OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Stochastic Adaptive Regularization Method with Cubics: A High Probability Complexity Bound

Katya Scheinberg KATYAS@CORNELL.EDU
Miaolan Xie MX229@CORNELL.EDU
Cornell University, 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 adaptive stochastic methods, such as trust region and line search. These oracles capture many settings, such as expected risk minimization, stochastic zeroth order 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 adaptive methods.

Keywords: nonlinear optimization, stochastic optimization, cubic regularization methods, high probability, complexity bound, stochastic oracles, random models.

1. Introduction

We consider the unconstrained optimization problem

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

where the objective function ϕ is assumed to be sufficiently smooth, but whose value and gradients are not computable exactly. Instead we assume that we have access to stochastic zeroth, first and second order oracles that, given x, produce random estimates of $\phi(x)$, $\nabla\phi(x)$ and $\nabla^2\phi(x)$ respectively. These oracles are as follows.

Stochastic zeroth order oracle ($\mathsf{SZO}(\epsilon_f,\nu,b)$). Given a point x, the oracle computes $f(x,\Xi(x))$, a (random) estimate of the function value $\phi(x)$. $\Xi(x)$ is a random variable (whose distribution is defined, given x, ϵ_f,ν and b). We assume the absolute value of the estimation error $E(x) = |f(x,\Xi(x)) - \phi(x)|$ (we omit the dependence on Ξ for brevity) to be a "one-sided" sub-exponential-like random variable[1] with parameters (ν, b), whose mean is bounded by some constant $\epsilon_f > 0$. Specifically,

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

[1] This is a weaker requirement than assuming $E(x)$ to be sub-exponential and is sufficient for our purposes.

We view x as the input to the oracle, $f(x, \Xi(x))$ as the output and the values (ϵ_f, ν, b) as values intrinsic to the oracle. To simplify notation, we use $f(x)$ instead of $f(x, \Xi(x))$ in the paper.

Stochastic first order oracle ($\mathsf{SFO}(\kappa_g)$. Given a point x and constants $M_1 > 0$, $\delta_1 \in [0, \frac{1}{2})$, the oracle computes $g(x, \Xi^1(x))$, a (random) estimate of the gradient $\nabla\phi(x)$, such that

$$\mathbb{P}(\|\nabla\phi(x) - g(x, \Xi^1(x))\| \leq \kappa_g M_1) \geq 1 - \delta_1.$$

The distribution of $\Xi^1(x)$ is defined, given x, M_1, δ_1 and κ_g. Here we view x, M_1 and δ_1 as inputs input to the oracle, while κ_g is intrinsic to the oracle. Similarly we define

Stochastic second order oracle ($\mathsf{SSO}(\kappa_H)$). Given a point x and constants $M_2 > 0$, $\delta_2 \in [0, \frac{1}{2})$, the oracle computes $H(x, \Xi^2(x))$, a (random) estimate of the Hessian $\nabla^2\phi(x)$, such that

$$\mathbb{P}(\|\nabla^2\phi(x) - H(x, \Xi^2(x))\| \leq \kappa_H M_2) \geq 1 - \delta_2.$$ [2]

The distribution of $\Xi^2(x)$ is defined, given x, M_2, δ_2 and κ_H. Here we view x, M_2 and δ_2 as inputs to the oracle, while κ_H is intrinsic to the oracle.

Similar oracles have been introduced and used recently in [17] and [8] for the analysis of high probability iteration complexity bounds for a stochastic adaptive step search (line search) method and a stochastic trust region method, respectively. Specific examples and settings under which the assumptions are satisfied can be found in [17] and [8].

Related work. In [17], the authors derived high probability iteration complexity results for a step search method that relies on stochastic, possibly biased zeroth and first order oracles. That paper extends results in [11], [6] and [21] in two key ways - allowing the use of possibly biased oracles and the derivation of high probability complexity bound.

On the other hand, [8] provides a similar extension to biased oracles for high probability iteration complexity for stochastic trust region method, compared to previous work in [1], [16], and [7, 15]

Adaptive regularization with cubics (ARC) method is known to have optimal complexity for finding ϵ stationary points of deterministic smooth functions [10, 13]. In that sense the method is superior to deterministic line search and trust region methods. In [11] the expected complexity analysis is derived for the case of stochastic first and second order oracles, but under the assumption that the values $\phi(x)$ can be computed exactly. Moreover, the definitions of the first and second order oracles imposed stronger conditions on those oracles than we require here.

There are many other variants of adaptive cubic regularization methods under various assumptions and requirements on the function value, gradient, and Hessian estimates. For example, in [2, 4, 5], bounds on expected complexity are provided under the assumption that function estimates are either exact or have a deterministically bounded error while the gradient and Hessian estimates are probabilistic. In [3, 14, 19, 20, 22, 23], the function, gradient, and Hessian estimates are assumed to have deterministically bounded errors whose magnitude can be adaptively changed in the course of the algorithm.

Our contributions. In this work we provide the first 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 accuracy is bounded by an adaptive quantity with sufficiently high probability, but otherwise can be arbitrarily bad. To the best of our knowledge, our

[2] The norm on the matrix is the operator norm.

Algorithm 1: Stochastic Adaptive Regularization with Cubics (SARC)

```
Initialization: Choose parameters $\gamma \in (0,1)$, $\theta \in (0,1)$, $\delta \in [0, \frac{1}{2})$ $\sigma_{\min} > 0$ and $\kappa_\theta \in (0,1)$.
    Pick initial $x_0$, $\kappa_g > 0$, $\kappa_H > 0$, $\bar{\epsilon} \geq 0$, $\epsilon'_f > 0$ and $\sigma_0 > \sigma_{\min}$.
Repeat for $k = 0, 1, \ldots$
  1. Compute a model trial step $s_k$: Compute gradient and Hessian estimations $g_k$, $H_k$ and a
     trial step $s_k$ that satisfies (6) and (7) via Algorithm 2, with parameters $\kappa_\theta$, $\bar{\epsilon}$, $\delta$, $\sigma_k$ 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 SZO, and set
         $\rho_k = \frac{f(x_k) - f(x_k^+) + 2\epsilon'_f}{f(x_k) - m_k(x_k^+)},$   (2)
     where
         $m_k(x_k^+) = f(x_k) + s_k^T g_k + \frac{1}{2} s_k^T H_k s_k + \frac{\sigma_k}{3}\|s_k\|^3.$   (3)
  3. Update the iterate: Set
         $x^{k+1} = \begin{cases} x_k^+ & \text{if} \quad \rho_k \geq \theta \qquad [k \text{ successful}] \\ x_k & \text{otherwise} \quad [k \text{ unsuccessful}] \end{cases}$   (4)
  4. Update the regularization parameter $\sigma_k$: Set
         $\sigma_{k+1} = \begin{cases} \max\{\gamma\sigma_k, \sigma_{\min}\} & \text{if} \quad \rho_k \geq \theta \\ \frac{1}{\gamma}\sigma_k & \text{otherwise.} \end{cases}$   (5)
```

work is the first to derive a complexity bound 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 [17] and [8] to the framework in [11]. 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.

For space reasons, all proofs are in the Appendix.

2. Adaptive regularization method with cubics (ARC) with probabilistic second-order models

The Stochastic Adaptive Regularization with Cubics (SARC) algorithm is presented below as Algorithm 1, with its subroutine presented as Algorithm 2.

Remark 1 *Algorithm 2 will always terminate in a finite number of iterations, if $\bar{\epsilon} > 0$. In particular, M_1 takes at most $\log\left(\frac{M}{\alpha_k \bar{\epsilon}}\right) + 1$ iterations to be less or equal to the required precision lower bound*

Algorithm 2: Subroutine for computing s_k

```
Input: Oracles SFO($\kappa_g$) and SSO($\kappa_H$), $x_k$, constants $\kappa_\theta$, $\bar{\epsilon}$, $\delta$, $\sigma_k$, $M_{1,0} = M_{2,0} = M > 0$.
Repeat for $i = 0, 1, \dots$
    1. Generate the model $m_{k,i}$: Generate $g_{k,i} = g(x_k, \xi^1_{k,i})$ and $H_{k,i} = H(x_k, \xi^2_{k,i})$ using
       SFO($\kappa_g$) and SSO($\kappa_H$) with $(M_{1,i}, \frac{\delta}{2})$ and $(M_{2,i}, \frac{\delta}{2})$ as inputs respectively. Form the model
           $m_{k,i}(x_k + s) = s^T g_{k,i} + \frac{1}{2} s^T H_{k,i} s + \frac{\sigma_k}{3} \|s\|^3.$
    2. Compute $s_{k,i}$: Compute an approximate minimizer $s_{k,i}$ of $m_{k,i}$ that satisfies (6) and (7)
       with parameter $\kappa_\theta$, using for example algorithms in [9, 12, 14].
           $(s_{k,i})^T g_{k,i} + (s_{k,i})^T H_{k,i} s_{k,i} + \sigma_k \|s_{k,i}\|^3 = 0$ and $(s_{k,i})^T H_{k,i} s_{k,i} + \sigma_k \|s_{k,i}\|^3 \geq 0$    (6)
       and
           $\|\nabla m_{k,i}(x_k + s_{k,i})\| \leq \kappa_\theta \min\{1, \|s_{k,i}\|\} \|g_{k,i}\|,$    (7)
    3a. Successful step: If $M_{1,i} \leq \max\left\{\|s_{k,i}\|^2, \frac{\bar{\epsilon}}{\sigma_k}\right\}$ and $M_{2,i} \leq \max\left\{\|s_{k,i}\|, \frac{\bar{\epsilon}}{\sigma_k \|s_{k,i}\|}\right\}$, end
        procedure and return $g_{k,i}$, $H_{k,i}$ and $s_{k,i}$.
    3b. Unsuccessful step: Otherwise, set $M_{1,i+1} \leftarrow \max\{\frac{M_{1,i}}{2}, \frac{\bar{\epsilon}}{\sigma_k}\}$,
        $M_{2,i+1} \leftarrow \max\{\frac{M_{2,i}}{2}, \sqrt{\frac{\bar{\epsilon}}{\sigma_k}}\}$ and go to step 1.
```

$\alpha_k \bar{\epsilon}$, and M_2 takes at most $\log\left(\frac{M}{\sqrt{\alpha_k \bar{\epsilon}}}\right) + 1$ *iterations to be less or equal to the required precision lower bound $\sqrt{\alpha_k \bar{\epsilon}}$. Note $\sqrt{\alpha_k \bar{\epsilon}}$ is the lower bound for* $\max\left\{\|s_{k,i}\|, \frac{\bar{\epsilon}}{\sigma_k \|s_{k,i}\|}\right\}$ *for any $\|s_{k,i}\|$.*

Remark 2 *Step 2 in Algorithm 2 can be replaced by the simple requirement that $s_{k,i}$ is an exact global optimizer of $m_{k,i}$, which is stronger than the requirement listed in Step 2.*

3. Elements of stochastic analysis

Algorithms 1 and 2 together generates a random stochastic process. Let i_k be the total number of iterations the loop in Algorithm 2 executes during iteration k of Algorithm 1. Let M_k denote $\left\{\Xi_k, \Xi_k^+, \Xi^1_{k,i_k}, \Xi^2_{k,i_k}\right\}$, whose realizations are $\left\{\xi_k, \xi_k^+, \xi^1_{k,i_k}, \xi^2_{k,i_k}\right\}$, where Ξ_k, Ξ_k^+ dictate the randomness of the function estimations at x_k and x_k^+.

A stochastic process $\left\{\left(G_k, \mathcal{H}_k, S_k, f(X_k, \Xi_k), f(X_k^+, \Xi_k^+), X_k, A_k = \frac{1}{\Sigma_k}\right)\right\}$ is generated by the algorithm, with realization $\left\{\left(g_k, H_k, s_k, f(x_k, \xi_k), f(x_k^+, \xi_k^+), x_k, \alpha_k = \frac{1}{\sigma_k}\right)\right\}$. It is adapted to $\{\mathcal{F}_k : k \geq 0\}$, where $\mathcal{F}_k = \sigma(M_0, M_1, \dots, M_k)$. At iteration k, X_k denotes the (random) iterate, G_k is the gradient apprpximation, $\mathcal{H}_k$ is the Hessian approximation, $A_k = \frac{1}{\Sigma_k}$ is the inverse of the model regularization parameter. S_k is the step computed by the Algorithm 2. $f(X_k, \Xi_k)$ and $f(X_k^+, \Xi_k^+)$ are the function estimates at the current point and the candidate point, respectively. Note that conditioned on X_k and M_{1,i_k}, G_k is dictated by Ξ^1_{k,i_k}. Similarly, conditioned on X_k and M_{2,i_k}, $\mathcal{H}_k$ is dictated by Ξ^2_{k,i_k}. The function estimates are dictated by Ξ_k, Ξ_k^+ in the zeroth order oracle.

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 $\Theta_k := \mathbb{1}\,\{\text{iteration } k \text{ is successful}\}$. The random variable Θ_k is clearly measurable with respect to the filtration $\mathcal{F}_k$.

By the construction of Algorithm 2, the stochastic model m_k at iteration k is "sufficiently accurate" with probability at least $1 - \delta$. Specifically, we have the following lemma.

Lemma 3 *By construction of Algorithm 2, given iterate x_k at iteration k, the indicator variable*

$$J_k\left(\Xi_k^1(x_k), \Xi_k^2(x_k)\right) = \mathbb{1}\{\|\nabla\phi(x_k) - g(x_k, \Xi_k^1(x_k))\| \le \kappa_g \max\left\{\|S_k\|^2, \alpha_k\bar{\epsilon}\right\},\ \textit{and}$$
$$\|(\nabla^2\phi(x_k) - H(x_k, \Xi_k^2(x_k)))S_k\| \le \kappa_H \max\left\{\|S_k\|^2, \alpha_k\bar{\epsilon}\right\}\ \}$$

satisfies the following submartingale-like condition

$$\mathbb{P}(J_k = 1 \mid \mathcal{F}_{k-1}) \ge 1 - \delta.$$

A key concept that will be used in the analysis is the concept of a *true iteration*.

Definition 4 (True iteration) *We say that iteration k is **true** if*

$$\|\nabla\phi(x_k) - g_k\| \le \kappa_g \max\left\{\|s_k\|^2, \alpha_k\bar{\epsilon}\right\},\ \|(\nabla^2\phi(x_k) - H_k)s_k\| \le \kappa_H \max\left\{\|s_k\|^2, \alpha_k\bar{\epsilon}\right\} \tag{8}$$
$$\textit{and}\ |f(x_k) - \phi(x_k)| + |f(x_k^+) - \phi(x_k^+)| \le 2\epsilon_f', \tag{9}$$

*and is **false** otherwise. I_k is the indicator random variable that iteration k is true.*

The stopping time of the algorithm is defined as follows.

Definition 5 (Stopping time)
For $\epsilon > 0$, $T_\epsilon := \min\left\{k : \left\|\nabla\phi(X_k^+)\right\| \le \epsilon\right\} + 1$, the iteration complexity of the algorithm for reaching a ϵ-stationary point. We will refer to T_ϵ as the stopping time *of the algorithm.*

It is easy to see that T_ϵ is a *stopping time* of the stochastic process with respect to $\mathcal{F}_k$. Given a level of accuracy ϵ, we aim to derive a bound on the iterations complexity T_ϵ with high probability. In particular, we will show the number of iterations until the stopping time T_ϵ is a sub-exponential random variable itself. The random variable Z_k is defined to measure the progress towards optimality.

Definition 6 (Measure of Progress) *For each $k \ge 0$, let $Z_k \ge 0$ 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.*

We make the following assumptions on the nonconvex objective ϕ and the algorithm input ϵ_f'.

Assumption 1 *ϕ is bounded below by a constant ϕ^*, $\phi \in \mathcal{C}^2(\mathbb{R}^n)$ and has globally L-Lipschitz continuous gradient and L_H-Lipschitz continuous Hessian.*

Assumption 2 $\epsilon_f' \ge \epsilon_f$.

Since the stochastic oracles are noisy and possibly biased, the algorithm cannot be expected to converge to a stationary point. Instead, it can only converge to an ϵ-stationary point where ϵ is dictated by $\bar{\epsilon}$ and ϵ'_f as follows.

Inequality 1 (Lower bound on the size of convergence neighborhood)

$$\epsilon > \max\left\{ \frac{1+\frac{\kappa_s}{\sigma_{\min}}}{1-\kappa_\theta}\bar{\epsilon}, \left(\frac{4\epsilon'_f}{\kappa_h(p-\frac{1}{2})}\right)^{\frac{2}{3}} \right\},$$

where $\kappa_h = \frac{\theta}{6}(1-\kappa_\theta)^{3/2}\frac{\sigma_{\min}}{(\sigma_c+\kappa_s)^{3/2}}$, $\sigma_c = \frac{2\kappa_g+\kappa_H+L+L_H}{1-\frac{1}{3}\theta}$, $\kappa_s = 2\kappa_g + \kappa_H + L + L_H$, and $p = 1-\delta-\exp\left(-\min\left\{\frac{u^2}{2\nu^2}, \frac{u}{2b}\right\}\right)$, where $u = \inf_x \left\{\epsilon'_f - \mathbb{E}[E(x)]\right\}$.

The stochastic process generated by the algorithm has the following properties.

Proposition 7 (Properties of the stochastic process) *Let Assumptions 1 and 2 hold, and suppose ϵ satisfies Inequality 1. For $\bar{\alpha} = \frac{1}{\sigma_c} > 0$ and the following non-decreasing function $h : \mathbb{R} \to \mathbb{R}$:*

$$h(\alpha) = \frac{\theta}{6}(1-\kappa_\theta)^{3/2}\frac{\sigma_{\min}}{(\frac{1}{\alpha}+\kappa_s)^{3/2}}\epsilon^{3/2},$$

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

(i) $h(\bar{\alpha}) > \frac{4\epsilon'_f}{p-\frac{1}{2}}$. (The lower bound of potential progress an iteration with step size $\bar{\alpha}$ can make.)

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

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

(iv) If $A_k \leq \bar{\alpha}$ and $I_k = 1$ then $\Theta_k = 1$. (Small and true iterations are also successful.)

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

4. High probability iteration complexity

We can now use essentially the same analysis as in [18] to obtain a high probability iteration bound. The only difference remains is the failure event is now $t+1 < T_\epsilon$ instead of $t < T_\epsilon$ with the new stopping time.

Theorem 8 *Suppose Assumptions 1 and 2 hold for Algorithm 1, and the Inequality 1 on ϵ is satisfied. Then we have the following bound on the iteration complexity: For any $s \geq 0$, $\hat{p} \in \left(\frac{1}{2} + \frac{4\epsilon'_f+s}{\kappa_h\epsilon^{3/2}}, p\right)$, and $t \geq \frac{R}{\hat{p}-\frac{1}{2}-\frac{4\epsilon'_f+s}{\kappa_h\epsilon^{3/2}}}$, we have*

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

where $R = \frac{\phi(x_0)-\phi^*}{\kappa_h\epsilon^{3/2}} + \max\left\{-\frac{\ln\alpha_0+\ln\sigma_c}{2\ln\gamma}, 0\right\}$, with κ_h, p, σ_c and κ_s as defined previously.

Remark 9

1. *Theorem 8 essentially shows the iteration complexity of the algorithm is $O(\epsilon^{-3/2})$ with overwhelmingly high probability, which matches its deterministic counterpart.*

2. *If $\bar{\epsilon} = 0$, the stopping time can be also defined as:*

$$T_\epsilon = \min \{k : \|\nabla \phi(X_k)\| \leq \epsilon\}.$$

3. *By [17], with an appropriately chosen γ with respect to p, $\alpha_k = \frac{1}{\sigma_k}$ will remain sufficiently large with high probability. As a result, the accuracy requirements for the first and second order oracles will remain reasonable to satisfy with high probability.*

References

[1] Afonso S Bandeira, Katya Scheinberg, and Luis Nunes Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization*, 24(3):1238–1264, 2014.

[2] Stefania Bellavia and Gianmarco Gurioli. Stochastic analysis of an adaptive cubic regularization method under inexact gradient evaluations and dynamic hessian accuracy. *Optimization*, 71(1):227–261, 2022.

[3] Stefania Bellavia, Gianmarco Gurioli, Benedetta Morini, and Philippe L Toint. Adaptive regularization algorithms with inexact evaluations for nonconvex optimization. *SIAM Journal on Optimization*, 29(4):2881–2915, 2019.

[4] Stefania Bellavia, Gianmarco Gurioli, Benedetta Morini, and Philippe L Toint. 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*, 2020.

[5] Stefania Bellavia, Gianmarco Gurioli, Benedetta Morini, and Ph L Toint. Adaptive regularization for nonconvex optimization using inexact function values and randomly perturbed derivatives. *Journal of Complexity*, 68:101591, 2022.

[6] Albert S Berahas, Liyuan Cao, and Katya Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *SIAM Journal on Optimization*, 31(2):1489–1518, 2021.

[7] Jose Blanchet, Coralia Cartis, Matt Menickelly, and Katya Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS journal on optimization*, 1(2):92–119, 2019.

[8] Liyuan Cao, Albert S Berahas, and Katya Scheinberg. First-and second-order high probability complexity bounds for trust-region methods with noisy oracles. *arXiv preprint arXiv:2205.03667*, 2022.

[9] Yair Carmon and John Duchi. Gradient descent finds the cubic-regularized nonconvex newton step. *SIAM Journal on Optimization*, 29(3):2146–2178, 2019.

[10] C. Cartis, N. Gould, and Philippe L. Toint. Optimal Newton-type methods for nonconvex smooth optimization problems. Technical Report Optimization Online, 2011.

[11] Coralia Cartis and Katya Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2):337–375, 2018.

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

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

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

[15] Ruobing Chen, Matt Menickelly, and Katya Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169(2):447–487, 2018.

[16] Serge Gratton, Clément W Royer, Luís N Vicente, and Zaikun Zhang. Complexity and global rates of trust-region methods based on probabilistic models. *IMA Journal of Numerical Analysis*, 38(3):1579–1597, 2018.

[17] Billy Jin, Katya Scheinberg, and Miaolan Xie. High probability step size lower bound for adaptive stochastic optimization.

[18] Billy Jin, Katya Scheinberg, and Miaolan Xie. High probability complexity bounds for adaptive step search based on stochastic oracles, 2021. URL https://arxiv.org/abs/2106.06454.

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

[20] Liu Liu, Xuanqing Liu, Cho-Jui Hsieh, and Dacheng Tao. Stochastic second-order methods for non-convex optimization with inexact hessian and gradient. *arXiv preprint arXiv:1809.09853*, 2018.

[21] Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

[22] Seonho Park, Seung Hyun Jung, and Panos M Pardalos. Combining stochastic adaptive cubic regularization with negative curvature for nonconvex optimization. *Journal of Optimization Theory and Applications*, 184(3):953–971, 2020.

[23] Zhe Wang, Yi Zhou, Yingbin Liang, and Guanghui Lan. A note on inexact gradient and hessian conditions for cubic regularized newton's method. *Operations Research Letters*, 47 (2):146–149, 2019.

Appendix A. Proof of Lemma 3

Proof We show $\mathbb{P}(J_k = 1 \mid \mathcal{F}_{k-1}) \geq 1 - \delta$ by showing

$$\mathbb{P}\left(\|\nabla\phi(x_k) - g(x_k, \Xi_k^1(x_k))\| \leq \kappa_g \max\left\{\|S_k\|^2, \alpha_k\bar{\epsilon}\right\} \mid \mathcal{F}_{k-1}\right) \geq 1 - \frac{\delta}{2}$$

and

$$\mathbb{P}\left(\|(\nabla^2\phi(x_k) - H(x_k, \Xi_k^2(x_k)))S_k\| \leq \kappa_H \max\left\{\|S_k\|^2, \alpha_k\bar{\epsilon}\right\} \mid \mathcal{F}_{k-1}\right) \geq 1 - \frac{\delta}{2}.$$

By definition of the oracles and construction of Algorithm 2 we have:

$$\mathbb{P}\left(\|\nabla\phi(x_k) - g(x_k, \Xi_k^1(x_k))\| \leq \kappa_g \max\left\{\|S_k\|^2, \alpha_k\bar{\epsilon}\right\}\right) \geq 1 - \frac{\delta}{2}, \tag{10}$$

and

$$\mathbb{P}\left(\|(\nabla^2\phi(x_k) - H(x_k, \Xi_k^2(x_k)))\| \leq \kappa_H \max\left\{\|S_k\|, \frac{\alpha_k\bar{\epsilon}}{\|S_k\|}\right\}\right) \geq 1 - \frac{\delta}{2}. \tag{11}$$

Inequality (11) implies

$$\mathbb{P}\left(\|(\nabla^2\phi(x_k) - H(x_k, \Xi_k^2(x_k)))S_k\| \leq \kappa_H \max\left\{\|S_k\|^2, \alpha_k\bar{\epsilon}\right\}\right) \geq 1 - \frac{\delta}{2}. \tag{12}$$

Inequality (10) and (12) together gives

$$\mathbb{P}(J_k = 1) \geq 1 - \delta.$$

Using the fact that conditioning on a set of x_k's, $(\Xi_k^1(x_k), \Xi_k^2(x_k))$'s are independent of each other, we obtain

$$\mathbb{P}(J_k = 1 \mid \mathcal{F}_{k-1}) = \mathbb{P}(J_k = 1) \geq 1 - \delta.$$

■

Appendix B. Lemmas and the proof of Proposition 7

The following lemmas provide useful properties of the stochastic process generated by Algorithm 1, which is essential for the convergence analysis.

Lemma 10 *Consider any realization of Algorithm 1, on each iteration k we have*

$$f(x_k) - m_k(x_k^+) \geq \frac{1}{6}\sigma_k\|s_k\|^3. \tag{13}$$

On every successful iteration k, we have

$$f(x_k) - f(x^{k+1}) \geq \frac{\theta}{6}\sigma_k\|s_k\|^3 - 2\epsilon_f', \tag{14}$$

or

$$\phi(x_k) - \phi(x^{k+1}) \geq \frac{\theta}{6}\sigma_k\|s_k\|^3 - e_k - e_k^+ - 2\epsilon_f'. \tag{15}$$

Proof The proof is similar to the proof of Lemma 3.3 in [12]. Clearly, (14) follows from (13) and the sufficient decrease condition (2)-(4):

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

and (15) follows from the definition of e_k and e_k^+.

It remains to prove (13). Combining the first condition on step s_k in (6), with the model expression for $s = s_k$, we can write

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

The second condition on s_k in (6) implies $(s_k)^T H_k s_k \geq -\sigma_k \|s_k\|^3$. Together with the above equation, we obtain (13). ■

Lemma 11 *Let Assumption 1 hold. For any realization of Algorithm 1, if iteration k is true (i.e., $I_k = 1$), and if*

$$\sigma_k \geq \sigma_c = \frac{2\kappa_g + \kappa_H + L + L_H}{1 - \frac{1}{3}\theta}, \tag{16}$$

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

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 (2), that

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

Notice that:

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

The second last inequality is by definition of e_k and e_k^+, and the last inequality is by definition of the iteration being true.

Taylor expansion and triangle inequalities give, for some $\xi_k \in [x_k, x_k^+]$,

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

where the last inequality follows from the fact that the iteration is true and hence (8) holds: $\|\nabla\phi(x_k)-g_k\| \le \kappa_g \max\left\{\|s_k\|^2, \frac{\bar{\epsilon}}{\sigma_k}\right\}$ and $\|(\nabla^2\phi(x_k)-H_k)s_k\| \le \kappa_H \max\left\{\|s_k\|^2, \frac{\bar{\epsilon}}{\sigma_k}\right\}$, and from Assumption 1. So as long as $\|s_k\|^2 \ge \frac{\bar{\epsilon}}{\sigma_k}$, we have

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

which together with (13) give that $1-\rho_k \le 1-\theta$ when σ_k satisfies (20).

■

Note that for the above lemma to hold σ_c does not have to depend on L. However, in what follows we will need another condition on σ_c, which will involve L; hence for simplicity of notation we introduced σ_c above to satisfy all necessary bounds.

Lemma 12 *Let Assumption 1 hold. Consider any realization of Algorithm 1. On each true iteration k we have*

$$\max\left\{\|s_k\|^2, \frac{\bar{\epsilon}}{\sigma_k}\right\} \ge \frac{1-\kappa_\theta}{\sigma_k+\kappa_s}\|\nabla\phi(x_k^+)\|, \tag{17}$$

where $\kappa_s = 2\kappa_g + \kappa_H + L + L_H$.

Proof

Triangle inequality, equality $\nabla m_k(x_k+s) = g_k + H_k s + \sigma_k\|s\|s$ and condition (7) on s_k together give

$$\begin{aligned}\|\nabla\phi(x_k^+)\| &\le \|\nabla\phi(x_k^+) - \nabla m_k(x_k^+)\| + \|\nabla m_k(x_k^+)\| \\ &\le \|\nabla\phi(x_k^+) - g_k - H_k s_k\| + \sigma_k\|s_k\|^2 + \kappa_\theta \min\{1, \|s_k\|\}\,\|g_k\|.\end{aligned} \tag{18}$$

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

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

where to get the second inequality, we also used (8) and Assumption 1.

We can bound $\|g_k\|$ as follows

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

Thus finally, we can bound all the terms on the right hand side of (18) in terms of $\|s_k\|^2$ and using the fact that $\kappa_\theta \in (0,1)$ we can write

$$\begin{aligned}(1-\kappa_\theta)\|\nabla\phi(x_k^+)\| &\le (2\kappa_g+\kappa_H)\max\left\{\frac{\bar\epsilon}{\sigma_k}, \|s_k\|^2\right\} + (L+L_H+\sigma_k)\|s_k\|^2 \\ &\le (2\kappa_g+\kappa_H)\max\left\{\frac{\bar\epsilon}{\sigma_k}, \|s_k\|^2\right\} + (L+L_H+\sigma_k)\max\left\{\frac{\bar\epsilon}{\sigma_k}, \|s_k\|^2\right\},\end{aligned}$$

which is equivalent to (17). ■

Lemma 13 *Let Assumption 1 hold. Consider any realization of Algorithm 1. Let $\bar\epsilon$ satisfy*

$$\bar\epsilon \le \frac{1-\kappa_\theta}{1+\frac{\kappa_s}{\sigma_{\min}}}\epsilon. \tag{19}$$

Then on each true iteration k, with $k < T_\epsilon - 1$ we have

$$\|s_k\|^2 \ge \frac{\bar\epsilon}{\sigma_k}.$$

Proof

If iteration k is true and $\|\nabla\phi(x_k^+)\| > \epsilon$ (since $k < T_\epsilon - 1$) then by Lemma 12:

$$\max\left\{\|s_k\|^2, \frac{\bar\epsilon}{\sigma_k}\right\} \ge \frac{1-\kappa_\theta}{\sigma_k+\kappa_s}\|\nabla\phi(x_k^+)\| > \frac{1-\kappa_\theta}{\sigma_k+\kappa_s}\epsilon,$$

but since

$$\bar\epsilon \le \frac{1-\kappa_\theta}{1+\frac{\kappa_s}{\sigma_{\min}}}\epsilon,$$

so

$$\frac{\bar\epsilon}{\sigma_k} \le \frac{1-\kappa_\theta}{\sigma_k+\frac{\kappa_s\sigma_k}{\sigma_{\min}}}\epsilon \le \frac{1-\kappa_\theta}{\sigma_k+\kappa_s}\epsilon.$$

Hence, we must have

$$\|s_k\|^2 > \frac{1-\kappa_\theta}{\sigma_k+\kappa_s}\epsilon.$$

Thus, $\|s_k\|^2 \ge \frac{\bar\epsilon}{\sigma_k}$. ■

Corollary 14 *Let Assumption 1 hold. Consider any realization of Algorithm 1. Let $\bar\epsilon$ satisfy (19), if $k < T_\epsilon - 1$ and iteration k is true (i.e., $I_k = 1$), then if*

$$\sigma_k \ge \sigma_c = \frac{2\kappa_g+\kappa_H+L+L_H}{1-\frac{1}{3}\theta}, \tag{20}$$

the iteration k is successful.

Proof

The result is straightforward by applying Lemma 11 and 13. ■

Lemma 15 *Let Assumption 1 hold. Consider any realization of Algorithm 1. Let $\bar{\epsilon}$ satisfy (19) and $k < T_\epsilon - 1$, then on each true and successful iteration k, we have*

$$\phi(x_k) - \phi(x^{k+1}) \geq \frac{\theta}{6}(1-\kappa_\theta)^{3/2}\frac{\sigma_k}{(\sigma_k+\kappa_s)^{3/2}}\|\nabla\phi(x^{k+1})\|^{3/2} - e_k - e_k^+ - 2\epsilon_f' \tag{21}$$

$$\geq \frac{\theta}{6}(1-\kappa_\theta)^{3/2}\frac{\sigma_{\min}}{(\sigma_k+\kappa_s)^{3/2}}\|\nabla\phi(x^{k+1})\|^{3/2} - e_k - e_k^+ - 2\epsilon_f' \tag{22}$$

$$\geq \frac{\kappa_f}{(\max\{\sigma_k,\sigma_c\})^{3/2}}\|\nabla\phi(x^{k+1})\|^{3/2} - e_k - e_k^+ - 2\epsilon_f', \tag{23}$$

where $\kappa_f := \frac{\theta}{12\sqrt{2}}(1-\kappa_\theta)^{3/2}\sigma_{\min}$ and σ_c is defined in (20).

Proof

Combining Lemma 12, 13, inequality (15) from Lemma 10 and the definition of successful iteration in Algorithm 1 we have, for all true and successful iterations k,

$$\phi(x_k) - \phi(x^{k+1}) \geq \frac{\theta}{6}\sigma_k\|s_k\|^3 - e_k - e_k^+ - 2\epsilon_f' \tag{24}$$

$$\geq \frac{\theta}{6}(1-\kappa_\theta)^{3/2}\frac{\sigma_k}{(\sigma_k+\kappa_s)^{3/2}}\|\nabla\phi(x^{k+1})\|^{3/2} - e_k - e_k^+ - 2\epsilon_f'. \tag{25}$$

Using that $\sigma_k \geq \sigma_{\min}$ and that $\kappa_s \leq \sigma_c$, the result follows. ■

Hence, if $k < T_\epsilon - 1$, and $\bar{\epsilon}$ satisfies (19), any true and successful iteration that has $\sigma_k \leq \sigma_c$ provides $\mathcal{O}(\epsilon^{3/2})$ reduction in $\phi(x)$.

B.1. Proof of Proposition 7

Proof

Part (i) follows easily from the definitions of $\bar{\alpha}$, $h(\alpha)$ and inequality 1.

Part (ii) has exactly the same proof as that of Proposition 1 part (ii) in [18].

Part(iii) follows directly from Lemma 15.

Part(iv) follows directly from Corollary 14.

Part (v) has exactly the same proof as that of Proposition 1 part (v) in [18]. ■