n}_k^h \geq \frac{1}{\delta_k \rho_k}$$

for all  $k \geq \bar{\ell}$ . Then, with probability  $\delta_*$ , the statements in part (i) are satisfied (for a possibly different  $\ell_*$ ) and  $(x^k)_k$  converges  $r$ -superlinearly to  $x^*$ .

*Proof.* The proof is split into several steps. First, we utilize the concentration results in Lemma 4.3 to quantify the occurrence and (conditional) probability of the events

$$\mathbf{G}_k^n(\varepsilon) := \{\omega \in \Omega : \mathcal{E}_k^g(z_n^{k-1}(\omega)) \leq \varepsilon\}, \quad \mathbf{G}_k^p(\varepsilon) := \{\omega \in \Omega : \mathcal{E}_k^g(z_p^{k-1}(\omega)) \leq \varepsilon\},$$

$\mathbf{G}_k(\varepsilon) := \mathbf{G}_k^n(\varepsilon) \cap \mathbf{G}_k^p(\varepsilon)$ , and  $\mathbf{H}_k(\varepsilon) := \{\omega \in \Omega : \mathcal{E}_k^h(x^k(\omega)) \leq \varepsilon\}$  for appropriately chosen  $\varepsilon > 0$ . As a result and to some extent, the local convergence analysis then reduces to a discussion of a highly inexact (rather than stochastic) version of the deterministic semismooth Newton method for convex composite programming. In particular, we can reuse some of the strategies presented in [52, 51] in our proof. Based on the invertibility result in Lemma 4.5, we establish convergence of the whole sequence  $(x^k)_k$  in step 2. Afterwards, in step 3 and 4, we show that the growth conditions (2.8)–(2.9) are always satisfied whenever  $k$  is sufficiently large. In the last steps we derive  $r$ -linear and  $r$ -superlinear convergence rates and prove part (ii) and (iii) of Theorem 4.6.

*Step 1: Probability bounds.* We want to show that the event  $\mathbf{E} := \bigcap_{k=\bar{\ell}}^\infty \mathbf{G}_k(\lambda_m \Gamma_k/2) \cap \mathbf{H}_k(\lambda_m \gamma_f/2)$  occurs with probability  $\mathbb{P}(\mathbf{E}) \geq \delta_*$ . Using the assumptions (E.1)–(E.2), Fact 3.4 and as demonstrated in the paragraph after Lemma 4.3, we can apply the concentration results in Lemma 4.3 via identifying  $\mathbf{X}_i \equiv \mathcal{G}(z_n^{k-1}, s_i^k) - \nabla f(z_n^{k-1})$ ,  $i \in [\mathbf{n}_k^g]$ , etc. Specifically, setting  $\tau = \sqrt{1/\delta_k}$  in part (i) and  $\tau = \sqrt{\kappa_n/\delta_k}$  in part (ii) of Lemma 4.3 and using the bounds (4.8), it easily follows  $\mathbb{P}(\mathbf{G}_k^n(\lambda_m \Gamma_k/2) \mid \mathcal{F}_{k-1}) \geq 1 - \delta_k$ ,

$$\mathbb{P}(\mathbf{G}_k^p(\lambda_m \Gamma_k/2) \mid \mathcal{F}_{k-1}) \geq 1 - \delta_k, \quad \text{and} \quad \mathbb{P}(\mathbf{H}_k(\lambda_m \gamma_f/2) \mid \hat{\mathcal{F}}_{k-1}) \geq 1 - \delta_k$$

almost everywhere and for all  $k \geq \bar{\ell}$ . Let us define  $\bar{\varepsilon}_k := \frac{\lambda_m \Gamma_k}{2}$  and  $\bar{\gamma} := \frac{\lambda_m \gamma_f}{2}$ . Then, by the tower property of the conditional expectation and by utilizing

$$\mathbb{1}_{\mathbf{G}_k(\bar{\varepsilon}_k)}(\omega) = \mathbb{1}_{\mathbf{G}_k^n(\bar{\varepsilon}_k) \cap \mathbf{G}_k^p(\bar{\varepsilon}_k)}(\omega) \geq \mathbb{1}_{\mathbf{G}_k^n(\bar{\varepsilon}_k)}(\omega) + \mathbb{1}_{\mathbf{G}_k^p(\bar{\varepsilon}_k)}(\omega) - 1, \quad \forall \omega \in \Omega,$$

and  $G_k^n(\bar{\varepsilon}_k), G_k^p(\bar{\varepsilon}_k) \in \hat{\mathcal{F}}_{k-1}, H_k(\bar{\gamma}) \in \mathcal{F}_k, k \geq \bar{\ell}$ , we inductively obtain

$$\begin{aligned} & \mathbb{P}\left(\bigcap_{k=\bar{\ell}}^L G_k(\bar{\varepsilon}_k) \cap H_k(\bar{\gamma})\right) \\ &= \mathbb{E}\left[\prod_{k=\bar{\ell}}^{L-1} \mathbb{1}_{G_k(\bar{\varepsilon}_k)} \mathbb{1}_{H_k(\bar{\gamma})} \left\{ \mathbb{E}[\mathbb{1}_{G_L^n(\bar{\varepsilon}_L) \cap G_L^p(\bar{\varepsilon}_L)} \mathbb{E}[\mathbb{1}_{H_L(\bar{\gamma})} \mid \hat{\mathcal{F}}_{L-1}] \mid \mathcal{F}_{L-1}] \right\}\right] \\ &\geq (1 - 2\delta_L)(1 - \delta_L) \cdot \mathbb{E}\left[\prod_{k=\bar{\ell}}^{L-1} \mathbb{1}_{G_k(\bar{\varepsilon}_k)} \mathbb{1}_{H_k(\bar{\gamma})}\right] \geq \dots \geq \prod_{k=\bar{\ell}}^L (1 - 2\delta_k)(1 - \delta_k) \end{aligned}$$

for any  $L > \bar{\ell}$ . Hence, taking the limit  $L \rightarrow \infty$ , this yields  $\mathbb{P}(E) \geq \delta_*$ . We now assume that the trajectories  $(x^k)_k, (z_n^k)_k$ , and  $(z_p^k)_k$  are generated by a sample point  $\bar{\omega} \in E$ , i.e., we have  $(x^k)_k \equiv (x^k(\bar{\omega}))_k$  etc. (As we have just shown this happens with probability at least  $\delta_*$ ).

*Step 2: Convergence of  $(x^k)_k$ .* Let us continue with the proof of the first part. Using the definition of the event  $G_k$ , we can infer

$$(4.11) \quad \mathcal{E}_k := \mathcal{E}_k^g(x^k) \leq \max\{\mathcal{E}_k^g(z_n^{k-1}), \mathcal{E}_k^g(z_p^{k-1})\} \leq \frac{\lambda_m \Gamma_k}{2} \leq \frac{\lambda_m \Upsilon_k}{2} \leq \frac{\lambda_m \varepsilon_k^1}{2}$$

for all  $k \geq \bar{\ell}$  and hence, by (B.2), the sequence  $(\mathcal{E}_k)_k$  is summable. Following the proof of [Theorem 3.10](#) and utilizing the boundedness of  $(\alpha_k)_k$ , this implies

$$(4.12) \quad F^{\Lambda*}(x^k) \rightarrow 0$$

and thus, in this situation every accumulation point of  $(x^k)_k$  is a stationary point of problem (1.1). Since the event  $H_k(\bar{\gamma})$  occurs for all  $k \geq \bar{\ell}$  by assumption, we can apply [Lemma 4.5](#) with  $\gamma_c := \gamma_f$  and  $\beta_c := \frac{1}{4}$ . Hence, there exists a constant  $\varepsilon_c$  (that neither depends on the samples  $s^k, t^k$  nor on  $k$ ) such that the statement in [Lemma 4.5](#) holds whenever we have  $x^k \in B_{\varepsilon_c}(x^*)$  and  $\mathcal{E}_k \leq \varepsilon_c$ . Furthermore, since  $(\mathcal{E}_k)_k$  converges to zero, there exists  $\bar{\ell} \geq \max\{\bar{\ell}, \bar{k}\}$  such that the inequality  $\mathcal{E}_k \leq \varepsilon_c$  is satisfied for all  $k \geq \bar{\ell}$ . Setting  $\mathcal{K}_{\varepsilon_c} := \{k \geq \bar{\ell} : x^k \in B_{\varepsilon_c}(x^*)\}$ , this now implies

$$(4.13) \quad \|M_k^{-1}\| \leq \frac{4C}{3}, \quad \forall M_k \in \mathcal{M}_{s^k, t^k}^{\Lambda*}(x^k), \quad \forall k \in \mathcal{K}_{\varepsilon_c}.$$

Next, let  $(x^k)_{k \in K}$  denote an arbitrary subsequence of  $(x^k)_k$  converging to  $x^*$ . Then, there exists  $\tilde{k} \in K$  such that  $\{k \in K : k \geq \tilde{k}\} \subset \mathcal{K}_{\varepsilon_c}$  and consequently, using (4.12), the estimate  $\|F_{s^k}^{\Lambda*}(x) - F^{\Lambda*}(x)\| \leq \lambda_m^{-1} \mathcal{E}_k^g(x)$ , and the summability of  $(\mathcal{E}_k)_k$ , we obtain

$$\|x^{k+1} - x^k\| \leq \min\{4C/3, 1\} \|F_{s^k}^{\Lambda*}(x^k)\| \leq \min\{4C/3, 1\} [\|F^{\Lambda*}(x^k)\| + \lambda_m^{-1} \mathcal{E}_k] \rightarrow 0$$

as  $K \ni k \rightarrow \infty$ . Let us emphasize that this limit behavior holds for any arbitrary subsequence  $(x^k)_K$  converging to  $x^*$ . Moreover, combining the assumptions (D.2) and (D.3) and as mentioned in [subsection 2.2](#), it follows that  $F^{\Lambda*}$  is semismooth at  $x^*$  with respect to the multifunction  $\mathcal{M}^{\Lambda*}$ . Thus, as in [59, Proposition 3], it can be shown that  $x^*$  is an isolated stationary point. Since condition (4.12) ensures that every accumulation point of  $(x^k)_k$  is a stationary point of problem (1.1), this also proves that  $x^*$  is an isolated accumulation point of  $(x^k)_k$ . Hence, a well-known result by Moré and Sorensen [55] yields convergence of the whole sequence  $(x^k)_k$  to  $x^*$ .

*Step 3: Acceptance of Newton steps.* Utilizing the assumptions (D.2)–(D.4), and (D.6), there exists  $\varepsilon_s > 0$  such that the following properties and inequalities hold simultaneously:

- Let us set  $\bar{\beta} := (12(1 + L\lambda_m^{-1}))^{-1}$ . Then, for all  $u \in B_{\varepsilon_s}(u^{\Lambda*}(x^*))$  and all generalized derivatives  $D(u) \in \partial \text{prox}_r^{\Lambda*}(u)$  we have

$$(4.14) \quad \|\text{prox}_r^{\Lambda^*}(u) - \text{prox}_r^{\Lambda^*}(u^{\Lambda^*}(x^*)) - D(u)(u - u^{\Lambda^*}(x^*))\| \leq \bar{\beta}\gamma_f \|u - u^{\Lambda^*}(x^*)\|.$$

- It holds  $\|\nabla f(x) - \nabla f(x^*) - \nabla^2 f(x^*)(x - x^*)\| \leq \frac{\lambda_m \gamma_f}{12} \|x - x^*\|$  and  $\|\nabla^2 f(x) - \nabla^2 f(x^*)\| \leq \frac{\lambda_m \gamma_f}{12}$  for all  $x \in B_{\varepsilon_s}(x^*)$ .
- The objective function  $\psi$  is Lipschitz continuous on  $B_{\varepsilon_s}(x^*)$  (with constant  $L_\psi$ ).
- The stationary point  $x^*$  is a global minimizer of  $\psi$  on  $B_{\varepsilon_s}(x^*)$ .

Now, let us define

$$\varepsilon_e := \min \left\{ \frac{2\lambda_m \varepsilon_s}{3\lambda_m + 2L}, \varepsilon_c \right\}, \quad \varepsilon_\psi := \min \left\{ \frac{1}{2} (2L_\psi C^{1-p})^{-\frac{1}{p}}, \frac{1}{2L_F} \right\}, \quad \varepsilon := \min \{ \varepsilon_e, \varepsilon_\psi \}.$$

Then, since  $(x^k)_k$  converges to  $x^*$  and (B.2) implies  $\Gamma_k \rightarrow 0$ , there exists  $\hat{\ell} \geq \tilde{\ell}$  such that  $x^k \in B_\varepsilon(x^*)$  and  $\min\{(\max\{C, 1\})^{-1}\varepsilon, \Gamma_k\} = \Gamma_k$  for all  $k \geq \hat{\ell}$ . Next, let  $x^k, k \geq \hat{\ell}$ , be an arbitrary iterate and let us consider the associated semismooth Newton step  $z_n^k = x^k + d^k$ ,  $d^k = -M_k^+ F_{s^k}^{\Lambda^*}(x^k)$  with  $M_k = I - D_k + D_k \Lambda_*^{-1} H_{t^k}(x^k)$  and  $D_k \in \partial \text{prox}_r^{\Lambda^*}(u_{s^k}^{\Lambda^*}(x^k))$ . By Lemma 4.5 and (4.13),  $M_k$  is invertible with  $\|M_k^{-1}\| \leq 2C$  for all  $k \geq \hat{\ell}$ . Since the events  $G_k(\lambda_m \Gamma_k/2)$  and  $G_{k+1}^n(\lambda_m \Gamma_{k+1}/2)$  occur for all  $k \geq \bar{\ell}$  by assumption, we can reuse the bounds (4.11), i.e., we have

$$(4.15) \quad \max\{\mathcal{E}_k, \mathcal{E}_{k+1}^g(z_n^k)\} \leq \frac{\lambda_m}{2} \min \left\{ \frac{\varepsilon}{\max\{C, 1\}}, \Upsilon_k \right\}.$$

Consequently, setting  $w^k := u_{s^k}^{\Lambda^*}(x^k) - u^{\Lambda^*}(x^*)$  and as shown in the proof of Lemma 4.5, it holds

$$(4.16) \quad \|w^k\| \leq (1 + L\lambda_m^{-1})\|x^k - x^*\| + \lambda_m^{-1}\mathcal{E}_k \leq (1 + L\lambda_m^{-1})\varepsilon_e + \frac{\varepsilon_e}{2} \leq \varepsilon_s.$$

Moreover, combining (4.14),  $\mathcal{E}_k^h(x^k) \leq \bar{\gamma}, \bar{\beta}\gamma_f \leq 1/48$ , and the last estimates, we can infer

$$\begin{aligned} \|z_n^k - x^*\| &= \|M_k^{-1}[F_{s^k}^{\Lambda^*}(x^k) - F^{\Lambda^*}(x^*) - M_k(x^k - x^*)]\| \\ &\leq (4C/3)\|\text{prox}_r^{\Lambda^*}(u_{s^k}^{\Lambda^*}(x^k)) - \text{prox}_r^{\Lambda^*}(u^{\Lambda^*}(x^*)) \\ &\quad - D_k(I - \Lambda_*^{-1}H_{t^k}(x^k))(x^k - x^*)\| \\ &\leq (4C/3)\|\text{prox}_r^{\Lambda^*}(u^{\Lambda^*}(x^*) + w^k) - \text{prox}_r^{\Lambda^*}(u^{\Lambda^*}(x^*)) - D_k w^k\| \\ &\quad + (4C/3)\|D_k \Lambda_*^{-1}[G_{s^k}(x^k) - \nabla f(x^*) - H_{t^k}(x^k)(x^k - x^*)]\| \\ &\leq (4C/3)\bar{\beta}\gamma_f \|w^k\| + (4C/3)\lambda_m^{-1}[\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^k - x^*)\| \\ &\quad + (\|\nabla^2 f(x^k) - \nabla^2 f(x^*)\| + \mathcal{E}_k^h(x^k))\|x^k - x^*\|] + \mathcal{E}_k \\ &\leq \frac{4C}{3} \left[ \bar{\beta}(1 + L\lambda_m^{-1}) + \frac{1}{12} + \frac{1}{12} + \frac{1}{2} \right] \gamma_f \|x^k - x^*\| + \frac{4C}{3}\lambda_m^{-1} \left[ \frac{1}{48} + 1 \right] \mathcal{E}_k \\ (4.17) \quad &\leq C\gamma_f \|x^k - x^*\| + 1.5C\lambda_m^{-1}\mathcal{E}_k. \end{aligned}$$

Let us note that due to (4.17), (4.15),  $C\gamma_f \leq \frac{1}{4}$ , and (D.4), we also have  $z_n^k \in B_\varepsilon(x^*) \subset \text{dom } r$ . The next steps essentially follow the proofs of [52, Theorem 4.8] and [51, Theorem 4.3.10]. In particular, our last result implies

$$\|x^k - x^*\| \leq \|z_n^k - x^*\| + \|d^k\| \leq C\gamma_f \|x^k - x^*\| + 1.5C\lambda_m^{-1}\mathcal{E}_k + (4C/3)\|F_{s^k}^{\Lambda^*}(x^k)\|$$

and thus, it follows

$$\|x^k - x^*\| \leq \frac{4C}{3(1 - C\gamma_f)} \|F_{s^k}^{\Lambda^*}(x^k)\| + \frac{3C\lambda_m^{-1}}{2(1 - C\gamma_f)} \mathcal{E}_k.$$

Furthermore, using the Lipschitz continuity of  $F^{\Lambda_*}$ ,  $F^{\Lambda_*}(x^*) = 0$ ,  $(1 - C\gamma_f)^{-1} \leq \frac{4}{3}$ , the definition of  $\gamma_f$ , and (4.15), we obtain

$$\begin{aligned} \|F_{s^{k+1}}^{\Lambda_*}(z_n^k)\| &\leq \|F^{\Lambda_*}(z_n^k)\| + \|F_{s^{k+1}}^{\Lambda_*}(z_n^k) - F^{\Lambda_*}(z_n^k)\| \leq L_F \|z_n^k - x^*\| + \lambda_m^{-1} \mathcal{E}_{k+1}^g(z_n^k) \\ &\leq L_F C \gamma_f \|x^k - x^*\| + 1.5 L_F C \lambda_m^{-1} \mathcal{E}_k + \lambda_m^{-1} \mathcal{E}_{k+1}^g(z_n^k) \\ &\leq \frac{4 L_F C^2 \gamma_f}{3(1 - C\gamma_f)} \|F_{s^k}^{\Lambda_*}(x^k)\| + \frac{3 L_F C \lambda_m^{-1}}{2(1 - C\gamma_f)} \mathcal{E}_k + \lambda_m^{-1} \mathcal{E}_{k+1}^g(z_n^k) \\ &\leq \eta \|F_{s^k}^{\Lambda_*}(x^k)\| + 2 L_F C \lambda_m^{-1} \mathcal{E}_k + \lambda_m^{-1} \mathcal{E}_{k+1}^g(z_n^k) \leq \eta \|F_{s^k}^{\Lambda_*}(x^k)\| + \varepsilon_k^1. \end{aligned}$$

Since the Newton step  $z_n^k$  is contained in  $B_\varepsilon(x^*)$ , we can utilize the semismoothness condition (4.14) for  $z_n^k$ . Moreover, Lemma 4.5 is also applicable for any matrix  $M_* = I - \bar{D}(z_n^k) + \bar{D}(z_n^k) \Lambda_*^{-1} \nabla^2 f(x^*)$  with  $\bar{D}(z_n^k) \in \partial \text{prox}_r^{\Lambda_*}(u^{\Lambda_*}(z_n^k))$ . This yields  $\|M_*^{-1}\| \leq (4C)/3$  and similar to the estimates in (4.17) (but simpler), we can get

$$\begin{aligned} \|z_n^k - x^*\| &= \|M_*^{-1}[F^{\Lambda_*}(z_n^k) - F^{\Lambda_*}(x^*) - M_*(z_n^k - x^*) - F^{\Lambda_*}(z_n^k)]\| \\ &\leq (4C/3)[\|\text{prox}_r^{\Lambda_*}(u^{\Lambda_*}(z_n^k)) - \text{prox}_r^{\Lambda_*}(u^{\Lambda_*}(x^*)) - \bar{D}(z_n^k)(I - \Lambda_*^{-1} \nabla^2 f(x^*)) \\ &\quad \cdot (z_n^k - x^*)\| + \|F_{s^{k+1}}^{\Lambda_*}(z_n^k)\| + \lambda_m^{-1} \mathcal{E}_{k+1}^g(z_n^k)] \\ &\leq (4C/3)[\|\text{prox}_r^{\Lambda_*}(u^{\Lambda_*}(z_n^k)) - \text{prox}_r^{\Lambda_*}(u^{\Lambda_*}(x^*)) - \bar{D}(z_n^k) \bar{w}^k\| + \|F_{s^{k+1}}^{\Lambda_*}(z_n^k)\| \\ &\quad + \|\bar{D}(z_n^k) \Lambda_*^{-1} [\nabla f(z_n^k) - \nabla f(x^*) - \nabla^2 f(x^*)(z_n^k - x^*)]\| + \lambda_m^{-1} \mathcal{E}_{k+1}^g(z_n^k)] \\ &\leq \frac{2}{9} C \gamma_f \|z_n^k - x^*\| + \frac{4C}{3} \|F_{s^{k+1}}^{\Lambda_*}(z_n^k)\| + \frac{4C}{3} \lambda_m^{-1} \mathcal{E}_{k+1}^g(z_n^k), \end{aligned}$$

where we used  $\bar{w}^k := u^{\Lambda_*}(z_n^k) - u^{\Lambda_*}(x^*)$ ,  $\|\bar{w}^k\| \leq (1 + L\lambda_m^{-1})\|z_n^k - x^*\|$ , (4.14), and the differentiability of  $\nabla f$ . This implies

$$\|z_n^k - x^*\| \leq \frac{12C}{9 - 2C\gamma_f} \|F_{s^{k+1}}^{\Lambda_*}(z_n^k)\| + \frac{12C\lambda_m^{-1}}{9 - 2C\gamma_f} \mathcal{E}_{k+1}^g(z_n^k).$$

Finally, using condition (D.4),  $(9 - 2C\gamma_f)^{-1} \leq \frac{1}{8}$ , the subadditivity of the mapping  $x \mapsto x^q$ ,  $q \in \{p, 1 - p\} \subset (0, 1)$ ,  $0.5^q + 1.5^q \leq 2$  for all  $q \in [0, 1]$ , and the fact that the stationary point  $x^*$  is a global minimum of the problem  $\min_{x \in B_\varepsilon(x^*)} \psi(x)$ , it follows

$$\begin{aligned} \psi(z_n^k) - \psi(x^k) &\leq \psi(z_n^k) - \psi(x^*) \leq L_\psi \|z_n^k - x^*\| = L_\psi \|z_n^k - x^*\|^{1-p} \|z_n^k - x^*\|^p \\ &\leq L_\psi C^{1-p} [\gamma_f^{1-p} \|x^k - x^*\|^{1-p} + (1.5\lambda_m^{-1} \mathcal{E}_k)^{1-p}] \|z_n^k - x^*\|^p \\ &\leq L_\psi C^{1-p} \left[ (2C\gamma_f \|F_{s^k}^{\Lambda_*}(x^k)\|)^{1-p} + ((2C\gamma_f)^{1-p} + 1.5^{1-p})(\lambda_m^{-1} \mathcal{E}_k)^{1-p} \right] \|z_n^k - x^*\|^p \\ &\leq L_\psi (2C^2 \gamma_f)^{1-p} \|F_{s^k}^{\Lambda_*}(x^k)\|^{1-p} [(1.5C)^p \|F_{s^{k+1}}^{\Lambda_*}(z_n^k)\|^p + (1.5C\lambda_m^{-1} \mathcal{E}_{k+1}^g(z_n^k))^p] \\ &\quad + 2L_\psi (C\lambda_m^{-1})^{1-p} \mathcal{E}_k^{1-p} \|z_n^k - x^*\|^p \\ &\leq 2L_\psi C (C\gamma_f)^{1-p} \|F_{s^k}^{\Lambda_*}(x^k)\|^{1-p} \|F_{s^{k+1}}^{\Lambda_*}(z_n^k)\|^p + 0.5\varepsilon_k^2 \\ &\quad + 2L_\psi C (C\gamma_f)^{1-p} (L_F \|x^k - x^*\| + \lambda_m^{-1} \mathcal{E}_k)^{1-p} \lambda_m^{-p} [\mathcal{E}_{k+1}^g(z_n^k)]^p \\ &\leq \beta \|F_{s^k}^{\Lambda_*}(x^k)\|^{1-p} \|F_{s^{k+1}}^{\Lambda_*}(z_n^k)\|^p + \frac{\varepsilon_k^2}{2} + \frac{1}{2} (2^{p-1} + 2^{p-1} \mu_p (\varepsilon_k^2)^{1-p}) 2^{-p} \mu_p (\varepsilon_k^2)^p \\ &\leq \beta \|F_{s^k}^{\Lambda_*}(x^k)\|^{1-p} \|F_{s^{k+1}}^{\Lambda_*}(z_n^k)\|^p + \varepsilon_k^2. \end{aligned}$$

*Step 4: Transition to fast local convergence.* Let us point out that the bounds and inequalities derived in the last part primarily depend on the occurrence of the events  $G_k$  and  $H_k$  and hold for any  $k \geq \hat{\ell}$ . Now, let  $k$  be any index  $k \geq \hat{\ell}$  with  $\|F_{s^k}^{\Lambda^*}(x^k)\| \leq \theta_k$ . Since the algorithm does not terminate after a finite number of steps and we have  $F_{s^k}^{\Lambda^*}(x^k) \rightarrow 0$ , there exist infinitely many such indices. Let  $\ell_* - 1$  be the smallest such index. Then, as shown in step 3, the semismooth Newton step  $z_n^{\ell_*-1}$  satisfies all of the acceptance criteria and it follows  $x^{\ell_*} = z_n^{\ell_*-1}$  and  $\theta_{\ell_*} = \|F_{s^{\ell_*}}^{\Lambda^*}(x^{\ell_*})\|$ . Inductively, this implies  $x^{k+1} = z_n^k$  for all  $k \geq \ell_* - 1$ . Since the success probability of the event  $E$  is at least  $\delta_*$ , this finishes the proof of the first part.

*Step 5: Proof of part (ii).* Revisiting the derivations in step 1, it is easy to see that the event  $E^\circ := \bigcap_{k=\bar{\ell}}^\infty G_k(\lambda_m \Gamma_k^\circ/2) \cap H_k(\lambda_m \gamma_f/2)$  also occurs with probability  $\delta_*$ . Moreover, due to  $E^\circ \subset E$ , all of the results and inequalities shown in step 2–4 remain valid. In particular, the estimate (4.17) holds for all  $k \geq \ell_* - 1$  with  $x^{k+1} = z_n^k$  and using  $L_F \geq 1$ , we obtain

$$\|x^{k+1} - x^*\| \leq C\gamma_f \|x^k - x^*\| + \frac{3}{4}C\Upsilon_k^\circ \leq \frac{1}{2} \left[ \frac{1}{2} \|x^k - x^*\| + \gamma_\eta^{k-\bar{\ell}} \right].$$

Let us set  $\bar{\tau} := \max\{\|x^{\ell_*-1} - x^*\|, \gamma_\eta^{\ell_*-\bar{\ell}-2}\}$  and  $\tau_k := \bar{\tau} \max\{\frac{1}{2}, \gamma_\eta\}^{k-\ell_*+1}$ . We now prove  $\|x^k - x^*\| \leq \tau_k$  for all  $k \geq \ell_* - 1$  by induction. For  $k = \ell_* - 1$ , it follows  $\tau_{\ell_*-1} = \bar{\tau}$  and the latter inequality is obviously true. Furthermore, it holds

$$\begin{aligned} \|x^{k+1} - x^*\| &\leq \frac{1}{2} \max\{\frac{1}{2}, \gamma_\eta\} [\tau_k + \gamma_\eta^{k-\bar{\ell}-1}] \\ &\leq \frac{1}{2} \max\{\frac{1}{2}, \gamma_\eta\}^{k-\ell_*+2} [\bar{\tau} + \gamma_\eta^{\ell_*-\bar{\ell}-2}] \leq \tau_{k+1}. \end{aligned}$$

Consequently, due to  $\tau_k \rightarrow 0$  and  $\tau_{k+1}/\tau_k = \max\{\frac{1}{2}, \gamma_\eta\} < 1$ , the sequence  $(x^k)_k$  converges  $r$ -linearly to  $x^*$  with rate  $\max\{\frac{1}{2}, \gamma_\eta\}$ .

*Step 6: Proof of part (iii).* Again, following our previous discussions, it can be easily shown that the event  $E^\circ := \bigcap_{k=\bar{\ell}}^\infty G_k(\lambda_m \Gamma_k^\circ/2) \cap H_k(\bar{\rho}\sqrt{\kappa_n \rho_k})$  occurs with probability  $\delta_*$ . Since  $(\rho_k)_k$  converges to zero, we have  $H_k(\bar{\rho}\sqrt{\kappa_n \rho_k}) \subset H_k(\lambda_m \gamma_f/2)$  for all  $k$  sufficiently large and hence, the results derived in step 2–4 still hold after possibly adjusting the constant  $\ell_*$ . In particular, we have  $x^k \rightarrow x^*$  as  $k \rightarrow \infty$ . Next, let us set

$$\begin{aligned} \beta_k^1 &= \frac{\|\text{prox}_r^{\Lambda^*}(u^{\Lambda^*}(x^*) + w^k) - \text{prox}_r^{\Lambda^*}(u^{\Lambda^*}(x^*)) - D_k w^k\|}{\|w^k\|}, \\ \beta_k^2 &= \frac{\|\nabla f(x^k) - \nabla f(x^*) - \nabla^2 f(x^*)(x^k - x^*)\|}{\|x^k - x^*\|}, \end{aligned}$$

and  $\beta_k^3 = \|\nabla^2 f(x^k) - \nabla^2 f(x^*)\| + \mathcal{E}_k^h(x^k)$ , where  $w^k$  and  $D_k$  have been defined in step 3. Since  $(w^k)_k$  converges to zero, the semismoothness of  $\text{prox}_r^{\Lambda^*}$  implies  $\beta_k^1 \rightarrow 0$ . Moreover, by the differentiability of  $f$  and using  $\rho_k \rightarrow 0$ , we obtain  $\beta_k^2 \rightarrow 0$  and  $\beta_k^3 \rightarrow 0$ . Without loss of generality, we now may assume  $x^{k+1} = z_n^k$  for all  $k \geq \ell_* - 1$  and as a consequence, by (4.16) and (4.17) and defining  $\vartheta_k := (4C/3)\lambda_m^{-1}[(\lambda_m + L)\beta_k^1 + \beta_k^2 + \beta_k^3]$ , we have

$$\begin{aligned} \|x^{k+1} - x^*\| &\leq (4C/3)\lambda_m^{-1}[(\lambda_m + L)\beta_k^1 + \beta_k^2 + \beta_k^3] \|x^k - x^*\| + (4C/3)\lambda_m^{-1}(\bar{\beta}\gamma_f + 1)\mathcal{E}_k \\ &\leq \vartheta_k \|x^k - x^*\| + (3C/4)\Upsilon_k^\circ \leq \vartheta_k \|x^k - x^*\| + \gamma_\eta^{k-\bar{\ell}} \end{aligned}$$

for all  $k \geq \ell_* - 1$ . Next, due to  $\vartheta_k, \gamma_k \rightarrow 0$ , there exists a constant  $\ell_\diamond \geq \ell_*$  such that  $\vartheta_k, \gamma_k \leq \frac{1}{2}$  for all  $k \geq \ell_\diamond$ . Let us set

$$\tau_{\ell_\diamond+1} := \max \left\{ \|x^{\ell_\diamond} - x^*\|, \gamma_{\ell_\diamond}^{(\ell_\diamond-\bar{\ell})/2} \right\}, \quad \tau_{k+1} := \max \left\{ (\vartheta_k + \gamma_k^{(k-\bar{\ell})/2}) \tau_k, \gamma_{k+1}^{(k+1-\bar{\ell})/2} \right\}$$

for all  $k > \ell_\diamond$ . Then, by induction and by using  $\tau_k \geq \gamma_k^{(k-\bar{\ell})/2}$ , it follows

$$\|x^{k+1} - x^*\| \leq \vartheta_k \|x^k - x^*\| + \gamma_k^{(k-\bar{\ell})/2} \tau_k \leq (\vartheta_k + \gamma_k^{(k-\bar{\ell})/2}) \tau_k \leq \tau_{k+1}$$

for all  $k > \ell_\diamond$ . Due to  $\vartheta_k, \gamma_k \leq \frac{1}{2}$ , this also establishes  $\|x^{\ell_\diamond+1} - x^*\| \leq \tau_{\ell_\diamond+1}$ . Finally, utilizing the boundedness of  $(\tau_k)_{k>\ell_\diamond}$  and the monotonicity of  $(\gamma_k)_k$ , we obtain  $\tau_k \rightarrow 0$  and

$$\frac{\tau_{k+1}}{\tau_k} = \max \left\{ \vartheta_k + \gamma_k^{\frac{k-\bar{\ell}}{2}}, \gamma_{k+1}^{\frac{1}{2}} \left[ \frac{\gamma_{k+1}}{\gamma_k} \right]^{(k-\bar{\ell})/2} \right\} \leq \max \left\{ \vartheta_k + \gamma_k^{\frac{k-\bar{\ell}}{2}}, \gamma_{k+1}^{\frac{1}{2}} \right\} \rightarrow 0,$$

which concludes the proof of [Theorem 4.6](#).  $\square$

Neglecting the dependence on  $\delta_k$  and on  $\varepsilon_k^1, \varepsilon_k^2$  for a moment, the results in [Theorem 4.6](#) can be summarized as follows. In order to guarantee r-linear convergence of the sequence  $(x^k)_k$ , it suffices to increase the sample size  $\mathbf{n}_k^g$  at a geometric rate and to choose  $\mathbf{n}_k^h$  sufficiently large. If  $\mathbf{n}_k^g$  is increased at a rate that is faster than geometric and we have  $\mathbf{n}_k^h \rightarrow \infty$ , then we obtain r-superlinear convergence with high probability. Similar results were established in [\[13, 72, 97\]](#) for stochastic Newton-type methods for smooth optimization problems. Clearly, the rate of convergence in [Theorem 4.6](#) (ii) can be further improved if  $\gamma_f$  is adjusted appropriately. Moreover, if the full gradient is used in the algorithm eventually, i.e., we have  $G_{s^k}(x^k) \equiv \nabla f(x^k)$  for all  $k$  sufficiently large, then the gradient related error terms  $\mathcal{E}_k^g$  vanish and we can derive q-linear and q-superlinear convergence, respectively. We continue with an additional remark.

**REMARK 4.7.** *Similar to [Remark 3.13](#), let us again assume that the function  $r$  has the special form  $r = \iota_C + \varphi$ , where  $\varphi$  is a real-valued, convex mapping and  $\iota_C$  is the indicator function of a closed, convex, and nonempty set  $C$ . In this case, as has already been discussed in [section 2](#), we can perform a projected Newton step  $z^k = \mathcal{P}_{\text{dom } r}(x^k + d^k) = \mathcal{P}_C(x^k + d^k)$  to obtain a feasible trial point  $z^k \in \text{dom } r$ . Due to the nonexpansiveness of the projection, this additional operation does also not affect our local convergence results. Moreover, since  $\varphi$  is locally Lipschitz continuous, assumption (D.4) is no longer needed in this situation.*

Based on the “light tail” assumption (E.3), we now present a straight-forward variant of [Theorem 4.6](#) with an improved dependence on the probabilities  $\delta_k$ .

**COROLLARY 4.8.** *Consider the setup discussed in [Theorem 4.6](#) and let us assume that the conditions in [Theorem 4.6](#) are fulfilled. Suppose that assumption (E.3) is satisfied. Then, the statements in [Theorem 4.6](#) (i) hold under the following improved sample size bounds*

$$(4.18) \quad \mathbf{n}_k^g \geq \left[ \left( 1 + \sqrt{3 \log(\delta_k^{-1})} \right) \frac{2\bar{\sigma}}{\lambda_m \Gamma_k} \right]^2, \quad \mathbf{n}_k^h \geq 3 \log(2n\delta_k^{-1}) \left[ \frac{2\bar{\rho}}{\lambda_m \gamma_f} \right]^2, \quad k \geq \bar{\ell}.$$

Furthermore, if (4.18) holds with  $\Gamma_k \equiv \Gamma_k^\circ$  or with  $\Gamma_k \equiv \Gamma_k^\diamond$  and  $\mathbf{n}_k^h \geq \log(2n\delta_k^{-1})\rho_k^{-1}$ , then the statements in [Theorem 4.6](#) (ii) and (iii) are satisfied, respectively.

[Corollary 4.8](#) can be shown directly by applying [Lemma 4.3](#). Finally, let us note that our local results can be further improved in the situation that was considered in [Theorem 3.12](#). We conclude this section with an example and discuss a specific choice of the parameters and sample sizes satisfying the assumptions in [Theorem 4.6](#) and [Corollary 4.8](#).

**EXAMPLE 4.9.** *Suppose that the conditions in [Corollary 4.8](#) are satisfied and let  $C_1, C_2 > 0$ , and  $\varpi > 0$  be given. Let the parameter sequences  $(\varepsilon_k^1)_k, (\varepsilon_k^2)_k$  be chosen via*

$$\varepsilon_1^k = C_1 k^{-(2+\frac{\varpi}{4})}, \quad \varepsilon_2^k = C_2 k^{-(1+\frac{\varpi}{8})}, \quad \forall k \in \mathbb{N},$$



and let us set  $p = \frac{1}{2}$  and  $\delta_k = \frac{1}{2k^8}$ . Then, setting  $\mathbf{n}_k^g = k^{4+\varpi} \log(k)$  and  $\mathbf{n}_k^h = \log(k)^{1+\varpi}$ , it can be shown that the statements in [Theorem 4.6](#) (i) hold with probability

$$\delta_* \geq \left[ \prod_{k=2}^{\infty} \left( 1 - \frac{1}{k^8} \right) \right]^2 = \left[ \frac{\sinh(\pi)(\cosh(\pi\sqrt{2}) - \cos(\pi\sqrt{2}))}{16\pi^3} \right]^2 \geq 0.99,$$

see [\[87\]](#). Additionally, if the gradient sample size increases geometrically, i.e., if we have  $\mathbf{n}_k^g = C_3 \ell^k$  for some  $\ell > 1$ ,  $C_3 > 0$ , the conditions in [Theorem 4.6](#) (ii) are satisfied and we can guarantee  $r$ -linear convergence with rate  $\max\{\frac{1}{2}, \frac{1}{\ell}\}$  with probability 99%.

**5. Numerical Results.** In this section, we demonstrate the efficiency of the proposed stochastic semismooth Newton framework and compare it with several state-of-the-art algorithms on a variety of test problems. All numerical experiments are performed using MATLAB R2017b on a desktop computer with Intel(R) Core(TM) i7-7700T 2.90GHz and 8GB memory.

**5.1. Logistic Regression.** In our first experiment, we consider the well-known empirical  $\ell_1$ -logistic regression problem

$$(5.1) \quad \min_{x \in \mathbb{R}^n} \psi(x) := f(x) + \mu \|x\|_1, \quad f(x) := \frac{1}{N} \sum_{i=1}^N f_i(x),$$

where  $f_i(x) := \log(1 + \exp(-b_i \cdot \langle a_i, x \rangle))$  denotes the logistic loss function and the data pairs  $(a_i, b_i) \in \mathbb{R}^n \times \{-1, 1\}$ ,  $i \in [N]$ , correspond to a given dataset or are drawn from a given distribution. The regularization parameter  $\mu > 0$  controls the level of sparsity of a solution of problem [\(5.1\)](#). In our numerical tests, we always choose  $\mu = 0.01$ .

**5.1.1. Algorithmic details and implementation.** Next, we describe the implementation details of our method and of the state-of-the-art algorithms used in our numerical comparison.

*Stochastic oracles.* In each iteration and similar to other stochastic second order methods, [\[16, 13, 18, 71\]](#), we generate stochastic approximations of the gradient and Hessian of  $f$  via first selecting two sub-samples  $\mathcal{S}_k, \mathcal{T}_k \subset [N]$  uniformly at random and without replacement from the index set  $\{1, \dots, N\}$ . We then define the following mini-batch-type stochastic oracles

$$(5.2) \quad G_{\mathcal{S}_k}(x) := \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} \nabla f_i(x), \quad H_{\mathcal{T}_k}(x) := \frac{1}{|\mathcal{T}_k|} \sum_{j \in \mathcal{T}_k} \nabla^2 f_j(x).$$

We refer to the variant of Algorithm 1 using the stochastic oracles [\(5.2\)](#) as S4N (sub-sampled semismooth Newton method). Furthermore, motivated by the recent success of variance reduction techniques [\[40, 90, 66, 5, 86\]](#), we will also work with a variance reduced stochastic gradient that can be calculated as follows

$$(5.3) \quad \begin{cases} 1 & \text{if } k \bmod m = 0 \text{ then set } \tilde{x} := x^k \text{ and calculate } \tilde{u} := \nabla f(\tilde{x}). \\ 2 & \text{Compute } G_{\mathcal{S}_k}(x^k) := \frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} (\nabla f_i(x^k) - \nabla f_i(\tilde{x})) + \tilde{u}. \end{cases}$$

Here,  $k \in \mathbb{N}$  is the current iteration and  $m \in \mathbb{N}$  denotes the number of iterations after which the full gradient  $\nabla f$  is evaluated at the auxiliary variable  $\tilde{x}$ . As in [\[66, 5, 86\]](#), this additional noise-free information is stored and utilized in the computation of the stochastic oracles for the following iterations.

*Overview of the tested methods.*

- **Adagrad**, [\[24\]](#). Adagrad is a stochastic proximal gradient method with a specific strategy for choosing the matrices  $\Lambda_k$ . We use the mini-batch gradient [\(5.2\)](#) as first order oracle in our implementation. This leads to the following update rule

$$(5.4) \quad x^{k+1} = \text{prox}_{\varphi}^{\Lambda_k}(x^k - \Lambda_k^{-1} G_{s^k}(x^k)), \quad \Lambda_k := \lambda^{-1} \text{diag}(\delta \mathbf{1} + \sqrt{G_k}),$$

where  $\delta, \lambda > 0$ ,  $G_k := G_{k-1} + G_{s^k}(x^k) \odot G_{s^k}(x^k)$ , and the multiplication “ $\odot$ ” and the square root “ $\sqrt{\cdot}$ ” are performed component-wise.

- **prox-SVRG**, [90]. Prox-SVRG is a variance reduced, stochastic proximal gradient method. Similar to [66, 67, 85], we substitute the basic variance reduction technique proposed in [40, 90] with the mini-batch version (5.3) to improve the performance of prox-SVRG.
- **S2N-D**. S2N-D is the deterministic version of the stochastic semismooth Newton method using the full gradient and Hessian of  $f$  instead of stochastic oracles.
- **S4N-HG**. S4N with both sub-sampled gradient and Hessian (5.2). In the numerical experiments, the maximum sample size  $|\mathcal{S}_k|$  of the stochastic oracle  $G_{s^k}$  is limited to 10%, 50% and 100% of the training data size  $N$ , respectively.
- **S4N-H**. This version of S4N uses the full gradient  $\nabla f$  and the sub-sampled Hessian  $H_{t^k}$  as defined in (5.2).
- **S4N-VR**. S4N-VR is a variant of S4N combining the variance reduced stochastic oracle (5.3) with the basic sub-sampling strategy (5.2) for the Hessian of  $f$ .

Adagrad and prox-SVRG are two popular and efficient first order stochastic optimization approaches for solving nonsmooth and possibly nonconvex problems of the form (1.1). We compare them with four different versions of our S4N method: S2N-D (deterministic), S4N-HG (sub-sampled gradient and Hessian), S4N-H (full gradient and sub-sampled Hessian), and S4N-VR (variance reduced stochastic gradient and sub-sampled Hessian).

*Implementational details.* For Adagrad, the sample size  $|\mathcal{S}_k|$  of the stochastic gradient is fixed to 5% of the training data size  $N$  and we set  $\delta = 10^{-7}$ . The parameter  $\lambda$  varies for the different tested datasets and is chosen from the set  $\{i \cdot 10^j : i \in [9], j \in \{-2, -1, 0, 1\}\}$  to guarantee optimal performance. The iterative scheme of prox-SVRG basically coincides with (5.4). Here, we also use a fixed sample size  $|\mathcal{S}_k| = \lfloor 0.01N \rfloor$  and we set  $m = 10$ . The parameter matrix  $\Lambda_k$  is defined via  $\Lambda_k := (1/\lambda_k)I$  and based on the full gradient values  $\nabla f(\tilde{x})$ ,  $\lambda_k$  is chosen adaptively to approximate the Lipschitz constant of the gradient  $\nabla f$ .

In S4N-HG, the initial sample size of the stochastic gradient is set to  $|\mathcal{S}_0| = \lfloor 0.01N \rfloor$ . The size of the mini-batch  $\mathcal{S}_k$  is then increased by a factor of 3.375 every 30 iterations until  $|\mathcal{S}_k|$  reaches the maximum sizes  $\lfloor 0.1N \rfloor$ ,  $\lfloor 0.5N \rfloor$ , and  $N$ , respectively. In the following, we will use S4N-HG 10%, S4N-HG 50%, and S4N-HG 100% to denote the different variants of S4N-HG. In S4N-VR, we use the fixed sample size  $|\mathcal{S}_k| = \lfloor 0.01N \rfloor$  for all  $k$  and  $m = 6$ . The mini-batch sizes of the stochastic Hessians are adjusted in a similar way. More specifically, in S4N-HG, S4N-H, and S4N-VR, we first set  $|\mathcal{T}_0| = \lfloor 0.01N \rfloor$ . As soon as the sample  $\mathcal{S}_k$  reaches its maximum size, we repeatedly increase the size of the set  $\mathcal{T}_k$  by a factor of 3.375 after 15 iterations. The upper limit of the Hessian sample size is set to 10% of the training data size, i.e., we have

$$|\mathcal{T}_k| \leq t_{\max}, \quad t_{\max} := \lfloor 0.1N \rfloor, \quad \forall k.$$

In S4N-HG 10% and different from the other methods, the size of  $\mathcal{T}_k$  is not changed, i.e., it holds  $t_{\max} = \lfloor 0.01N \rfloor$ . As in prox-SVRG, we use  $\Lambda_k := (1/\lambda_k)I$  and choose  $\lambda_k$  adaptively to estimate the Lipschitz constant of the gradient. In particular, we compute

$$\lambda_k^1 = \frac{\|x^k - x^{k-1}\|}{\|G_{s^k}(x^k) - G_{s^{k-1}}(x^{k-1})\|}, \quad \lambda_k^2 = \max\{10^{-3}, \min\{10^4, \lambda_k^1\}\}.$$

In order to prevent outliers, we calculate a weighted mean of  $\lambda_k^2$  and of the previous parameters  $\lambda_j$ ,  $j \in [k-1]$ . This mean is then used as the new step size parameter  $\lambda_k$ . The initial

Data Set	Data Points $N$	Variables $n$	Density	Reference
CINA	16033	132	29.56%	[89]
gisette	6000	5000	12.97%	[36]
MNIST	60000	784	19.12%	[43]
rcv1	20242	47236	0.16%	[45]

Table 1: A description of binary datasets used in the experiments

step size is set to  $\lambda_0 = 0.1$ .

The proximity operator of the  $\ell_1$ -norm has the explicit representation  $\text{prox}_{\mu\|\cdot\|_1}^{\Lambda_k}(u) = u - \mathcal{P}_{[-\mu\lambda_k, \mu\lambda_k]}(u)$  and is also known as the shrinkage operator or soft-thresholding function. Similar to [52, 61, 91], we will work with the following generalized Jacobian of  $\text{prox}_{\mu\|\cdot\|_1}^{\Lambda_k}$  at some  $u \in \mathbb{R}^n$

$$D(u) := \text{diag}(d(u)), \quad d(u) \in \mathbb{R}^n, \quad d(u)_i := \begin{cases} 1 & |u_i| > \mu\lambda_k, \\ 0 & \text{otherwise.} \end{cases}$$

The generalized derivatives of  $F_{s^k}^{\Lambda_k}$  are then built as in (2.7). As described in [52, 61, 91], we can exploit the structure of the resulting semismooth Newton system and reduce it to a smaller and symmetric linear system of equations. We utilize an early terminated conjugate gradient (CG) method to solve this system approximately. The maximum number of iterations and the desired accuracy of the CG method are adjusted adaptively depending on the computed residual  $\|F_{s^k}^{\Lambda_k}(x^k)\|$ . When the residual is large, only few iterations are performed to save time. The initial relative tolerance and the initial maximum number of iterations are set to 0.01 and 2, respectively. The total maximum number of CG-iterations is restricted to 12. In order to numerically robustify the computation of the Newton step  $z_n^k$ , we also consider the following, regularized version of the Newton system

$$(M_k + \rho_k I) \cdot d^k = -F_{s^k}^{\Lambda_k}(x^k), \quad M_k \in \mathcal{M}_{s^k, t^k}^{\Lambda_k}(x^k),$$

where  $\rho_k > 0$  is a small positive number. We adjust  $\rho_k$  according to the norm of the residual  $F_{s^k}^{\Lambda_k}(x^k)$  so that  $\rho_k \rightarrow 0$  as  $\|F_{s^k}^{\Lambda_k}(x^k)\| \rightarrow 0$ .

Finally, in our implementation of S4N, we only check the first growth condition (2.8) to measure the quality of the Newton step  $z_n^k$ . Although both growth conditions (2.8) and (2.9) are generally required to guarantee global convergence, this adjustment does not affect the globalization process and convergence of S4N in the numerical experiments. Moreover, as we have shown in Theorem 3.12 and in Theorem 4.6, the condition (2.9) is actually not necessary for strongly convex problems and it is satisfied locally close to a stationary point of problem (1.1) under certain assumptions. These different observations motivate us to restrict the acceptance test of the semismooth Newton steps to the cheaper condition (2.8). We use the following parameters  $\eta = 0.85$ ,  $\nu_k = \varepsilon_k^1 = c_\nu k^{-1.1}$ ,  $c_\nu = 500$ , and  $\alpha_k = 10^{-2}$ .

**5.1.2. Numerical comparison.** The datasets tested in our numerical comparison are summarized in Table 1. We linearly scale the entries of the data-matrix  $(a_1, \dots, a_N)$  to  $[0, 1]$  for each dataset. The datasets for multi-class classification have been manually divided into two types or features. For instance, the MNIST dataset is used for classifying even and odd digits. For all methods, we choose  $x^0 = 0$  as initial point.

In Figure 1 and Figure 2, we show the performances of all methods for solving the logistic regression problem (5.1). The change of the *relative error*  $(\psi(x) - \psi(x^*)) / \max\{1, |\psi(x^*)|\}$

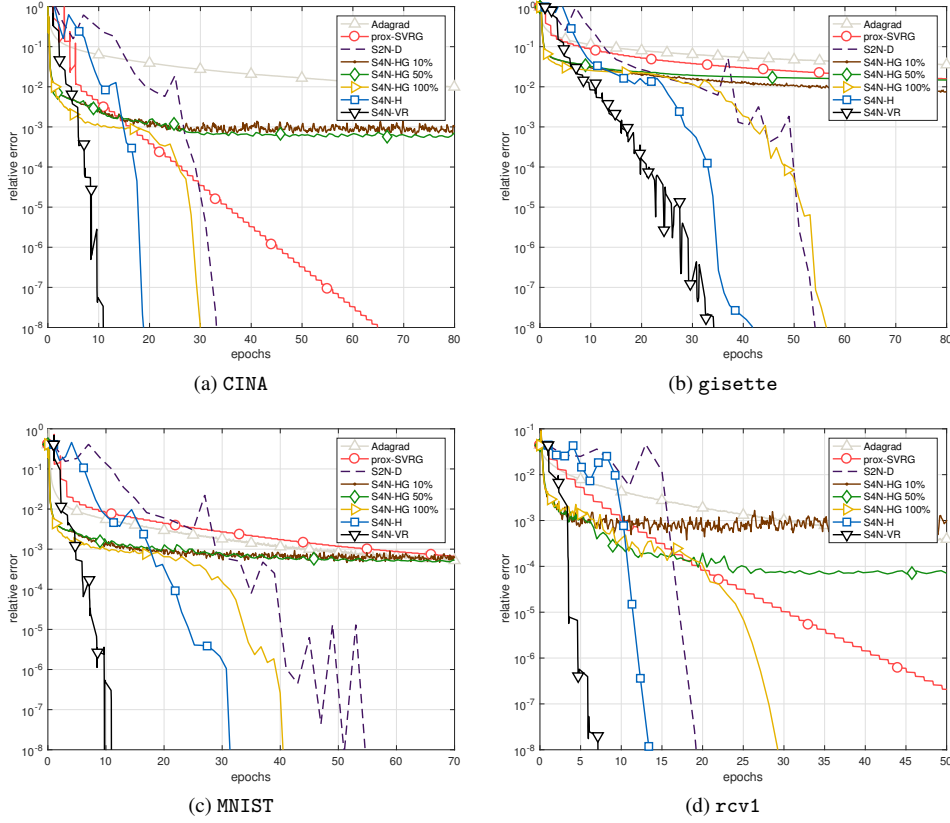


Fig. 1: Change of the relative error with respect to the required epochs for solving the  $\ell_1$ -logistic regression problem (5.1). (Averaged over 50 independent runs).

is reported with respect to *epochs* and *cpu-time*, respectively. Here,  $x^*$  is a reference solution of problem (5.1) generated by S2N-D with stopping criterion  $\|F^I(x)\| \leq 10^{-12}$ . Moreover, one epoch denotes a full pass over a dataset. The results presented in Figure 1 and Figure 2 are averaged over 50 independent runs.

At first, we observe that S2N-D, S4N-HG 100%, S4N-H, and S4N-VR outperform the first order method Adagrad both with respect to the required number of epochs and *cpu-time*. Furthermore, the different variants S4N and S2N-D seem to be especially well-suited for recovering high accuracy solutions.

The deterministic semismooth Newton method S2N-D decreases slowly in the early stage of the iteration process, but converges rapidly when the iterates are close to an optimal solution. The results show that in the early stage the performance of S2N-D is inferior to the performance of the other stochastic methods. If a higher precision is required, then S2N-D becomes more efficient and behaves similar to S4N-HG 100%. Overall, S2N-D is not competitive with the stochastic variants S4N-H and S4N-VR and converges slower. These observations indicate the strength of stochastic algorithms in general.

Our numerical experiments show that the different performances of the stochastic methods can be roughly split into two categories. The first category includes Adagrad, S4N-HG 10%, and S4N-HG 50%, while the second category consists of S4N-H, S4N-HG 100%, and

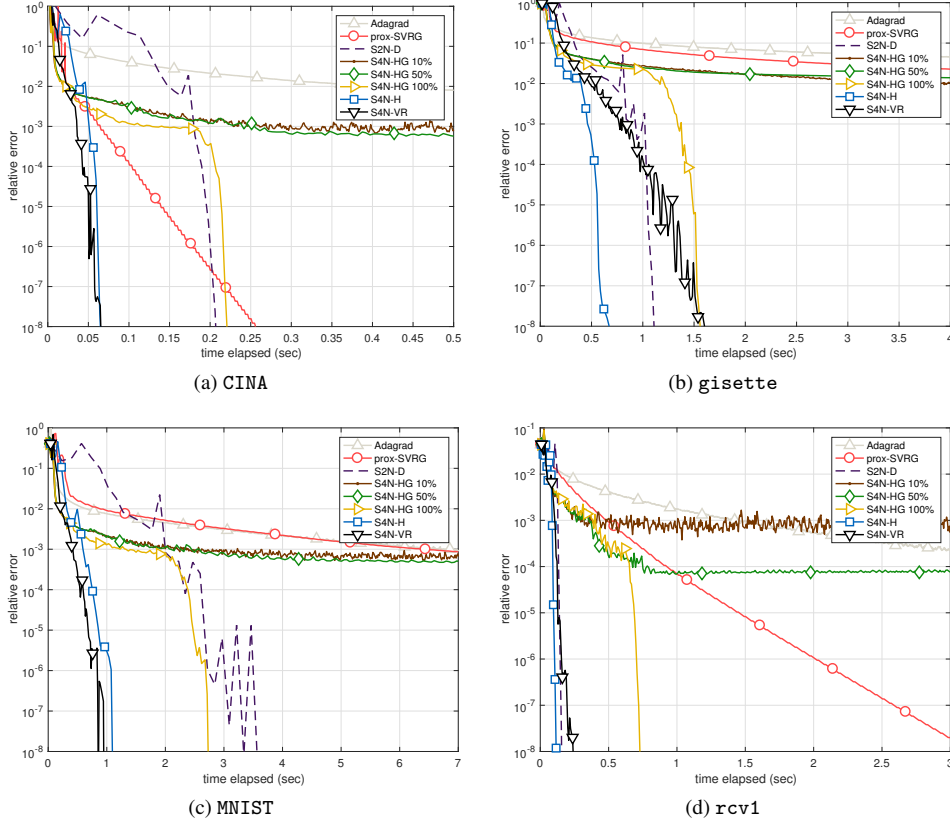


Fig. 2: Change of the relative error with respect to the cpu-time for solving the  $\ell_1$ -logistic regression problem (5.1). (Averaged over 50 independent runs).

S4N-VR. The performance of prox-SVRG depends on the tested datasets. While in *gisette* and *MNIST*, it converges slowly and performs similarly to Adagrad, prox-SVRG shows much faster convergence on the datasets *CINA* and *rcv1*. Comparing the results in Figure 1, it appears that the performance of S4N-HG 10% and S4N-HG 50% is comparable to the one of the first order method Adagrad. Since the maximum sample set size of the stochastic gradient in S4N-HG 10% and S4N-HG 50% is limited to  $[0.1N]$  and  $[0.5N]$ , the associated gradient error terms still might be too large, preventing transition to fast local convergence and causing stagnation of the methods. Thus and similar to the observations for stochastic quasi-Newton methods [18, 86], the performance of S4N is greatly affected by the sampling strategy and the accuracy of the gradient approximation. This is also partly illustrated by the acceptance of the growth condition (2.8). While in S2N-D, S4N-HG 100%, S4N-H, and S4N-VR usually every semismooth Newton step is accepted as a new iterate, a small number of Newton steps is rejected in S4N-HG 10% and S4N-HG 50%. This situation typically occurs in the *rcv1* dataset, when either S4N-HG 10% or S4N-HG 50% stagnates and the stochastic Newton step does not provide sufficient progress to be accepted. The results in Figure 1 and Figure 2 demonstrate that the performance of S4N can be further improved by increasing the sample size of the gradient gradually to its full size, as in S4N-HG 100%, or by introducing an addi-

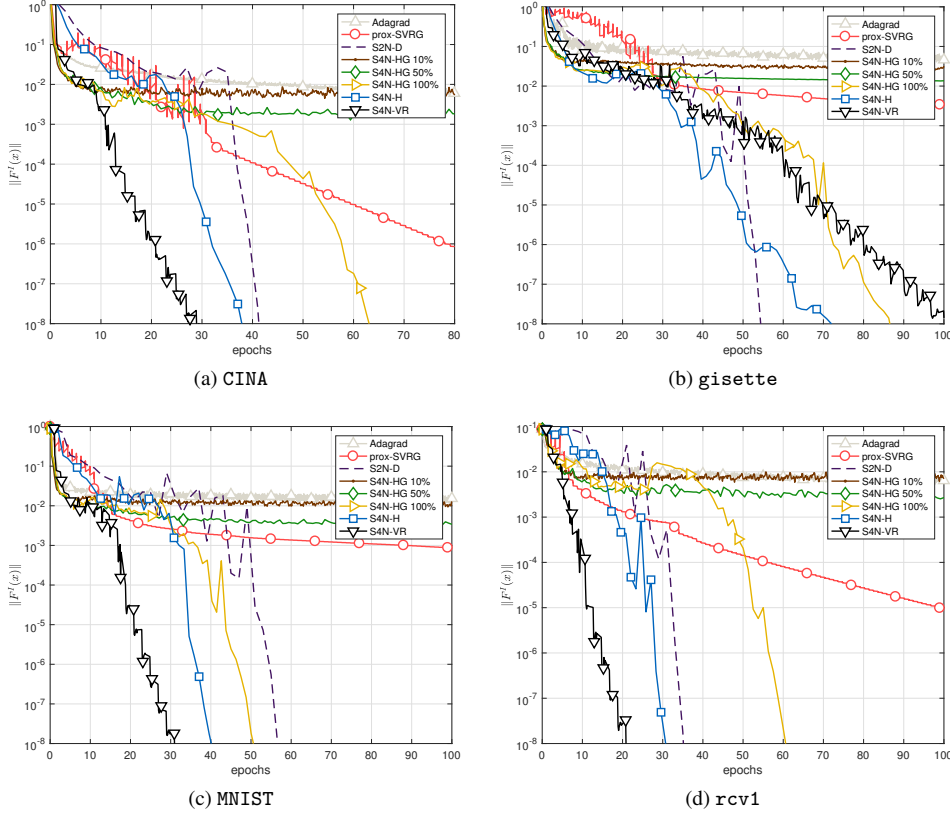


Fig. 3: Change of the residual  $\|F^I(x)\|$  with respect to epochs for solving the nonconvex binary classification problem (5.5). (Averaged over 50 independent runs).

tional variance reduction technique as in S4N-VR. We also observe that S4N-VR outperforms most of the other methods (especially with respect to number of required epoch evaluations) which indicates that the combination of second order information and variance reduction is advantageous and very promising.

**5.2. Nonconvex Binary Classification.** In this subsection, we consider the following nonconvex, binary classification problem [49, 86]

$$(5.5) \quad \min_{x \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^N [1 - \tanh(b_i \cdot \langle a_i, x \rangle)] + \mu \|x\|_1,$$

where  $f_i(x) := 1 - \tanh(b_i \langle a_i, x \rangle)$  is the sigmoid loss function and  $\mu = 0.01$  is a regularization parameter. We test the same datasets as in subsection 5.1 to evaluate the performance of the different versions of S4N.

The sampling strategy and parameters are adjusted as follows. For all S4N methods, the initial mini-batch size of the stochastic gradient is increased to  $|\mathcal{S}_0| = \lfloor 0.05N \rfloor$ . For S4N-HG 10%, S4N-HG 50%, and S4N-VR, we set  $|\mathcal{T}_0| = \lfloor 0.025N \rfloor$ . The other variants of S4N start with the initial sample size  $|\mathcal{T}_0| = \lfloor 0.05N \rfloor$ . We set  $t_{\max} = \lfloor 0.25N \rfloor$ , except for S4N-HG 10% where  $t_{\max} = \lfloor 0.05N \rfloor$  is used. We utilize the minimal residual method (MINRES) to



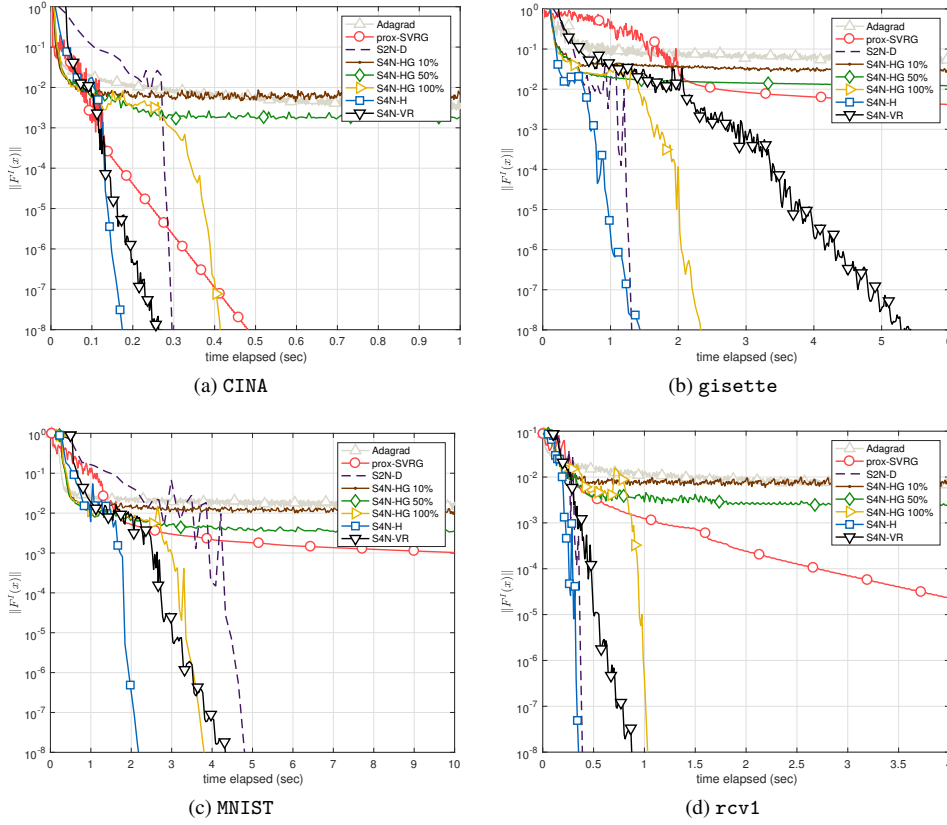


Fig. 4: Change of the residual  $\|F^I(x)\|$  with respect to cpu-time for solving the nonconvex binary classification problem (5.5). (Averaged over 50 independent runs).

solve the reduced Newton system and the maximum number of MINRES-iterations is set to 32. We choose  $c_\nu = 2500$  and in S4N-VR, the parameter  $m$  is adjusted to  $m = 8$ . For the *gisette* dataset, we changed the initial value for  $\lambda_0$  to 5, which significantly improved the performance of the S4N methods. All remaining parameters and strategies follow the setup discussed in the last subsection.

The numerical results are presented in Figure 3 and Figure 4. We report the change of the residual  $\|F^I(x)\|$  with respect to the required epochs and cpu-time. In general, the overall performance of the different methods is similar to the results shown in the last subsection. However, in contrast to the convex logistic regression problem, the more accurate approximation of the Hessian seems to be beneficial and can accelerate convergence. This observation is also supported by the slightly improved performance of the deterministic semismooth Newton method S2N-D. Our results show that prox-SVRG now consistently outperforms S4N-HG 10% and S4N-HG 50% in recovering high precision solutions. Similar to the convex example, S4N-HG 100% manages to significantly reduce the residual  $\|F^I(x)\|$  in the first iterations. As soon as the stochastic error in the gradient approximation becomes negligible, the behavior of S4N-HG 100% changes and fast local converges can be observed. The methods S4N-H and S4N-VR still compare favorably with the other stochastic approaches. With respect to the epochs, S4N-VR again outperforms the other methods (except for the dataset *gisette*). Re-

garding the required cpu-time, S4N-H achieves good results and converges quickly to highly accurate solutions.

**6. Conclusion.** In this paper, we investigate a stochastic semismooth Newton method for solving nonsmooth and nonconvex minimization problems. In the proposed framework, the gradient and Hessian of the smooth part of the objective function are approximated by general stochastic first and second order oracles. This allows the application of various sub-sampling and variance reduction techniques or other stochastic approximation schemes. The method is based on stochastic semismooth Newton steps, stochastic proximal gradient steps, and growth conditions and a detailed discussion of the global convergence properties is provided. Under suitable assumptions, transition to fast local convergence is established. More specifically, we show that the method converges locally with high probability with an  $r$ -linear or  $r$ -superlinear rate if the sample sizes of the stochastic gradient and Hessian are increased sufficiently fast. The approach is tested on an  $\ell_1$ -logistic regression and a nonconvex binary classification problem on a variety of datasets. The numerical comparisons indicate that our algorithmic framework and especially the combination of (generalized) second order information and variance reduction are promising and competitive.

## Appendix A. Proofs.

**A.1. Proof of Fact 3.4.** We start with some preparatory definitions. Let  $(\Omega, \mathcal{F})$  and  $(\Xi, \mathcal{X})$  be measurable spaces, then  $\mathcal{F} \otimes \mathcal{X}$  denotes the usual product  $\sigma$ -algebra of the product space  $\Omega \times \Xi$ . We use  $\mathcal{B}(\mathbb{R}^n)$  to denote the Borel  $\sigma$ -algebra of  $\mathbb{R}^n$ . Let  $\Gamma : \Omega \rightrightarrows \mathbb{R}^n$  be a multifunction. Then, for some set  $C \subset \mathbb{R}^n$ , we define  $\Gamma^{-1}(C) := \{\omega \in \Omega : \Gamma(\omega) \cap C \neq \emptyset\}$ . Following [2], the multifunction  $\Gamma$  is called *weakly measurable* if for all *open* sets  $C \subset \mathbb{R}^n$ , the set  $\Gamma^{-1}(C)$  is measurable. The function  $\Gamma$  is called *measurable* if  $\Gamma^{-1}(C)$  is measurable for all *closed* sets  $C \subset \mathbb{R}^n$ . If  $\Gamma$  is closed-valued then these two concepts coincide and are actually equivalent, see [70]. The *graph* of  $\Gamma$  is given by  $\text{gra } \Gamma := \{(\omega, x) \in \Omega \times \mathbb{R}^n : x \in \Gamma(\omega)\}$  and the *closure*  $\bar{\Gamma}$  of  $\Gamma$  is defined via  $\bar{\Gamma}(\omega) := \text{cl } \Gamma(\omega)$  for all  $\omega \in \Omega$ . The next result is straightforward.

**LEMMA A.1.** *Let  $(\Omega, \mathcal{F})$  be a measurable space and let  $S : \Omega \rightrightarrows \mathbb{R}^n$  be a multifunction. Then, the indicator function  $\iota : \Omega \times \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $\iota(\omega, x) := \mathbb{1}_{S(\omega)}(x)$  is jointly  $\mathcal{F} \otimes \mathcal{B}(\mathbb{R}^n)$ -measurable if and only if  $\text{gra } S \in \mathcal{F} \otimes \mathcal{B}(\mathbb{R}^n)$ .*

**LEMMA A.2.** *Let  $(\Omega, \mathcal{F})$  be a measurable space and let  $f : \Omega \times \mathbb{R}^n \rightarrow \mathbb{R}$  be a given Carathéodory function. Moreover, suppose that the mapping  $\varphi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$  is convex, lower semicontinuous, and proper and let us consider the multifunction*

$$S_\gamma : \Omega \rightrightarrows \mathbb{R}^n, \quad S_\gamma(\omega) := \{x \in \mathbb{R}^n : \varphi(x) + f(\omega, x) \leq \gamma\}$$

for some  $\gamma \in \mathbb{R}$ . Then,  $S_\gamma$  has measurable graph, i.e., it holds  $\text{gra } S_\gamma \in \mathcal{F} \otimes \mathcal{B}(\mathbb{R}^n)$ .

*Proof.* Since  $\varphi$  is convex, lower semicontinuous, and proper, it can be written as the supremum of its affine minorants. In particular, by [9, Corollary 13.36 and Proposition 7.11] there exists a countable, dense subset  $\mathcal{D} := \{(y_1, \tau_1), (y_2, \tau_2), \dots\}$  of  $\text{epi } \varphi^*$  such that

$$\varphi(x) = \varphi^{**}(x) = \sup_{(y, \tau) \in \text{epi } \varphi^*} \langle y, x \rangle - \tau = \sup_{k \in \mathbb{N}} \langle y_k, x \rangle - \tau_k, \quad \forall x \in \mathbb{R}^n.$$

(Here, we also used the closedness of the epigraph  $\text{epi } \varphi^*$ ). Next, let us define the family of multifunctions  $S_{\gamma, k} : \Omega \rightrightarrows \mathbb{R}^n$ ,

$$S_{\gamma, k}(\omega) := \{x \in \mathbb{R}^n : \langle y_k, x \rangle - \tau_k + f(\omega, x) < \gamma\}, \quad k \in \mathbb{N}.$$

Since the mapping  $\tilde{f}_k(\omega, x) := \langle y_k, x \rangle - \tau_k + f(\omega, x)$  is obviously a Carathéodory function, it follows from [2, Lemma 18.7] that each multifunction  $S_{\gamma,k}$  is (weakly) measurable. Moreover, [2, Theorem 18.6] implies that the closures  $\bar{S}_{\gamma,k}$  have measurable graph. Thus, due to  $\text{gra } S_\gamma = \bigcup_{k=1}^\infty \text{gra } \bar{S}_{\gamma,k}$ , we can infer  $\text{gra } S_\gamma \in \mathcal{F} \otimes \mathcal{B}(\mathbb{R}^n)$ .  $\square$

LEMMA A.3. *Let  $(\Omega, \mathcal{F})$  be a measurable space and let  $T : \Omega \rightarrow \mathbb{R}^{n \times n}$  and  $y : \Omega \rightarrow \mathbb{R}^n$  be measurable functions. Then, the mapping  $\omega \mapsto \zeta(\omega) := T(\omega)^+ y(\omega)$  is measurable.*

*Proof.* By [32], we have  $\lim_{\lambda \rightarrow 0} (A^\top A + \lambda I)^{-1} A^\top = A^+$  for any matrix  $A \in \mathbb{R}^{n \times n}$ . Now, let  $(\lambda_k)_k \subset \mathbb{R}$  be an arbitrary sequence converging to zero. Then, a continuity argument implies that the mapping  $\omega \mapsto \zeta_k(\omega) := (T(\omega)^\top T(\omega) + \lambda_k I)^{-1} T(\omega)^\top y(\omega)$  is  $\mathcal{F}$ -measurable for all  $k \in \mathbb{N}$ . Since  $\zeta$  is the point-wise limit of the sequence  $(\zeta_k)_k$ , this finishes the proof  $\square$

We now turn to the proof of Fact 3.4.

*Proof.* Since the stochastic oracles  $\mathcal{G}$  and  $\mathcal{H}$  are Carathéodory functions, it follows that  $\mathcal{G}$  and  $\mathcal{H}$  are jointly measurable, i.e., it holds

$$(A.1) \quad \mathcal{G} \in \mathcal{B}(\mathbb{R}^n) \otimes \mathcal{X}, \quad \mathcal{H} \in \mathcal{B}(\mathbb{R}^n) \otimes \mathcal{X},$$

see, e.g., [7, Lemma 8.2.6] or [2, Section 4.10]. We now prove the slightly extended claim

$$(A.2) \quad z_p^k, z_n^k \in \mathcal{F}_k, \quad Y_{k+1}, \theta_{k+1} \in \hat{\mathcal{F}}_k, \quad \text{and} \quad x^{k+1} \in \hat{\mathcal{F}}_k$$

inductively. Let us suppose that the statement (A.2) holds for  $k-1, k \in \mathbb{N}$ . Then, due to (A.1) and  $x^k \in \hat{\mathcal{F}}_{k-1} \subset \mathcal{F}_k$  and using the  $(\mathcal{F}_k, \mathcal{B}(\mathbb{R}^n) \otimes \mathcal{X})$ -measurability of the functions  $\xi_i^k : \Omega \rightarrow \mathbb{R}^n \times \Xi$ ,  $\xi_i^k(\omega) := (x^k(\omega), s_i^k(\omega))$  and  $\tau_j^k : \Omega \rightarrow \mathbb{R}^n \times \Xi$ ,  $\tau_j^k(\omega) := (x^k(\omega), t_j^k(\omega))$  for  $i = 1, \dots, n_g^k$  and  $j = 1, \dots, n_h^k$ , it follows

$$G_{s^k}(x^k) = \frac{1}{n_g^k} \sum_{i=1}^{n_g^k} \mathcal{G}(\xi_i^k) \in \mathcal{F}_k \quad \text{and} \quad H_{t^k}(x^k) = \frac{1}{n_h^k} \sum_{j=1}^{n_h^k} \mathcal{H}(\tau_j^k) \in \mathcal{F}_k.$$

Since the proximity operator  $\text{prox}_r^{\Lambda_k}$  is (Lipschitz) continuous, this also implies  $u_{s^k}^{\Lambda_k}(x^k) \in \mathcal{F}_k$ ,  $p_{s^k}^{\Lambda_k}(x^k) \in \mathcal{F}_k$ , and  $F_{s^k}^{\Lambda_k}(x^k) \in \mathcal{F}_k$  and thus, we have  $z_p^k \in \mathcal{F}_k$ . Moreover, by assumption (C.1) and Lemma A.3, we can infer  $z_n^k \in \mathcal{F}_k$ . Next, we study the measurability of the set  $S_k$  used in the definition of  $Y_{k+1}$ . Since lower semicontinuous functions are (Borel) measurable and we have  $\theta_k, x^k \in \hat{\mathcal{F}}_{k-1} \subset \mathcal{F}_k$ , the mapping  $p_k(\omega, a, b) := -\beta \theta_k(\omega)^{1-p} \|b\|^p - \psi(x^k(\omega))$  is a jointly  $\mathcal{F}_k \otimes \mathcal{B}(\mathbb{R}^n) \otimes \mathcal{B}(\mathbb{R}^n)$ -measurable Carathéodory function. Thus, by Lemma A.2, the multifunction  $\text{lev}_{\varepsilon_k^2} P_k : \Omega \rightrightarrows \mathbb{R}^n \times \mathbb{R}^n$ ,

$$(\text{lev}_{\varepsilon_k^2} P_k)(\omega) = \{(a, b) \in \mathbb{R}^n \times \mathbb{R}^n : \psi(a) + p_k(\omega, a, b) \leq \varepsilon_k^2\}$$

has measurable graph. Similarly, since  $q_k(\omega, a) := \|a\| - (\nu_k + \eta)\theta_k(\omega)$  is a Carathéodory function, we can argue that  $\text{lev}_{\varepsilon_k^1} Q_k : \Omega \rightrightarrows \mathbb{R}^n \times \mathbb{R}^n$  has measurable graph. (More precisely, since the set  $(\text{lev}_\gamma Q_k)(\omega)$  is compact-valued for all  $\gamma \in \mathbb{R}$  and  $\omega \in \Omega$ , we can even infer that  $\text{lev}_{\varepsilon_k^1} Q_k$  is measurable, see [2, Lemma 18.4]). This easily yields  $\text{gra } [\text{dom } r \times \text{lev}_{\varepsilon_k^1} Q_k] \in \mathcal{F}_k \otimes \mathcal{B}(\mathbb{R}^n) \otimes \mathcal{B}(\mathbb{R}^n)$  and hence, by Lemma A.1, the indicator function  $\mathbb{1}_{S_k}(\cdot, \cdot)$  is jointly  $\mathcal{F}_k \otimes \mathcal{B}(\mathbb{R}^n) \otimes \mathcal{B}(\mathbb{R}^n)$ -measurable. As before, due to  $\mathcal{F}_k \subset \hat{\mathcal{F}}_k$ , we have  $G_{s^{k+1}}(z_n^k) \in \hat{\mathcal{F}}_k$ ,

$$u_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k) \in \hat{\mathcal{F}}_k, \quad p_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k) \in \hat{\mathcal{F}}_k, \quad \text{and} \quad F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k) \in \hat{\mathcal{F}}_k,$$

which finally implies that the binary variable  $Y_{k+1}$  is  $\hat{\mathcal{F}}_k$ -measurable. Using the representation (3.7), it now follows  $x^{k+1} \in \hat{\mathcal{F}}_k$  and, due to  $\theta_{k+1} = Y_{k+1} \|F_{s^{k+1}}^{\Lambda_{k+1}}(z_n^k)\| + (1 - Y_{k+1})\theta_k$ ,

we also have  $\theta_{k+1} \in \hat{\mathcal{F}}_k$ . Since the constant random variables  $x^0$  and  $\theta_0$  are trivially  $\mathcal{F}_0$ -measurable, we can use the same argumentation for the base case  $k = 0$ . This finishes the proof of [Fact 3.4](#).  $\square$

We conclude with a remark on the existence of measurable selections of the multifunction  $\mathcal{M}_k$ . Due to [\[19\]](#), the generalized derivative  $\partial \text{prox}_r^{\Lambda_k} : \mathbb{R}^n \rightrightarrows \mathbb{R}^{n \times n}$  is an upper semi-continuous, compact-valued function for all  $k \in \mathbb{N}$ . Hence, by [\[83, Lemma 4.4\]](#),  $\partial \text{prox}_r^{\Lambda_k}$  is (Borel) measurable for all  $k$ . As we have shown inductively, the function  $\omega \mapsto u_{s^k(\omega)}^{\Lambda_k}(x^k(\omega))$  is  $\mathcal{F}_k$ -measurable and thus, by [\[83, Lemma 4.5\]](#), the multifunction

$$\mathcal{D}_k : \Omega \rightrightarrows \mathbb{R}^{n \times n}, \quad \mathcal{D}_k(\omega) := \partial \text{prox}_r^{\Lambda_k}(u_{s^k(\omega)}^{\Lambda_k}(x^k(\omega))),$$

is nonempty, closed-valued, and measurable with respect to the  $\sigma$ -algebra  $\mathcal{F}_k$  for all  $k$ . The Kuratowski-Ryll-Nardzewski Selection Theorem, [\[70, 7, 2\]](#), now implies that  $\mathcal{D}_k$  admits an  $\mathcal{F}_k$ -measurable selection  $D_k : \Omega \rightarrow \mathbb{R}^{n \times n}$ . Using  $H_{t^k}(x^k) \in \mathcal{F}_k$ , this also implies that  $M_k := I - D_k(I - \Lambda_k^{-1}H_{t^k}(x^k))$  is an  $\mathcal{F}_k$ -measurable selection of  $\mathcal{M}_k$ .

### A.2. Proof of [Lemma 3.7](#).

*Proof.* Since  $r$  is subdifferentiable at  $p_s^\Lambda(x)$  with  $\Lambda(u_s^\Lambda(x) - p_s^\Lambda(x)) \in \partial r(p_s^\Lambda(x))$  we have  $r'(p_s^\Lambda(x); h) \geq \langle \Lambda F_s^\Lambda(x) - G_s(x), h \rangle$  for all  $x, h \in \mathbb{R}^n$ , see, e.g., [\[9, Proposition 17.17\]](#). Now, using the convexity of  $f - \frac{\mu_f}{2} \|\cdot\|^2$ , the  $\mu_r$ -strong convexity of  $r$ , [\(3.12\)](#) with  $f \equiv 0$  and  $\mu_f = 0$ , and the descent lemma [\(3.3\)](#), it follows

$$\begin{aligned} \psi(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu_f}{2} \|y - x\|^2 + r(p_s^\Lambda(x)) \\ &\quad + \langle \Lambda F_s^\Lambda(x) - G_s(x), y - p_s^\Lambda(x) \rangle + \frac{\mu_r}{2} \|y - p_s^\Lambda(x)\|^2 \\ &\geq \psi(p_s^\Lambda(x)) + \langle \nabla f(x) - G_s(x), F_s^\Lambda(x) \rangle - \frac{L}{2} \|F_s^\Lambda(x)\|^2 + \frac{\mu_f}{2} \|y - x\|^2 \\ &\quad + \|F_s^\Lambda(x)\|_\Lambda^2 + \langle \Lambda F_s^\Lambda(x) + \nabla f(x) - G_s(x), y - x \rangle + \frac{\mu_r}{2} \|y - p_s^\Lambda(x)\|^2 \\ &= \psi(p_s^\Lambda(x)) + \langle \nabla f(x) - G_s(x), F_s^\Lambda(x) \rangle + \frac{1}{2}(\mu_r - L) \|F_s^\Lambda(x)\|^2 \\ &\quad + \|F_s^\Lambda(x)\|_\Lambda^2 + \langle (\Lambda + \mu_r I) F_s^\Lambda(x) + \nabla f(x) - G_s(x), y - x \rangle + \frac{\bar{\mu}}{2} \|y - x\|^2 \end{aligned}$$

for all  $x, y \in \mathbb{R}^n$ . Now, setting  $\mathcal{E}_s(x) := \|\nabla f(x) - G_s(x)\|$ ,  $b_2 = (\lambda_M + \mu_r)^2 \bar{\mu}^{-1}$ , and applying Young's inequality twice for some  $\tau, \alpha > 0$ , we get

$$\begin{aligned} &|\langle (\Lambda + \mu_r I) F_s^\Lambda(x) + \nabla f(x) - G_s(x), x^* - x \rangle| \\ &\leq \frac{1}{2}(1 + \alpha)b_2 \|F_s^\Lambda(x)\|^2 + \frac{(1 + \tau)(1 + \alpha)\bar{\mu}}{2(1 + \tau)(1 + \alpha)} \|x - x^*\|^2 + \frac{(1 + \tau)(1 + \alpha)}{2\tau\alpha\bar{\mu}} \cdot \mathcal{E}_s(x)^2, \end{aligned}$$

Setting  $y = x^*$  in our first derived inequality, recalling  $b_1 = L - 2\lambda_m - \mu_r$ , and using the optimality of  $x^*$ ,  $\Lambda \succeq \lambda_m I$ , Young's inequality, and the latter estimate, it holds

$$\begin{aligned} &\frac{\alpha\bar{\mu}}{2(1 + \tau)(1 + \alpha)} \|x - x^*\|^2 \\ &\leq \frac{1}{2}(L - \mu_r + (1 + \alpha)b_2) \|F_s^\Lambda(x)\|^2 - \|F_s^\Lambda(x)\|_\Lambda^2 \\ &\quad - \langle F_s^\Lambda(x), \nabla f(x) - G_s(x) \rangle + (1 + \tau)(1 + \alpha)(2\tau\alpha\bar{\mu})^{-1} \cdot \mathcal{E}_s(x)^2 \\ &\leq \frac{1}{2}(b_1 + \tau + (1 + \alpha)b_2) \|F_s^\Lambda(x)\|^2 + \frac{\alpha\bar{\mu} + (1 + \tau)(1 + \alpha)}{2\tau\alpha\bar{\mu}} \cdot \mathcal{E}_s(x)^2. \end{aligned}$$

Multiplying both sides with  $2(1+\tau)(1+\alpha)(\alpha\bar{\mu})^{-1}$  and choosing  $\alpha := (b_1 + b_2 + \tau)^{\frac{1}{2}} b_2^{-\frac{1}{2}}$  (this minimizes the factor in front of  $\|F_s^\Lambda(x)\|^2$ ), we finally obtain (3.2) with  $B_2(\tau) := (\tau\alpha^2\bar{\mu}^2)^{-1}(1+\tau)(1+\alpha)(\alpha\bar{\mu} + (1+\tau)(1+\alpha))$ . If the full gradient is used, we do not need to apply Young's inequality for  $\tau > 0$ . Thus, we can set  $B_1(\tau) \equiv B_1(0)$ .  $\square$

### A.3. Proof of Lemma 4.3 (ii).

*Proof.* As in [41, Proposition 4.2], we first use the inequality  $e^x \leq x + \exp(9x^2/16)$ ,  $x \in \mathbb{R}$ . Hence, for any matrix  $X \in \mathbb{S}^n$  it follows

$$e^X \preceq X + \exp(9X^2/16) \quad \text{and} \quad e^X \preceq \|e^X\| \cdot I \preceq e^{\|X\|} \cdot I,$$

where “ $\preceq$ ” denotes the usual semidefinite order on the space of symmetric matrices. By Jensen's inequality and under the condition  $\theta \leq \frac{4}{3\bar{\sigma}_k}$ , this implies

$$\begin{aligned} \mathbb{E}[e^{\theta X_k} \mid \mathcal{U}_{k-1}] &\preceq \mathbb{E}[\theta X_k \mid \mathcal{U}_{k-1}] + \mathbb{E}[\exp(9\theta^2 X_k^2/16) \mid \mathcal{U}_{k-1}] \\ &\preceq \mathbb{E}[\exp(9\theta^2 \|X_k\|^2/16) \mid \mathcal{U}_{k-1}] \cdot I \preceq \exp(9\theta^2 \bar{\sigma}_k^2/16) \cdot I \end{aligned}$$

with probability 1. On the other hand, by Young's inequality, we have  $\theta x \leq \frac{3\theta^2 \bar{\sigma}_k^2}{8} + \frac{2x^2}{3\bar{\sigma}_k^2}$  for all  $\theta, x \in \mathbb{R}$ , which easily yields

$$\mathbb{E}[e^{\theta X_k} \mid \mathcal{U}_{k-1}] \preceq \mathbb{E}[e^{\theta \|X_k\|} \mid \mathcal{U}_{k-1}] \cdot I \preceq \exp(3\theta^2 \bar{\sigma}_k^2/8 + 2/3) \cdot I.$$

As in [41], we can now combine the last two estimates and it follows

$$\mathbb{E}[e^{\theta X_k} \mid \mathcal{U}_{k-1}] \preceq \exp(3\theta^2 \bar{\sigma}_k^2/4) \cdot I, \quad \text{a.e.,} \quad \forall \theta \geq 0.$$

The rest of the proof is similar to [81, Theorem 7.1]. Specifically, following the arguments and steps in the proof of [81, Theorem 7.1], we can show

$$\begin{aligned} \mathbb{E}\left[\text{tr} \exp\left(\sum_{k=1}^m \theta X_k\right) \mid \mathcal{U}_0\right] &\leq \mathbb{E}\left[\text{tr} \exp\left(\sum_{k=1}^{m-1} \theta X_k + \log \mathbb{E}[e^{\theta X_m} \mid \mathcal{U}_{m-1}]\right) \mid \mathcal{U}_0\right] \\ &\leq \mathbb{E}\left[\text{tr} \exp\left(\sum_{k=1}^{m-1} \theta X_k + \frac{3\theta^2 \bar{\sigma}_m^2}{4} I\right) \mid \mathcal{U}_0\right] \\ &\leq \dots \leq \exp(3\theta^2 \|\bar{\sigma}\|^2/4) \cdot n. \end{aligned}$$

This estimate can be used in the (conditional) Laplace transform bound [81, Proposition 3.1]. Optimizing with respect to  $\theta$  and setting  $\theta = \frac{2\tau}{3\|\bar{\sigma}\|^2}$ , we then get

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_k X_k\right) \geq \tau \mid \mathcal{U}_0\right) \leq e^{-\theta\tau} \cdot \mathbb{E}\left[\text{tr} \exp\left(\sum_k \theta X_k\right) \mid \mathcal{U}_0\right] \leq n \cdot \exp\left(-\frac{\tau^2}{3\|\bar{\sigma}\|^2}\right).$$

Finally, by rescaling  $\tau$  and applying this result to the *dilations* of the matrices  $X_k$ , [81, Section 2.4], we obtain (4.3).  $\square$

## REFERENCES

- [1] N. AGARWAL, B. BULLINS, AND E. HAZAN, *Second order stochastic optimization in linear time*. <https://arxiv.org/abs/1602.03943v5>, 11 2017.
- [2] C. D. ALIPRANTIS AND K. C. BORDER, *Infinite dimensional analysis*, Springer, Berlin, third ed., 2006. A hitchhiker's guide.

- [3] Z. ALLEN-ZHU, *Katyusha: The first direct acceleration of stochastic gradient methods*, in Proceedings of 49th Annual ACM SIGACT Symposium on the Theory of Computing, 2017, pp. 1200–1205.
- [4] ———, *Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter*, in Proceedings of the 34th Int. Conf. on Mach. Learn., vol. 70, 2017, pp. 89–97.
- [5] Z. ALLEN-ZHU AND E. HAZAN, *Variance reduction for faster non-convex optimization*, in Proceedings of the 33rd Int. Conf. on Mach. Learn., 2016, pp. 699–707.
- [6] Y. F. ATCHADÉ, G. FORT, AND E. MOULINES, *On perturbed proximal gradient algorithms*, J. Mach. Learn. Res., 18 (2017), pp. 310–342.
- [7] J.-P. AUBIN AND H. FRANKOWSKA, *Set-valued analysis*, Modern Birkhäuser Classics, Birkhäuser Boston, Inc., Boston, MA, 2009. Reprint of the 1990 edition.
- [8] F. BACH, R. JENATTON, J. MAIRAL, AND G. OBOZINSKI, *Optimization with sparsity-inducing penalties*, Foundations and Trends® in Machine Learning, 4 (2011), pp. 1–106.
- [9] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex analysis and monotone operator theory in Hilbert spaces*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, New York, 2011.
- [10] A. S. BERAHAS, R. BOLLAPRAGADA, AND J. NOCEDAL, *An investigation of newton-sketch and subsampled newton methods*. <https://arxiv.org/abs/1705.06211>, 05 2017.
- [11] D. P. BERTSEKAS, *Convex Optimization Algorithms*, Athena Scientific, 2015.
- [12] R. BHATTACHARYA AND E. C. WAYMIRE, *A basic course in probability theory*, Universitext, Springer, Cham, second ed., 2016.
- [13] R. BOLLAPRAGADA, R. BYRD, AND J. NOCEDAL, *Exact and inexact subsampled Newton methods for optimization*. <https://arxiv.org/abs/1609.08502>, 9 2016.
- [14] A. BORDES, L. BOTTOU, AND P. GALLINARI, *SGD-QN: careful quasi-Newton stochastic gradient descent*, J. Mach. Learn. Res., 10 (2009), pp. 1737–1754.
- [15] L. BOTTOU, F. E. CURTIS, AND J. NOCEDAL, *Optimization methods for large-scale machine learning*. <https://arxiv.org/abs/1606.04838>, 06 2016.
- [16] R. H. BYRD, G. M. CHIN, W. NEVEITT, AND J. NOCEDAL, *On the use of stochastic Hessian information in optimization methods for machine learning*, SIAM J. Optim., 21 (2011), pp. 977–995.
- [17] R. H. BYRD, G. M. CHIN, J. NOCEDAL, AND Y. WU, *Sample size selection in optimization methods for machine learning*, Math. Program., 134 (2012), pp. 127–155.
- [18] R. H. BYRD, S. L. HANSEN, J. NOCEDAL, AND Y. SINGER, *A stochastic quasi-Newton method for large-scale optimization*, SIAM J. Optim., 26 (2016), pp. 1008–1031.
- [19] F. H. CLARKE, *Optimization and nonsmooth analysis*, vol. 5 of Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second ed., 1990.
- [20] P. L. COMBETTES AND V. R. WAJS, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200 (electronic).
- [21] C. D. DANG AND G. LAN, *Stochastic block mirror descent methods for nonsmooth and stochastic optimization*, SIAM J. Optim., 25 (2015), pp. 856–881.
- [22] A. DEFAZIO, F. BACH, AND S. LACOSTE-JULIEN, *SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives*, in Adv. in Neural Inf. Process. Syst., 2014, pp. 1646–1654.
- [23] L. DENG AND D. YU, *Deep learning: Methods and applications*, Foundations and Trends® Signal Processing, 7 (2014), pp. 197–387.
- [24] J. DUCHI, E. HAZAN, AND Y. SINGER, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learn. Res., 12 (2011), pp. 2121–2159.
- [25] M. A. ERDOGDU AND A. MONTANARI, *Convergence rates of sub-sampled newton methods*, in Adv. in Neural Inf. Process. Syst., 2015, pp. 3034–3042.
- [26] F. FACCHINEI AND J.-S. PANG, *Finite-dimensional variational inequalities and complementarity problems. Vol. II*, Springer-Verlag, New York, 2003.
- [27] M. P. FRIEDLANDER AND M. SCHMIDT, *Hybrid deterministic-stochastic methods for data fitting*, SIAM J. Sci. Comput., 34 (2012), pp. A1380–A1405.
- [28] M. FUKUSHIMA AND H. MINE, *A generalized proximal point algorithm for certain nonconvex minimization problems*, Internat. J. Systems Sci., 12 (1981), pp. 989–1000.
- [29] S. GHADIMI AND G. LAN, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim., 23 (2013), pp. 2341–2368.
- [30] ———, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), pp. 59–99.
- [31] S. GHADIMI, G. LAN, AND H. ZHANG, *Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization*, Math. Program., 155 (2016), pp. 267–305.
- [32] G. H. GOLUB AND C. F. VAN LOAN, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, fourth ed., 2013.
- [33] I. GOODFELLOW, Y. BENGIO, AND A. COURVILLE, *Deep Learning*, MIT Press, 2016.
- [34] R. GOWER, D. GOLDFARB, AND P. RICHTARIK, *Stochastic block BFGS: Squeezing more curvature out of data*, in Proceedings of the 33rd Int. Conf. on Mach. Learn., 2016, pp. 1869–1878.

- [35] M. GÜRBÜZBALABAN, A. OZDAGLAR, AND P. PARRILO, *A globally convergent incremental Newton method*, Math. Program., 151 (2015), pp. 283–313.
- [36] I. GUYON, S. GUNN, A. BEN-HUR, AND G. DROR, *Result analysis of the NIPS 2003 feature selection challenge*, in Adv. in Neural Inf. Process. Syst. 17, MIT Press, 2004, pp. 545–552.
- [37] T. HASTIE, R. TIBSHIRANI, AND J. FRIEDMAN, *The elements of statistical learning*, Springer Series in Statistics, Springer, New York, second ed., 2009. Data mining, inference, and prediction.
- [38] J.-B. HIRIART-URRUTY, J.-J. STRODIOT, AND V. H. NGUYEN, *Generalized Hessian matrix and second-order optimality conditions for problems with  $C^{1,1}$  data*, Appl. Math. Optim., 11 (1984), pp. 43–56.
- [39] A. N. IUSEM, A. JOFRÉ, R. I. OLIVEIRA, AND P. THOMPSON, *Extragradient method with variance reduction for stochastic variational inequalities*, SIAM J. Optim., 27 (2017), pp. 686–724.
- [40] R. JOHNSON AND T. ZHANG, *Accelerating stochastic gradient descent using predictive variance reduction*, in Adv. in Neural Inf. Process. Syst., 2013, pp. 315–323.
- [41] A. B. JUDITSKY AND A. S. NEMIROVSKI, *Large deviations of vector-valued martingales in 2-smooth normed spaces*. Preprint available at <https://arxiv.org/abs/0809.0813>, 09 2008.
- [42] Y. LECUN, Y. BENGIO, AND G. HINTON, *Deep learning*, Nature, 521 (2015), pp. 436–444.
- [43] Y. LECUN, C. CORTES, AND C. J. C. BURGESS, *The mnist database of handwritten digits*. <http://yann.lecun.com/exdb/mnist/>. 2010.
- [44] J. D. LEE, Y. SUN, AND M. A. SAUNDERS, *Proximal Newton-type methods for minimizing composite functions*, SIAM J. Optim., 24 (2014), pp. 1420–1443.
- [45] D. D. LEWIS, Y. YANG, T. G. ROSE, AND F. LI, *RCV1: A new benchmark collection for text categorization research*, J. Mach. Learn. Res., 5 (2004), pp. 361–397.
- [46] H. LIN, J. MAIRAL, AND Z. HARCHAOU, *A universal catalyst for first-order optimization*, in Adv. in Neural Inf. Process. Syst., 2015, pp. 3384–3392.
- [47] Q. LIN, Z. LU, AND L. XIAO, *An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization*, SIAM J. Optim., 25 (2015), pp. 2244–2273.
- [48] J. MAIRAL, F. BACH, J. PONCE, AND G. SAPIRO, *Online dictionary learning for sparse coding*, in Proceedings of the 26th Annual Int. Conf. on Mach. Learn., ICML '09, New York, NY, USA, 2009, pp. 689–696.
- [49] L. MASON, J. BAXTER, P. BARTLETT, AND M. FREAN, *Boosting algorithms as gradient descent in function space*, in Proceedings of the 12th Int. Conf. on Neural Inf. Process. Syst., NIPS'99, 1999, pp. 512–518.
- [50] F. MENG, D. SUN, AND G. ZHAO, *Semismoothness of solutions to generalized equations and the Moreau-Yosida regularization*, Math. Program., 104 (2005), pp. 561–581.
- [51] A. MILZAREK, *Numerical methods and second order theory for nonsmooth problems*, PhD Dissertation, Technische Universität München, 2016.
- [52] A. MILZAREK AND M. ULBRICH, *A semismooth Newton method with multidimensional filter globalization for  $l_1$ -optimization*, SIAM J. Optim., 24 (2014), pp. 298–333.
- [53] H. MINE AND M. FUKUSHIMA, *A minimization method for the sum of a convex function and a continuously differentiable function*, J. Optim. Theory Appl., 33 (1981), pp. 9–23.
- [54] A. MOKHTARI AND A. RIBEIRO, *RES: regularized stochastic BFGS algorithm*, IEEE Trans. Signal Process., 62 (2014), pp. 6089–6104.
- [55] J. J. MORÉ AND D. C. SORESENSEN, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
- [56] J.-J. MOREAU, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.
- [57] P. MORITZ, R. NISHIHARA, AND M. JORDAN, *A linearly-convergent stochastic L-BFGS algorithm*, in Proceedings of the 19th Int. Conf. on Artif. Intelligence and Statistics, 2016, pp. 249–258.
- [58] A. NEMIROVSKI, A. JUDITSKY, G. LAN, AND A. SHAPIRO, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim., 19 (2008), pp. 1574–1609.
- [59] J.-S. PANG AND L. QI, *Nonsmooth equations: motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.
- [60] N. PARIKH AND S. BOYD, *Proximal algorithms*, Foundations and Trends® in Optimization, 1 (2014), pp. 127–239.
- [61] P. PATRINOS, L. STELLA, AND A. BEMPORAD, *Forward-backward truncated Newton methods for convex composite optimization*. <https://arxiv.org/abs/1402.6655>, 2 2014.
- [62] M. PILANCI AND M. J. WAINWRIGHT, *Newton sketch: a near linear-time optimization algorithm with linear-quadratic convergence*, SIAM J. Optim., 27 (2017), pp. 205–245.
- [63] M. J. D. POWELL, *A fast algorithm for nonlinearly constrained optimization calculations*, (1978), pp. 144–157. Lecture Notes in Math., Vol. 630.
- [64] L. QI, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res., 18 (1993), pp. 227–244.
- [65] L. QI AND J. SUN, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.
- [66] S. J. REDDI, A. HEFNY, S. SRA, B. PÓCZOS, AND A. J. SMOLA, *Stochastic variance reduction for non-convex optimization*, in Proceedings of the 33rd Int. Conf. on Mach. Learn., 2016, pp. 314–323.



- [67] S. J. REDDI, S. SRA, B. PÓCZOS, AND A. J. SMOLA, *Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization*, in Adv. in Neural Inf. Process. Syst. 29, 2016, pp. 1145–1153.
- [68] P. RICHTÁRIK AND M. TAKÁČ, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Math. Program., 144 (2014), pp. 1–38.
- [69] H. ROBBINS AND S. MONRO, *A stochastic approximation method*, Ann. Math. Statistics, 22 (1951), pp. 400–407.
- [70] R. T. ROCKAFELLAR, *Integral functionals, normal integrands and measurable selections*, (1976), pp. 157–207. Lecture Notes in Math., Vol. 543.
- [71] F. ROOSTA-KHORASANI AND M. W. MAHONEY, *Sub-sampled Newton methods I: Globally convergent algorithms*. <https://arxiv.org/abs/1601.04737>, 2 2016.
- [72] ———, *Sub-sampled Newton methods II: Local convergence rates*. <https://arxiv.org/abs/1601.04738>, 2 2016.
- [73] J. SCHMIDHUBER, *Deep learning in neural networks: An overview*, Neural Networks, 61 (2015), pp. 85 – 117.
- [74] M. SCHMIDT, N. LE ROUX, AND F. BACH, *Minimizing finite sums with the stochastic average gradient*, Math. Program., 162 (2017), pp. 83–112.
- [75] N. N. SCHRAUDOLPH, J. YU, AND S. GÜNTHER, *A stochastic quasi-newton method for online convex optimization*, in Proceedings of the 11th Int. Conf. on Artif. Intelligence and Statistics, vol. 2, PMLR, 2007, pp. 436–443.
- [76] S. SHALEV-SHWARTZ AND S. BEN-DAVID, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, New York, NY, USA, 2014.
- [77] S. SHALEV-SHWARTZ AND T. ZHANG, *Stochastic dual coordinate ascent methods for regularized loss minimization*, J. Mach. Learn. Res., 14 (2013), pp. 567–599.
- [78] ———, *Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization*, Math. Program., 155 (2016), pp. 105–145.
- [79] Z. SHI AND R. LIU, *Large scale optimization with proximal stochastic Newton-type gradient descent*, in Mach. Learn. and Knowl. Disc. in Databases, vol. 9284, Springer Int. Publishing, 2015, pp. 691–704.
- [80] L. STELLA, A. THEMELIS, AND P. PATRINOS, *Forward-backward quasi-Newton methods for nonsmooth optimization problems*, Comput. Optim. Appl., 67 (2017), pp. 443–487.
- [81] J. A. TROPP, *User-friendly tail bounds for sums of random matrices*, Found. Comput. Math., 12 (2012), pp. 389–434.
- [82] P. TSENG AND S. YUN, *A coordinate gradient descent method for nonsmooth separable minimization*, Math. Program., 117 (2009), pp. 387–423.
- [83] M. ULBRICH, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2002), pp. 805–842 (2003).
- [84] ———, *Semismooth Newton methods for variational inequalities and constrained optimization problems in function spaces*, vol. 11 of MOS-SIAM Series on Optimization, SIAM and MOS, Philadelphia, 2011.
- [85] J. WANG AND T. ZHANG, *Improved optimization of finite sums with minibatch stochastic variance reduced proximal iterations*. <https://arxiv.org/abs/1706.07001>, 06 2017.
- [86] X. WANG, S. MA, D. GOLDFARB, AND W. LIU, *Stochastic Quasi-Newton Methods for Nonconvex Stochastic Optimization*, SIAM J. Optim., 27 (2017), pp. 927–956.
- [87] E. W. WEISSTEIN, *Infinite product*. From MathWorld—A Wolfram Web Resource, <http://mathworld.wolfram.com/InfiniteProduct.html>. Accessed: 2017-10-26.
- [88] D. WILLIAMS, *Probability with martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, 1991.
- [89] C. WORKBENCH TEAM, *A marketing dataset*. <http://www.causality.inf.ethz.ch/data/CINA.html>, 09 2008.
- [90] L. XIAO AND T. ZHANG, *A proximal stochastic gradient method with progressive variance reduction*, SIAM J. Optim., 24 (2014), pp. 2057–2075.
- [91] X. XIAO, Y. LI, Z. WEN, AND L. ZHANG, *A regularized semi-smooth Newton method with projection steps for composite convex programs*. <https://arxiv.org/abs/1603.07870>, 03 2016.
- [92] P. XU, F. ROOSTA-KHORASANI, AND M. W. MAHONEY, *Newton-type methods for non-convex optimization under inexact Hessian information*. <https://arxiv.org/abs/1708.07164>, 8 2017.
- [93] P. XU, F. ROOSTA-KHORASANI, AND M. W. MAHONEY, *Second-order optimization for non-convex machine learning: An empirical study*. <https://arxiv.org/abs/1708.07827>, 8 2017.
- [94] P. XU, J. YANG, F. ROOSTA-KHORASANI, C. RÉ, AND M. W. MAHONEY, *Sub-sampled Newton methods with non-uniform sampling*, in Adv. in Neural Inf. Process. Syst., 2016, pp. 3000–3008.
- [95] Y. XU AND W. YIN, *Block stochastic gradient iteration for convex and nonconvex optimization*, SIAM J. Optim., 25 (2015), pp. 1686–1716.
- [96] Z. YAO, P. XU, F. ROOSTA-KHORASANI, AND M. W. MAHONEY, *Inexact non-convex Newton-type methods*. <https://arxiv.org/abs/1802.06925>, 2 2017.
- [97] H. YE, L. LUO, AND Z. ZHANG, *Approximate Newton methods and their local convergence*, in Proceedings of the 34th Int. Conf. on Mach. Learn., vol. 70, PMLR, 2017, pp. 3931–3939.