Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

STOCHASTIC ADAPTIVE REGULARIZATION METHOD WITH CUBICS: A HIGH PROBABILITY COMPLEXITY BOUND

Katya Scheinberg
Miaolan Xie

Operations Research and Information Engineering
Cornell University
136 Hoy Rd
Ithaca, NY 14853, USA

ABSTRACT

We present a high probability complexity bound for a stochastic adaptive regularization method with cubics, also known as regularized Newton method. The method makes use of stochastic zeroth-, first- and second-order oracles that satisfy certain accuracy and reliability assumptions. Such oracles have been used in the literature by other stochastic adaptive methods, such as trust region and line search. These oracles capture many settings, such as expected risk minimization, simulation optimization, and others. In this paper, we give the first high probability iteration bound for stochastic cubic regularization, and show that just as in the deterministic case, it is superior to other stochastic adaptive methods.

1 INTRODUCTION

We are interested in unconstrained optimization problems of the form

$$\min_{x\in\mathbb{R}^n} \phi(x),$$

where ϕ is possibly nonconvex and satisfies the following condition:

Assumption 1 ϕ is bounded from below by a constant ϕ^*, is twice continuously differentiable, and has globally L-Lipschitz continuous gradient and L_H-Lipschitz continuous Hessian.

We present and analyze a stochastic adaptive cubic regularization algorithm for computing a point x such that $\|\nabla\phi(x)\| \leq \varepsilon$, for some $\varepsilon > 0$, when $\phi(x)$ or its derivatives are not computable exactly. Specifically, we assume access to stochastic zeroth-, first- and second-order oracles, which are defined as follows.

Stochastic zeroth-order oracle ($\mathsf{SZO}(\varepsilon_f,\lambda,a)$). Given a point x, the oracle computes $f(x,\Xi(x))$, where $\Xi(x)$ is a random variable, whose distribution may depend on x, ε_f,λ and a, that satisfies

$$\mathbb{E}_{\Xi(x)}\left[|\phi(x)-f(x,\Xi(x))|\right] \leq \varepsilon_f \quad \text{and} \quad \mathbb{P}_{\Xi(x)}\left(|\phi(x)-f(x,\Xi(x))| < t\right) \geq 1-e^{\lambda(a-t)}, \tag{1}$$

for any $t > 0$.

We view x as the input to the oracle, $f(x,\Xi(x))$ as the output and the values $(\varepsilon_f,\lambda,a)$ as values intrinsic to the oracle. Thus $|f(x,\Xi(x))-\phi(x)|$ is a sub-exponential random variable with parameters (λ,a), whose mean is bounded by some constant $\varepsilon_f > 0$.

Stochastic first-order oracle ($\mathsf{SFO}(\kappa_g)$. Given a point x and constants $\mu_1 > 0$, $\delta_1 \in [0,\frac{1}{2})$, the oracle computes $g(x,\Xi^1(x))$, such that

$$\mathbb{P}_{\Xi^1(x)}(\|\nabla\phi(x)-g(x,\Xi^1(x))\| \leq \kappa_g\mu_1) \geq 1-\delta_1, \tag{2}$$

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

where $\Xi^1(x)$ is a random variable whose distribution may depend on x, μ_1, δ_1 and κ_g. We view x, μ_1 and δ_1 as inputs to the oracle, while κ_g is intrinsic to the oracle.

Stochastic second-order oracle ($\mathsf{SSO}(\kappa_H)$). Given a point x and constants $\mu_2 > 0$, $\delta_2 \in [0, \frac{1}{2})$, the oracle computes $H(x, \Xi^2(x))$, such that

$$\mathbb{P}_{\Xi^2(x)}(\|\nabla^2\phi(x) - H(x, \Xi^2(x))\| \leq \kappa_H \mu_2) \geq 1 - \delta_2, \tag{3}$$

where $\Xi^2(x)$ is a random variable whose distribution may depend on x, μ_2, δ_2 and κ_H. The norm on the matrix is the operator norm. x, μ_2 and δ_2 are inputs to the oracle, while κ_H is intrinsic to the oracle.

Wherever possible, we will omit the dependence on x and write Ξ, Ξ^1, and Ξ^2 instead of $\Xi(x), \Xi^1(x)$, and $\Xi^2(x)$, and we will use $f(x), g(x)$ and $H(x)$ to denote the outputs of the stochastic oracles for brevity.

Related work. Several stochastic adaptive optimization methods have been studied in recent years under various stochastic oracle assumptions, similar to the ones we present above. Specifically, Cartis and Scheinberg (2018) bounds the expected iteration complexity for an adaptive step search (line search) method and an adaptive cubic regularization method, with a similar first-order oracle, but with an exact zeroth-order oracle. In Paquette and Scheinberg (2020), an expected iteration complexity result is derived for a variant of a step search method under a stochastic zeroth-order oracle. In Jin et al. (2021), the results of Paquette and Scheinberg (2020) are strengthened under a somewhat more restrictive zeroth-order oracle, which is similar to the SZO in this paper, and a high probability complexity bound is derived.

Similarly, in Bandeira et al. (2014), and Gratton et al. (2018), a trust region method is analyzed under essentially SFO and SSO, but with an exact zeroth-order oracle. Later in Chen et al. (2018) and Blanchet et al. (2019) an expected complexity bound is derived for a trust region method with a stochastic zeroth-order oracle. In Cao et al. (2022) a high probability iteration complexity bound for first- and second-order stochastic trust region methods is derived under the same oracles we use. Recently, the same oracles were used within a stochastic quasi-Newton method in Menickelly et al. (2023), and a stochastic SQP-based method for nonlinear equality constrained problems in Berahas et al. (2023).

Adaptive regularization with cubics (ARC) methods are theoretically superior alternatives to line search and trust region methods, when applied to deterministic smooth functions, because of their optimal complexity of $O(\varepsilon^{-3/2})$ vs $O(\varepsilon^{-2})$ for finding ε stationary points (Cartis et al. 2011c; Cartis et al. 2011a). There are many variants of adaptive cubic regularization methods under various assumptions and requirements on the function value, gradient, and Hessian estimates. Specifically, in Cartis et al. (2011b), Liu et al. (2018), Bellavia et al. (2019), Wang et al. (2019), Kohler and Lucchi (2017), Park et al. (2020), the oracles are assumed to be deterministically bounded, with adaptive magnitude or errors. In Cartis and Scheinberg (2018), Bellavia and Gurioli (2022), Bellavia et al. (2022), Bellavia et al. (2020), bounds on expected complexity are provided under exact or deterministically bounded zeroth-order oracle and the gradient and Hessian oracles similar to SFO and SSO. A cubicly regularized method in a fully stochastic setting is analyzed in Tripuraneni et al. (2018). The method is not adaptive, relying on the knowledge of the Lipschitz constants of $\nabla\phi(x)$, and therefore not requiring a zeroth-order oracle at all. However, the results in that paper are simply derived assuming that stochastic gradient and Hessian estimates are sufficiently accurate at each iteration. The final complexity bound only applies with probability that this holds true. No expected complexity bound can thus be derived.

Our contributions. In this work we provide the first high probability analysis of a stochastic ARC method (SARC) that allows 1. Stochastic function estimates that can have arbitrarily large errors, and 2. Stochastic gradient and Hessian approximations whose error is bounded by an adaptive quantity with sufficiently high probability, but otherwise can be arbitrarily large. To the best of our knowledge, our work is the first to derive an iteration complexity bound that matches the deterministic iteration complexity of $O(\varepsilon^{-3/2})$ in this setting with an overwhelmingly high probability. We show that our variant of stochastic ARC, while more general than those in prior literature, still maintains its optimal iteration complexity.

The analysis presented here extends the stochastic settings and high probability results in Jin et al. (2021) and Cao et al. (2022) to the framework in Cartis and Scheinberg (2018). However, this extension is far from trivial, as it requires careful modification of most of the elements of the existing analysis. We point out these modifications in the appropriate places in the paper.

The oracles used in this paper are essentially the same as in Jin et al. (2021) and Cao et al. (2022). However, our assumption on the oracles is a bit stronger in this paper than in these two previous works. In particular, we assume that SFO and SSO are implementable for arbitrarily small values of μ_1 and μ_2, respectively. In contrast, the analysis in Jin et al. (2021) and Cao et al. (2022) allows for the case when these oracles cannot be implemented for arbitrarily small error bounds. We will discuss further in the paper, that even though SARC may impose small values of μ_1 and μ_2, this happens only with small probability.

We do not discuss the numerical performance of our method in this paper. Although deterministic implementations of ARC can be competitive with trust-region and line search methods when implemented with care, their efficiency in practice is highly dependent on the subproblem solver used. We expect similar behavior in the stochastic case and leave this study as a subject of future research.

2 STOCHASTIC ADAPTIVE REGULARIZATION METHOD WITH CUBICS (SARC) WITH PROBABILISTIC SECOND-ORDER MODELS

The Stochastic Adaptive Regularization with Cubics (SARC) method is presented below as Algorithm 1. At each iteration k, given gradient estimate g_k, Hessian estimate H_k, and a regularization parameter $\sigma_k > 0$, the following model is approximately minimized with respect to s to obtain the trial step s_k:

$$m_k(x_k+s) = \phi(x_k) + s^T g_k + \frac{1}{2} s^T H_k s + \frac{\sigma_k}{3}\|s\|^3. \tag{4}$$

The constant term $\phi(x_k)$ is never computed and is used simply for presentation purposes. In the case of SARC, g_k and H_k are computed using SFO and SSO so as to satisfy certain accuracy requirements with sufficiently high probability, which will be specified in Section 3. We require the trial step s_k to be an “approximate minimizer" of $m_k(x_k+s)$ in the sense that it has to satisfy:

$$(s_k)^T g_k + (s_k)^T H_k s_k + \sigma_k \|s_k\|^3 = 0 \text{ and } (s_k)^T H_k s_k + \sigma_k \|s_k\|^3 \geq 0 \tag{5}$$

and

$$\|\nabla m_k(x_k+s_k)\| \leq \eta \min\{1, \|s_k\|\} \|g_k\|, \tag{6}$$

where $\eta \in (0,1)$ is a user-chosen constant. The conditions are typical in the literature, e.g., in (Cartis et al. 2011c) and can be satisfied, for example, using algorithms in (Cartis et al. 2011b; Carmon and Duchi 2019) to approximately minimize the model (4), as well as by any global minimizer of $m_k(x_k+s)$.

As in any variant of the ARC method, once s_k is computed, the trial step is accepted (and σ_k is decreased) if the estimated function value of $x_k^+ = x_k + s_k$ is sufficiently smaller than that of x_k, when compared to the model value decrease. We call these iterations *successful*. Otherwise, the trial step is rejected (and σ_k is increased). We call these iterations *unsuccessful*. In the case of SARC, however, function value estimates are obtained via SZO and the step acceptance criterion is modified by adding an "error correction" term of $2\varepsilon_f'$. This is because function value estimates have an irreducible error, so without this correction term, the algorithm may always reject improvement steps.

3 DETERMINISTIC PROPERTIES OF ALGORITHM 1

Algorithm 1 generates a stochastic process and we will analyze it in the next section. First, however, we state and prove several lemmas that establish the behavior of the algorithm *for every realization*.

A key concept that will be used in the analysis is the concept of a *true iteration*. Let $e_k = |f(x_k) - \phi(x_k)|$ and $e_k^+ = |f(x_k^+) - \phi(x_k^+)|$.

Algorithm 1: Stochastic Adaptive Regularization with Cubics (SARC)

```
Input: Oracles $\mathsf{SZO}(\varepsilon_f, \lambda, a)$, $\mathsf{SFO}(\kappa_g)$ and $\mathsf{SSO}(\kappa_H)$; initial iterate $x_0$, parameters $\gamma \in (0,1)$,
    $\theta \in (0,1)$, $\delta_1, \delta_2 \in [0, \frac{1}{2})$, $\sigma_{\min} > 0$, $\eta \in (0,1)$, $\mu \geq 0, \varepsilon_f' > 0$ and $\sigma_0 \geq \sigma_{\min}$.
Repeat for $k = 0, 1, \dots$
    1. Compute a model trial step $s_k$: Generate $g_k = g(x_k, \xi_k^1)$ and $H_k = H(x_k, \xi_k^2)$ using $\mathsf{SFO}(\kappa_g)$
       and $\mathsf{SSO}(\kappa_H)$ with $(\frac{\mu}{\sigma_k}, \delta_1)$, and $(\sqrt{\frac{\mu}{\sigma_k}}, \delta_2)$ as inputs, respectively. Compute a trial step $s_k$ that
       satisfies (5) and (6) with parameters $\eta$ and $\sigma_k$ at $x_k$.
    2. Check sufficient decrease: Let $x_k^+ = x_k + s_k$. Compute function value estimations $f(x_k) =$
       $f(x_k, \xi_k)$ and $f(x_k^+) = f(x_k^+, \xi_k^+)$ using the $\mathsf{SZO}(\varepsilon_f, \lambda, a)$, and set
           $$\rho_k = \frac{f(x_k) - f(x_k^+) + 2\varepsilon_f'}{m(x_k) - m_k(x_k^+)}, \tag{7}$$
    3. Update the iterate: Set
           $$x_{k+1} = \begin{cases} x_k^+ & \text{if} \quad \rho_k \geq \theta \quad \text{[successful iteration]} \\ x_k, & \text{otherwise} \quad \text{[unsuccessful iteration]} \end{cases} \tag{8}$$
    4. Update the regularization parameter $\sigma_k$: Set
           $$\sigma_{k+1} = \begin{cases} \max\{\gamma\sigma_k, \sigma_{\min}\} & \text{if} \quad \rho_k \geq \theta \\ \frac{1}{\gamma}\sigma_k, & \text{otherwise.} \end{cases}$$
```

Definition 1 (True iteration) We say that iteration k is **true** if

$$\|\nabla\phi(x_k) - g_k\| \leq \kappa_g \max\left\{\frac{\mu}{\sigma_k}, \|s_k\|^2\right\}, \ \|(\nabla^2\phi(x_k) - H_k)s_k\| \leq \kappa_H \max\left\{\frac{\mu}{\sigma_k}, \|s_k\|^2\right\} \tag{9}$$
$$\text{and } e_k + e_k^+ \leq 2\varepsilon_f'.$$

Remark 1 We will show in Lemma 6 that by using SFO and SSO with the respective inputs, $\mu_1 = \frac{\mu}{\sigma_k}$ in (2) and $\mu_2 = \sqrt{\frac{\mu}{\sigma_k}}$ in (3), each iteration of Algorithm 1 satifies (9) with probability at least $1 - \delta_1 - \delta_2$. However, we note that the probabilistic requirement of (9) can be implied by more relaxed inputs that use $\mu_1 = \max\{\frac{\mu}{\sigma_k}, \|s_k\|^2\}$ in (2), and $\mu_2 = \max\{\frac{\mu}{\sigma_k\|s_k\|}, \|s_k\|\}$ in (3), instead of $\mu_1 = \frac{\mu}{\sigma_k}$ and $\mu_2 = \sqrt{\frac{\mu}{\sigma_k}}$. Since s_k depends on the output of the oracles, implementing such a relaxed version is not trivial, and may require modification of Algorithm 1. We leave it as a subject of future research.

We will now prove a sequence of results that hold for each realization of Algorithm 1, and are essential for the complexity analysis. The two key results are Corollary 1 and Lemma 5, where Corollary 1 shows that until an ε-stationary point is reached, every true iteration with large enough σ_k is successful, and Lemma 5 establishes the lower bound on function improvement on true and successful iterations. Lemmas 1 to 4 lay the building blocks for them: On every successful iteration, the function improvement is lower bounded in terms of the norm of the step (Lemma 1). There is a threshold value of σ_k where any true iteration is either always successful or results in a very small step (Lemma 2). When an iteration is true and the step is not very small, the norm of the step is lower bounded in terms of $\|\nabla\phi(x_k^+)\|$ (Lemma 3). Until an ε-stationary point is reached, the step cannot be too small on true iterations (Lemma 4).

Lemma 1 (Improvement on successful iterations) Consider any realization of Algorithm 1. For each iteration k, we have

$$m_k(x_k) - m_k(x_k^+) \geq \frac{1}{6}\sigma_k \|s_k\|^3. \tag{10}$$

On every successful iteration k, we have

$$f(x_k) - f(x_{k+1}) \geq \frac{\theta}{6}\sigma_k \|s_k\|^3 - 2\varepsilon_f', \tag{11}$$

which implies

$$\phi(x_k) - \phi(x_{k+1}) \geq \frac{\theta}{6}\sigma_k \|s_k\|^3 - e_k - e_k^+ - 2\varepsilon_f'. \tag{12}$$

Proof. The proof is similar to the proof of Lemma 3.3 in Cartis et al. (2011b). Clearly, (11) follows from (10) and the sufficient decrease condition (7)-(8):

$$\frac{f(x_k) - f(x_k^+) + 2\varepsilon_f'}{m_k(x_k) - m_k(x_k^+)} \geq \theta,$$

and (12) follows from the definition of e_k and e_k^+.

It remains to prove (10). Combining the first condition on step s_k in (5), with the model expression for $s = s_k$, we can write

$$m_k(x_k) - m_k(x_k^+) = \frac{1}{2}(s_k)^T H_k s_k + \frac{2}{3}\sigma_k \|s_k\|^3.$$

The second condition on s_k in (5) implies $(s_k)^T H_k s_k \geq -\sigma_k \|s_k\|^3$. Together with the above equation, we obtain (10).

□

Lemma 2 (Large σ_k guarantees success or small step) Let Assumption 1 hold. For any realization of Algorithm 1, if iteration k is true, and if

$$\sigma_k \geq \bar{\sigma} = \frac{2\kappa_g + \kappa_H + L + L_H}{1 - \frac{1}{3}\theta}, \tag{13}$$

then iteration k is either successful or produces s_k such that $\|s_k\|^2 < \frac{\mu}{\sigma_k}$.

Proof. Clearly, if $\rho_k - 1 \geq 0$, then k is successful by definition. Let us consider the case when $\rho_k < 1$; then if $1 - \rho_k \leq 1 - \theta$, k is successful. We have from (7), that

$$1 - \rho_k = \frac{m_k(x_k) - m_k(x_k^+) - f(x_k) + f(x_k^+) - 2\varepsilon_f'}{m_k(x_k) - m_k(x_k^+)}.$$

Notice that:

$$\begin{aligned} m_k(x_k) - m_k(x_k^+) - f(x_k) + f(x_k^+) - 2\varepsilon_f' &= f(x_k^+) - \left(f(x_k) + s_k^T g_k + \frac{1}{2}s_k^T H_k s_k + \frac{\sigma_k}{3}\|s_k\|^3\right) - 2\varepsilon_f' \\ &\leq \phi(x_k^+) - \left(\phi(x_k) + s_k^T g_k + \frac{1}{2}s_k^T H_k s_k + \frac{\sigma_k}{3}\|s_k\|^3\right) - 2\varepsilon_f' + e_k + e_k^+ \\ &\leq \phi(x_k^+) - \phi(x_k) - s_k^T g_k - \frac{1}{2}s_k^T H_k s_k - \frac{\sigma_k}{3}\|s_k\|^3. \end{aligned}$$

The second to last inequality follows from the definition of e_k and e_k^+, and the last inequality due to the iteration being true.

Taylor expansion and Cauchy-Schwarz inequalities give, for some $\tau \in [x_k, x_k^+]$,

$$\begin{aligned}
&\phi(x_k^+) - \phi(x_k) - s_k^T g_k - \tfrac{1}{2} s_k^T H_k s_k - \tfrac{\sigma_k}{3}\|s_k\|^3 \\
&= [\nabla\phi(x_k) - g_k]^T s_k + \tfrac{1}{2}(s_k)^T[\nabla^2\phi(\tau) - \nabla^2\phi(x_k)]s_k + \tfrac{1}{2}(s_k)^T[\nabla^2\phi(x_k) - H_k]s_k - \tfrac{1}{3}\sigma_k\|s_k\|^3 \\
&\leq \|\nabla\phi(x_k) - g_k\| \cdot \|s_k\| + \tfrac{1}{2}\|\nabla^2\phi(\tau) - \nabla^2\phi(x_k)\| \cdot \|s_k\|^2 + \tfrac{1}{2}\|(\nabla^2\phi(x_k) - H_k)s_k\| \cdot \|s_k\| - \tfrac{1}{3}\sigma_k\|s_k\|^3 \\
&\leq (\kappa_g + \tfrac{\kappa_H}{2}) \max\left\{\tfrac{\mu}{\sigma_k}, \|s_k\|^2\right\}\|s_k\| + \left(\tfrac{L_H}{2} - \tfrac{1}{3}\sigma_k\right)\|s_k\|^3
\end{aligned}$$

where the last inequality follows from the fact that the iteration is true and hence (9) holds: $\|\nabla\phi(x_k) - g_k\| \leq \kappa_g \max\left\{\frac{\mu}{\sigma_k}, \|s_k\|^2\right\}$ and $\|(\nabla^2\phi(x_k) - H_k)s_k\| \leq \kappa_H \max\left\{\frac{\mu}{\sigma_k}, \|s_k\|^2\right\}$ and from Assumption 1. So as long as $\|s_k\|^2 \geq \frac{\mu}{\sigma_k}$, we have

$$m_k(x_k) - m_k(x_k^+) - f(x_k) + f(x_k^+) - 2\varepsilon_f' \leq \left(\kappa_g + \frac{\kappa_H}{2} + \frac{L_H}{2} - \frac{1}{3}\sigma_k\right)\|s_k\|^3 = (6\kappa_g + 3L_H + 3\kappa_H - 2\sigma_k)\frac{1}{6}\|s_k\|^3,$$

which, together with (10), gives that $1 - \rho_k \leq 1 - \theta$ when σ_k satisfies (13). □

Note that for the above lemma to hold, $\bar\sigma$ does not need to include L in the numerator. However, we will need another condition on $\bar\sigma$ later that will involve L; hence for simplicity of notation we introduced $\bar\sigma$ above to satisfy all necessary bounds.

Lemma 3 (Lower bound on step norm in terms of $\|\nabla\phi(x_k^+)\|$) Let Assumption 1 hold. For any realization of Algorithm 1, if k is a true iteration we have

$$\max\left\{\|s_k\|^2, \frac{\mu}{\sigma_k}\right\} \geq \frac{1-\eta}{\sigma_k + (1 - \frac{\theta}{3})\bar\sigma}\|\nabla\phi(x_k^+)\|. \tag{14}$$

Proof. The triangle inequality, the equality $\nabla m_k(x_k + s) = g_k + H_k s + \sigma_k\|s\|s$ and condition (6) on s_k together give

$$\begin{aligned}
\|\nabla\phi(x_k^+)\| &\leq \|\nabla\phi(x_k^+) - \nabla m_k(x_k^+)\| + \|\nabla m_k(x_k^+)\| \\
&\leq \|\nabla\phi(x_k^+) - g_k - H_k s_k\| + \sigma_k\|s_k\|^2 + \eta\min\{1, \|s_k\|\}\|g_k\|.
\end{aligned} \tag{15}$$

Recalling Taylor expansion of $\nabla\phi(x_k^+)$: $\nabla\phi(x_k^+) = \nabla\phi(x_k) + \int_0^1 \nabla^2\phi(x_k + ts_k)s_k dt$, and applying triangle inequality again, we have

$$\begin{aligned}
\|\nabla\phi(x_k^+) - g_k - H_k s_k\| &\leq \|\nabla\phi(x_k) - g_k\| + \left\|\int_0^1 [\nabla^2\phi(x_k + ts_k) - \nabla^2\phi(x_k)]s_k dt\right\| + \|\nabla^2\phi(x_k)s_k - H_k s_k\| \\
&\leq (\kappa_g + \kappa_H)\max\left\{\tfrac{\mu}{\sigma_k}, \|s_k\|^2\right\} + \tfrac{1}{2}L_H\|s_k\|^2,
\end{aligned}$$

where to get the second inequality, we also used (9) and Assumption 1. We can bound $\|g_k\|$ as follows:

$$\|g_k\| \leq \|g_k - \nabla\phi(x_k)\| + \|\nabla\phi(x_k) - \nabla\phi(x_k^+)\| + \|\nabla\phi(x_k^+)\| \leq \kappa_g\max\left\{\frac{\mu}{\sigma_k}, \|s_k\|^2\right\} + L\|s_k\| + \|\nabla\phi(x_k^+)\|.$$

Thus finally, we can bound all the terms on the right hand side of (15) in terms of $\|s_k\|^2$ and using the fact that $\eta \in (0,1)$ we can write

$$\begin{aligned}
(1-\eta)\|\nabla\phi(x_k^+)\| &\leq (2\kappa_g + \kappa_H)\max\left\{\frac{\mu}{\sigma_k}, \|s_k\|^2\right\} + (L + L_H + \sigma_k)\|s_k\|^2 \\
&\leq (2\kappa_g + \kappa_H + L + L_H + \sigma_k)\max\left\{\frac{\mu}{\sigma_k}, \|s_k\|^2\right\},
\end{aligned}$$

which is equivalent to (14). □

Lemma 4 (Lower bound on step norm until ε-accuracy is reached) Let Assumption 1 hold. Consider any realization of Algorithm 1. Let ε satisfy

$$\mu \leq \frac{1-\eta}{1+\frac{(1-\frac{\theta}{3})\bar{\sigma}}{\sigma_{\min}}}\varepsilon. \tag{16}$$

Then on each true iteration k such that $\|\nabla\phi(x_k^+)\| \geq \varepsilon$, we have

$$\|s_k\|^2 \geq \frac{\mu}{\sigma_k}.$$

Proof. If iteration k is true and $\|\nabla\phi(x_k^+)\| > \varepsilon$, then by Lemma 3:

$$\max\left\{\|s_k\|^2, \frac{\mu}{\sigma_k}\right\} \geq \frac{1-\eta}{\sigma_k+(1-\frac{\theta}{3})\bar{\sigma}}\|\nabla\phi(x_k^+)\| > \frac{1-\eta}{\sigma_k+(1-\frac{\theta}{3})\bar{\sigma}}\varepsilon,$$

but since

$$\mu \leq \frac{1-\eta}{1+\frac{(1-\frac{\theta}{3})\bar{\sigma}}{\sigma_{\min}}}\varepsilon,$$

so

$$\frac{\mu}{\sigma_k} \leq \frac{1-\eta}{\sigma_k+\frac{(1-\frac{\theta}{3})\bar{\sigma}\sigma_k}{\sigma_{\min}}}\varepsilon \leq \frac{1-\eta}{\sigma_k+(1-\frac{\theta}{3})\bar{\sigma}}\varepsilon.$$

Hence, we must have

$$\|s_k\|^2 > \frac{1-\eta}{\sigma_k+(1-\frac{\theta}{3})\bar{\sigma}}\varepsilon \geq \frac{\mu}{\sigma_k}.$$

□

Corollary 1 (True iteration with large σ_k must be successful) Let Assumption 1 hold. Consider any realization of Algorithm 1. Let ε satisfy (16) and

$$\sigma_k \geq \bar{\sigma} = \frac{2\kappa_g+\kappa_H+L+L_H}{1-\frac{1}{3}\theta},$$

then if iteration k is true and $\|\nabla\phi(x_k^+)\| > \varepsilon$, then iteration k is successful.

Proof. The result is straightforward by applying Lemma 2 and 4. □

Lemma 5 (Minimum improvement achieved by true and successful iterations) Let Assumption 1 hold. Consider any realization of Algorithm 1. Let ε satisfy (16). Then on each true and successful iteration k for which $\|\nabla\phi(x_{k+1})\| > \varepsilon$, we have

$$\begin{aligned}\phi(x_k)-\phi(x_{k+1}) &\geq \frac{\theta}{6}(1-\eta)^{3/2}\frac{\sigma_k}{(\sigma_k+(1-\frac{\theta}{3})\bar{\sigma})^{3/2}}\|\nabla\phi(x_{k+1})\|^{3/2}-e_k-e_k^+-2\varepsilon_f' \\ &\geq \frac{\theta}{6}(1-\eta)^{3/2}\frac{\sigma_{\min}}{(\sigma_k+(1-\frac{\theta}{3})\bar{\sigma})^{3/2}}\|\nabla\phi(x_{k+1})\|^{3/2}-e_k-e_k^+-2\varepsilon_f',\end{aligned}$$

where $\bar{\sigma}$ is defined in (13).

Proof. Combining Lemma 3, 4, inequality (12) from Lemma 1 and the definition of successful iteration in Algorithm 1 we have, for all true and successful iterations k,

$$\begin{aligned}\phi(x_k)-\phi(x_{k+1}) &\geq \frac{\theta}{6}\sigma_k\|s_k\|^3-e_k-e_k^+-2\varepsilon_f' \\ &\geq \frac{\theta}{6}(1-\eta)^{3/2}\frac{\sigma_k}{(\sigma_k+(1-\frac{\theta}{3})\bar{\sigma})^{3/2}}\|\nabla\phi(x_{k+1})\|^{3/2}-e_k-e_k^+-2\varepsilon_f'.\end{aligned}$$

Using the fact that $\sigma_k \geq \sigma_{\min}$, the result follows. □

We can now use these important properties of Algorithm 1 to show that stochastic process that the algorithm generates fits into the framework analyzed in Jin et al. (2021).

4 STOCHASTIC PROPERTIES OF ALGORITHM 1

Algorithm 1 generates a stochastic process. Let M_k denote the collection of random variables $\{\Xi_k,\Xi_k^+,\Xi_k^1,\Xi_k^2\}$, whose realizations are $\{\xi_k,\xi_k^+,\xi_k^1,\xi_k^2\}$. Let $\{\mathscr{F}_k : k\geq 0\}$ denote the filtration generated by $M_0,M_1,\ldots,M_k$. At iteration k, X_k denotes the random iterate, G_k is the random gradient approximation, H_k is the random Hessian approximation, Σ_k is the random model regularization parameter. S_k is the step computed for the random model, $f(X_k,\Xi_k)$ and $f(X_k^+,\Xi_k^+)$ are the random function estimates at the current point and the candidate point, respectively.

Conditioned on X_k and Σ_k, the random quantities G_k and H_k are determined by Ξ_k^1 and Ξ_k^2 respectively. The realization of S_k depends on the realizations of G_k and H_k. The function estimates $f(X_k,\Xi_k)$ and $f(X_k^+,\Xi_k^+)$ are determined by Ξ_k,Ξ_k^+, conditioned on X_k and X_k^+. In summary, the stochastic process $\{(G_k,\mathsf{H}_k,S_k,f(X_k,\Xi_k),f(X_k^+,\Xi_k^+),X_k,\Sigma_k)\}$ generated by the algorithm, with realization $\{(g_k,H_k,s_k,f(x_k,\xi_k),f(x_k^+,\xi_k^+),x_k,\sigma_k)\}$, is adapted to the filtration $\{\mathscr{F}_k : k\geq 0\}$.

We further define $E_k := |f(X_k,\Xi_k)-\phi(X_k)|$ and $E_k^+ := |f(X_k^+,\Xi_k^+)-\phi(X_k^+)|$, with realizations e_k and e_k^+. Let $\Theta_k := \mathbb{1}\{\text{iteration } k \text{ is successful}\}$, and let $I_k := \mathbb{1}\{\text{iteration } k \text{ is true}\}$. The indicator random variables Θ_k and I_k are clearly measurable with respect to the filtration $\mathscr{F}_k$.

The next lemma shows that by construction of the algorithm, the stochastic model m_k at iteration k is "sufficiently accurate" with probability at least $1-\delta_1-\delta_2$.

Lemma 6 The indicator variable

$$\begin{aligned}J_k = \mathbb{1}\Bigg\{ &\|\nabla\phi(X_k)-g(X_k,\Xi_k^1(X_k))\| \leq \kappa_g \max\left\{\frac{\mu}{\Sigma_k},\|S_k\|^2\right\}, \text{ and} \\ &\|(\nabla^2\phi(X_k)-H(X_k,\Xi_k^2(X_k)))S_k\| \leq \kappa_H \max\left\{\frac{\mu}{\Sigma_k},\|S_k\|^2\right\}\Bigg\}\end{aligned}$$

satisfies the following submartingale-like condition

$$\mathbb{P}(J_k = 1 \mid \mathscr{F}_{k-1}) \geq 1-\delta_1-\delta_2.$$

Proof. By the properties of oracles SFO and SSO and the choices of the inputs for them in step 1 of the algorithm, we have:

$$\mathbb{P}\left(\|\nabla\phi(X_k)-g(X_k,\Xi_k^1(X_k))\| \leq \kappa_g\frac{\mu}{\Sigma_k} \mid \mathscr{F}_{k-1}\right) \geq 1-\delta_1, \tag{17}$$

and

$$\mathbb{P}\left(\|(\nabla^2\phi(X_k)-H(X_k,\Xi_k^2(X_k)))\| \leq \kappa_H\sqrt{\frac{\mu}{\Sigma_k}} \mid \mathscr{F}_{k-1}\right) \geq 1-\delta_2. \tag{18}$$

Inequality (17) implies

$$\mathbb{P}\left(\|\nabla\phi(X_k)-g(X_k,\Xi_k^1(X_k))\|\leq\kappa_g\max\left\{\frac{\mu}{\Sigma_k},\|S_k\|^2\right\}\mid\mathscr{F}_{k-1}\right)\geq 1-\delta_1,$$

and inequality (18) implies

$$\mathbb{P}\left(\|(\nabla^2\phi(X_k)-H(X_k,\Xi_k^2(X_k)))S_k\|\leq\kappa_H\max\left\{\frac{\mu}{\Sigma_k},\|S_k\|^2\right\}\mid\mathscr{F}_{k-1}\right)\geq 1-\delta_2.$$

Thus, we conclude that $\mathbb{P}(J_k=1\mid\mathscr{F}_{k-1})\geq 1-\delta_1-\delta_2$ by the union bound. □

Recall Definition 1 and that we denote the event of iteration k being true by indicator random variable I_k. It is crucial for our analysis that $\mathbb{P}(I_k=1\mid\mathscr{F}_{k-1})\geq p>\frac{1}{2}$ for all k. We will later combine Lemma 6 with the properties of SZO for a bound on $\delta_1+\delta_2$ to ensure $p>\frac{1}{2}$.

The iteration complexity of our algorithm is defined as the following stopping time.

Definition 2 (Stopping time) For $\varepsilon>0$, $T_\varepsilon:=\min\left\{k:\left\|\nabla\phi(X_k^+)\right\|\leq\varepsilon\right\}+1$, the iteration complexity of the algorithm for reaching an ε-stationary point. We will refer to T_ε as the *stopping time* of the algorithm.

It is important to note that even if for some iteration k, $\left\|\nabla\phi(X_k^+)\right\|\leq\varepsilon$, this iteration may not be successful and thus $\|\nabla\phi(X_{k+1})\|$ may be greater than ε. This is a consequence of the complexity analysis of cubic regularization methods that measure progress in terms of the gradient at the trial point and not at the current iterate, and thus is not specific to SARC. The stopping time is thus defined as the first time at which the algorithm *computes* a point at which the gradient of ϕ is less than ε.

It is easy to see that T_ε is a *stopping time* of the stochastic process with respect to $\mathscr{F}_k$. Given a level of accuracy ε, we aim to derive a bound on the iterations complexity T_ε with overwhelmingly high probability. In particular, we will show the number of iterations until the stopping time T_ε is a sub-exponential random variable, whose value (with high probability) scales as $O(\varepsilon^{-3/2})$, similarly to the deterministic case. Towards that end, we define stochastic process Z_k to measure the progress towards optimality.

Definition 3 (Measure of Progress) For each $k\geq 0$, let $Z_k\geq 0$ be a random variable measuring the progress of the algorithm at step k: $Z_k=\phi(X_k)-\phi^*$, where ϕ^* is a lower bound of ϕ.

Armed with these definitions, we will be able to state properties of the stochastic process generated by Algorithm 1, which lead to the desired bounds on T_ε. These properties hold under certain conditions on the parameters used by Algorithm 1. We state these conditions here.

Assumption 2 Define $u=\varepsilon_f'-\varepsilon_f$ and $K=C\max\{\frac{1}{\lambda},\frac{\ln(2)}{a}\}$, C is a universal constant and $p=1-\delta_1-\delta_2-\exp\left(-\min\left\{\frac{u^2}{2K^2},\frac{u}{2K}\right\}\right)$.

(a) $\varepsilon_f'>\varepsilon_f$,

(b) $\delta_1+\delta_2$ are chosen sufficiently small so that $p>\frac{1}{2}$,

(c)

$$\varepsilon>\max\left\{\frac{1+\frac{(1-\frac{\theta}{3})\bar{\sigma}}{\sigma_{\min}}}{1-\eta}\mu,\frac{((2-\frac{\theta}{3})\bar{\sigma})}{1-\eta}\left(\frac{24\varepsilon_f'}{(p-\frac{1}{2})\theta\sigma_{\min}}\right)^{\frac{2}{3}}\right\}. \tag{19}$$

Remark 2 Assumption 2 (c) gives a lower bound on the best accuracy the algorithm can achieve given the accuracy parameters related to the stochastic oracles SFO/SSO and SZO. Specifically, ε_f' is lower bounded by ε_f, which is the "irreducible" error of the zeroth-order oracle. We observe that, if $\varepsilon_f'\approx\varepsilon_f$ then the term involving the error of the zeroth-order oracle in the lower bound of ε for SARC is $O(\varepsilon_f^{\frac{2}{3}})$, which is better dependency than those of SASS in Jin et al. (2021) and the stochastic trust region algorithms in Cao et al. (2022), where ε is lower bounded by $O\left(\sqrt{\varepsilon_f}\right)$.

The dependence of ε on μ has a somewhat more complicated interpretation: μ can be chosen arbitrarily by the algorithm, as long as oracles $\mathsf{SFO}/\mathsf{SSO}$ can deliver appropriate accuracy. Recall that in the algorithm, the accuracy input μ_1 for SFO is $\frac{\mu}{\sigma_k}$ and the accuracy input μ_2 for SSO is $\sqrt{\frac{\mu}{\sigma_k}}$. If σ_k is bounded from above by a constant, then essentially ε is proportional to the best accuracy required of SFO during the algorithm procedure and it is proportional to the square of the best accuracy required of SSO during the algorithm procedure. This dependency is the same as in deterministic inexact algorithms as well as in Jin et al. (2021) and the stochastic trust region algorithms in Cao et al. (2022). We will comment on the existence of the upper bound on σ_k after our main complexity result.

The following theorem establishes key properties of the stochastic process generated by Algorithm 1, that are essential for the convergence analysis. Similar properties used in Jin et al. (2021) obtain high probability iteration complexity for a stochastic step search method. To be consistent with the notation in Jin et al. (2021), we define the random variable $A_k := \frac{1}{\Sigma_k}$, with realization $\alpha_k = \frac{1}{\sigma_k}$, and a constant $\bar{\alpha} = \frac{1}{\bar{\sigma}}$.

Theorem 2 Let Assumptions 1 and 2 hold. For $\bar{\alpha} = \frac{1}{\bar{\sigma}}$ and the following non-decreasing function $h : \mathbb{R} \to \mathbb{R}$:

$$h(\alpha) = \frac{\theta}{6}(1-\eta)^{3/2} \frac{\sigma_{\min}}{(\frac{1}{\alpha} + \frac{(1-\frac{\theta}{3})}{\bar{\alpha}})^{3/2}} \varepsilon^{3/2},$$

the following hold for all $k < T_\varepsilon - 1$:

(i) $\mathbb{P}(I_k = 1 \mid \mathscr{F}_{k-1}) \geq p$ for all k. (Conditioning on the past, each iteration is true with probability at least p.)

(ii) If $A_k \leq \bar{\alpha}$ and $I_k = 1$ then $\Theta_k = 1$. (True iterations with sufficiently small α_k are successful.)

(iii) If $I_k \Theta_k = 1$ then $Z_{k+1} \leq Z_k - h(A_k) + 4\varepsilon'_f$. (True, successful iterations make progress.)

(iv) $h(\bar{\alpha}) > \frac{4\varepsilon'_f}{p-\frac{1}{2}}$. (The lower bound of potential progress for an iteration with parameter $\bar{\alpha}$.)

(v) $Z_{k+1} \leq Z_k + 2\varepsilon'_f + E_k + E_k^+$ for all k. (The "damage" at each iteration is bounded.)

Proof. Part (i) follows from the assumptions on p and the definition of the true iteration.

Part (ii) follows directly from Corollary 1.

Part (iii) follows from Lemma 5.

Part (iv) follows from the definitions of $\bar{\alpha}$, $h(\alpha)$, and inequality (19). Specifically, plugging in the definitions of $\bar{\alpha}$ and $h(\cdot)$, one can show that the inequality $h(\bar{\alpha}) \geq \frac{4\varepsilon'_f}{p-\frac{1}{2}}$ is equivalent to $\varepsilon > \frac{((2-\frac{\theta}{3})\bar{\sigma})}{1-\eta} \left(\frac{24\varepsilon'_f}{(p-\frac{1}{2})\theta\sigma_{\min}}\right)^{\frac{2}{3}}$, which holds by Assumption 2.

Part (v) has exactly the same proof as that of Proposition 1 part (v) in (Jin et al. 2021) and is easily derived from the step acceptance condition of Algorithm 1.

□

5 HIGH PROBABILITY ITERATION COMPLEXITY RESULT

In Jin et al. (2021) a high probability bound on T_ε is derived for a stochastic process with properties stated in Theorem 2. Thus, we can simply apply this theorem here. We first observe that

$$\mathbb{E}_\Xi \left[\exp\left\{\tau(E(x) - \mathbb{E}[E(x)])\right\}\right] \leq \exp\left(\frac{\tau^2 \nu^2}{2}\right), \quad \forall \tau \in \left[0, \frac{1}{b}\right],$$

with $\nu = b = K$, where $K = C\max\{\frac{1}{\lambda}, \frac{\ln(2)}{a}\}$, C is a universal constant. This follows from (1) of SZO by applying Proposition 2.7.1 of Vershynin (2018). Another minor modification of the result in Jin et al. (2021) is that it now applies to the event of $T_\varepsilon \leq t+1$ instead of the event of $T_\varepsilon \leq t$, due to the different definitions of the stopping time.

Theorem 3 Suppose Assumptions 1 and 2 hold for Algorithm 1, then we have the following bound on the iteration complexity: for any $s \geq 0$, $\hat{p} \in \left(\frac{1}{2} + \frac{4\varepsilon'_f + s}{c_1 \varepsilon^{3/2}}, p\right)$, and $t \geq \frac{R}{\hat{p} - \frac{1}{2} - \frac{4\varepsilon'_f + s}{c_1 \varepsilon^{3/2}}}$, we have

$$\mathbb{P}\left(T_\varepsilon \leq t+1\right) \geq 1 - \exp\left(-\frac{(p-\hat{p})^2}{2p^2} t\right) - \exp\left(-\min\left\{\frac{s^2 t}{8K^2}, \frac{st}{4K}\right\}\right),$$

where $c_1 = \frac{\theta}{6}(1-\eta)^{3/2} \frac{\sigma_{\min}}{((2-\frac{\theta}{3})\bar{\sigma})^{3/2}}$, $K = C\max\{\frac{1}{\lambda}, \frac{\ln(2)}{a}\}$, C is a universal constant, $R = \frac{\phi(x_0)-\phi^*}{c_1 \varepsilon^{3/2}} + \max\left\{-\frac{\ln \alpha_0 + \ln \bar{\sigma}}{2\ln \gamma}, 0\right\}$, with p and $\bar{\sigma}$ as defined previously.

Remark 3 The following are some remarks about Theorem 3.

1. Theorem 3 shows the iteration complexity of Algorithm 1 is $O(\varepsilon^{-3/2})$ with overwhelmingly high probability, which matches the deterministic counterpart.
2. The SARC algorithm encounters an ε-stationary point in a finite number of iterations with probability 1. This is a direct consequence of the Borel–Cantelli lemma.
3. Since the probabilities of the failure events $\{T_\varepsilon > t+1\}$ are exponentially decaying for all $t \geq \Theta(\varepsilon^{-3/2})$, this implies a complexity bound of $O(\varepsilon^{-3/2})$ in expectation for SARC.

5.1 Upper Bound on σ_k

While the penalty parameters Σ_k form a stochastic process, this process has nice properties. Specifically it is upper bounded by a one-sided geometric random walk. This random walk is analyzed in Jin et al. (2023) and it is shown that for any given number of iterations t, and for γ chosen appropriately dependent on t, $\max_{1 \leq k \leq t}\{\sigma_k\} \leq O(\bar{\sigma})$ with high probability. A consequence of this fact is that with high probability, for Algorithm 1, there exists a lower bound on all of the accuracy requirements μ_1 and μ_2, which are inputs to the oracles SFO and SSO as in (2) and (3). This, in turn, can give rise to a total "sample" complexity bound for Algorithm 1. For examples of such analyses, we refer the reader to Jin et al. (2023).

ACKNOWLEDGMENTS

This work was partially supported by NSF Grants CCF 20-08434, TRIPODS 17-40796 and ONR award N00014-22-1-215. The authors are grateful to the anonymous referees for their reviews that helped us improve the paper.

REFERENCES

Bandeira, A. S., K. Scheinberg, and L. N. Vicente. 2014. “Convergence of Trust-Region Methods based on Probabilistic Models”. *SIAM Journal on Optimization* 24(3):1238–1264.

Bellavia, S., and G. Gurioli. 2022. “Stochastic Analysis of an Adaptive Cubic Regularization Method under Inexact Gradient Evaluations and Dynamic Hessian Accuracy”. *Optimization* 71(1):227–261.

Bellavia, S., G. Gurioli, B. Morini, and P. L. Toint. 2019. “Adaptive Regularization Algorithms with Inexact Evaluations for Nonconvex Optimization”. *SIAM Journal on Optimization* 29(4):2881–2915.

Bellavia, S., G. Gurioli, B. Morini, and P. L. Toint. 2020. “A Stochastic Cubic Regularisation Method with Inexact Function Evaluations and Random Derivatives for Finite Sum Minimisation”. In *Thirty-seventh International Conference on Machine Learning: ICML2020*.

Bellavia, S., G. Gurioli, B. Morini, and P. L. Toint. 2022. “Adaptive Regularization for Nonconvex Optimization Using Inexact Function Values and Randomly Perturbed Derivatives”. *Journal of Complexity* 68:101591.

Berahas, A. S., M. Xie, and B. Zhou. 2023. “A Sequential Quadratic Programming Method with High Probability Complexity Bounds for Nonlinear Equality Constrained Stochastic Optimization”. *arxiv preprint arxiv:2301.00477*.

Blanchet, J., C. Cartis, M. Menickelly, and K. Scheinberg. 2019. “Convergence Rate Analysis of a Stochastic Trust-Region Method via Supermartingales”. *INFORMS journal on optimization* 1(2):92–119.

Cao, L., A. S. Berahas, and K. Scheinberg. 2022. “First-and Second-Order High Probability Complexity Bounds for Trust-Region Methods with Noisy Oracles”. *arXiv preprint arXiv:2205.03667*.

Carmon, Y., and J. Duchi. 2019. “Gradient Descent Finds the Cubic-Regularized Nonconvex Newton Step”. *SIAM Journal on Optimization* 29(3):2146–2178.

Cartis, C., N. Gould, and P. L. Toint. 2011a. “Optimal Newton-type Methods for Nonconvex Smooth Optimization Problems”. Technical Report Optimization Online.

Cartis, C., N. I. M. Gould, and P. L. Toint. 2011b. “Adaptive Cubic Regularisation Methods for Unconstrained Optimization. Part I: Motivation, Convergence and Numerical Results”. *Math. Program.* 127(2):245–295.

Cartis, C., N. I. M. Gould, and P. L. Toint. 2011c. “Adaptive Cubic Regularisation Methods for Unconstrained Optimization. Part II: Worst-case Function- and Derivative-Evaluation Complexity”. *Math. Program.* 130(2):295–319.

Cartis, C., and K. Scheinberg. 2018. “Global Convergence Rate Analysis of Unconstrained Optimization Methods Based on Probabilistic Models”. *Mathematical Programming* 169(2):337–375.

Chen, R., M. Menickelly, and K. Scheinberg. 2018. “Stochastic Optimization Using a Trust-Region Method and Random Models”. *Mathematical Programming* 169(2):447–487.

Gratton, S., C. W. Royer, L. N. Vicente, and Z. Zhang. 2018. “Complexity and Global Rates of Trust-Region Methods based on Probabilistic Models”. *IMA Journal of Numerical Analysis* 38(3):1579–1597.

Jin, B., K. Scheinberg, and M. Xie. 2021. “High Probability Complexity Bounds for Adaptive Step Search Based on Stochastic Oracles”. *arXiv preprint arXiv:2106.06454*.

Jin, B., K. Scheinberg, and M. Xie. 2023. “Sample Complexity Analysis for Adaptive Optimization Algorithms with Stochastic Oracles”. *arXiv preprint arXiv:2303.06838*.

Kohler, J. M., and A. Lucchi. 2017. “Sub-sampled Cubic Regularization for Non-convex Optimization”. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, Volume 70 of *Proceedings of Machine Learning Research*: PMLR.

Liu, L., X. Liu, C.-J. Hsieh, and D. Tao. 2018. “Stochastic Second-Order Methods for Non-Convex Optimization with Inexact Hessian and Gradient”. *arXiv preprint arXiv:1809.09853*.

Menickelly, M., S. M. Wild, and M. Xie. 2023. “A Stochastic Quasi-Newton Method in the Absence of Common Random Numbers”. *arXiv preprint arXiv:2302.09128*.

Paquette, C., and K. Scheinberg. 2020. “A Stochastic Line Search Method with Expected Complexity Analysis”. *SIAM Journal on Optimization* 30(1):349–376.

Park, S., S. H. Jung, and P. M. Pardalos. 2020. “Combining Stochastic Adaptive Cubic Regularization with Negative Curvature for Nonconvex Optimization”. *Journal of Optimization Theory and Applications* 184(3):953–971.

Tripuraneni, N., M. Stern, C. Jin, J. Regier, and M. I. Jordan. 2018. “Stochastic Cubic Regularization for Fast Nonconvex Optimization”. In *Advances in Neural Information Processing Systems*, 2899–2908.

Vershynin, R. 2018. *High-Dimensional Probability: An Introduction with Applications in Data Science*, Volume 47. Cambridge University Press.

Wang, Z., Y. Zhou, Y. Liang, and G. Lan. 2019. “A Note on Inexact Gradient and Hessian Conditions for Cubic Regularized Newton’s Method”. *Operations Research Letters* 47(2):146–149.

AUTHOR BIOGRAPHIES

KATYA SCHEINBERG is a Professor at the School of Operations Research and Information Engineering at Cornell University. Her main research areas are related to developing computationally efficient algorithms and their theoretical analysis for various problems in continuous optimization. She is an Informs Fellow, a recipient of the Lagrange Prize from SIAM and MOS, the Farkas Prize from Informs Optimization Society and the Outstanding Simulation Publication award from Informs Simulation Society. Katya is currently the editor-in-chief of Mathematics of Operations Research, and co-editor of Mathematical Programming. She served as the Chair of SIAM Activity Group on Optimization from 2020 until 2022. Her email address is ks2375@cornell.edu and her homepage is https://scheinberg.engineering.cornell.edu/

MIAOLAN XIE is a 5th-year Ph.D. Candidate in the Operations Research and Information Engineering Department at Cornell University, advised by Katya Scheinberg. Her research interests lie in the intersection of stochastic optimization and data science. In particular, she is interested in designing user-friendly and adaptive optimization algorithms that operate under realistic assumptions while still having good theoretical guarantees with tools in optimization, stochastic processes, and statistics. Miaolan received her Bachelor’s and Master’s degrees in Mathematics from the University of Waterloo. Before joining Cornell, she worked in Alibaba’s supply chain group as a data scientist. Her email address is mx229@cornell.edu and her homepage is https://miaolan.github.io/