SIAM J. Optim.
Vol. 30, No. 1, pp. 349–376

A STOCHASTIC LINE SEARCH METHOD WITH EXPECTED COMPLEXITY ANALYSIS[*]

COURTNEY PAQUETTE[†] AND KATYA SCHEINBERG[‡]

Abstract. For deterministic optimization, line search methods augment algorithms by providing stability and improved efficiency. Here we adapt a classical backtracking Armijo line search to the stochastic optimization setting. While traditional line search relies on exact computations of the gradient and values of the objective function, our method assumes that these values are available up to some dynamically adjusted accuracy which holds with some sufficiently large, but fixed, probability. We bound the expected number of iterations to reach a desired first-order accuracy in the nonconvex, convex, and strongly convex cases and show that this bound matches the complexity bound of deterministic gradient descent up to constants.

Key words. line search, complexity, stochastic

AMS subject classifications. 90C30, 90C15

DOI. 10.1137/18M1216250

1. Introduction.

We consider the stochastic optimization problem

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1.1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a C^1-smooth function with L-Lipschitz continuous gradients. We assume the function $f(x)$ is not computable, but instead the objective function is approximated by stochastic estimates $\tilde{f}(x;\xi)$. Here ξ is a random variable obeying some probability distribution, P. The most common case addressed in recent literature is the expected loss formulation

$$\min_{x\in\mathbb{R}^n} \left\{ f(x) = \mathbb{E}_{\xi\sim P}[\tilde{f}(x;\xi)] \right\}. \tag{1.2}$$

In this paper we do not specifically require that $\tilde{f}(x;\xi)$ be an unbiased estimate of $f(x)$; instead we require $f(x)$ to be approximated sufficiently accurately, for example, by sampling ξ and averaging these estimates. We will also require that $\nabla f(x)$ be approximated sufficiently accurately. As we discuss later in the paper, this approximation can be obtained by sampling ξ and averaging the estimates $\nabla\tilde{f}(x;\xi)$, if they are available, or by generalized finite difference approximations using $\tilde{f}(x;\xi)$, when estimates $\nabla\tilde{f}(x;\xi)$ are not available.

The most widely used method to solve (1.2) is the stochastic gradient descent (SGD) [18]. Due to its low iteration cost, SGD is often preferred to the standard gradient descent (GD) method for empirical risk minimization. Despite the prevalent use of SGD, it has known challenges and inefficiencies. Firstly, the step direction may

[*] Received by the editors September 24, 2018; accepted for publication (in revised form) October 21, 2019; published electronically February 6, 2020.
https://doi.org/10.1137/18M1216250

Funding: The work of the first author was partially supported by NSF TRIPODS grant 17-40796 and DMS 18-03289. The work of the second author was partially supported by NSF grants CCF 16-18717 and TRIPODS 17-40796 and by DARPA Lagrange award HR-001117S0039.

[†] Department of Industrial and Systems Engineering, Lehigh University, Harold S. Mohler Laboratory, 200 West Packer Avenue, Bethlehem, PA 18015-1582 (cop318@lehigh.edu).

[‡] School of Operations Research and Information Engineering, Cornell University, Rhodes Hall, Ithaca, NY 14850 (katyas@cornell.edu).

not be a descent direction, and secondly, the method is sensitive to the choice of the step size (learning rate) which usually needs to be tuned by hand. Various authors have attempted to address this last issue; see [9, 11, 13, 14]. Motivated by these facts, we turn to the deterministic optimization approach for adaptively selecting step sizes—GD with Armijo backtracking line search.

Related work. In general, GD with backtracking requires exact gradient *and* function evaluations which are impossible for the general problem (1.1). On the other hand, the iteration complexity for GD is superior to SGD making it an attractive alternative. Several works have attempted to transfer ideas from deterministic GD to the stochastic setting by using dynamic gradient sampling (see, e.g., [5, 10, 12]); however, these works address only the convex setting. Moreover to obtain convergence rates matching those of GD in expectation, a constant step size must be determined in advance based on the Lipschitz constant, and hence possibly underestimated, while to decrease the variance of gradient estimates the sample size needs to be increased at a predescribed rate, possibly overestimated. Recently in [4], an adaptive sample size selection strategy was proposed where the sample size is selected based on the reduction of the gradient (and not predescribed). For convergence rates to be derived, however, an assumption had to be made that these sample sizes can be selected based on the size of the true gradient, which is, of course, unknown. In [19, 17], second-order methods that average the sampled gradients and Hessians, a procedure known as subsampling, are proposed; however, the sample sizes are simply assumed to be sufficiently large so that essentially, the methods behave as deterministic inexact methods with some probability. In fact, the convergence rate analysis in [19] is carried out under the assumption that gradient and Hessian estimates are accurate at *every* iteration up until one has reached an ϵ-accurate solution. Thus the probability of reaching this accuracy decays exponentially with the complexity bound. In contrast, our analysis provides a bound on the expected number of iterations to reach an ϵ-accurate solution, and we show the gradient of the iterates converge a.s. to zero.

In [4] and [10], two practical backtracking line search methods were proposed that use their respective heuristic sample sizes to select approximate gradients and function estimates. In both cases, the backtracking is based on the Armijo line search condition applied to function estimates that are computed on the same batch as the gradient estimates. A different type of line search method based on a probabilistic Wolfe condition is proposed in [15]; however, it aims at improving step size selection for SGD and provides no theoretical guarantees. In summary, while several practical line search methods or sampling procedures have been proposed, there has not been a practical stochastic line search method with convergence rate guarantees.

The complexity bounds we derive in this work are different from the complexity guarantees for a majority of other stochastic methods. For instance, in the convex setting, SGD complexity is often derived as the number of iterations until the expected function gap is small. In the nonconvex case the bound is on the expected *sum of the norm squared of the gradients* over all iterates generated by the algorithm. We, instead, bound, in terms of ϵ, the expected number of iterations to achieve an ϵ-accurate solution. Thus we will denote this complexity bound as the *expected complexity* of the line search algorithm, while typical analysis bounds the expected *convergence rates*.

Our analysis relies on a general framework proposed in [3] that uses results from martingale theory to derive a bound on the stopping time of a stochastic process. Using that framework a stochastic trust region method was analyzed in [3] and an expected complexity bound was derived. The analysis from trust region methods does not readily extend to line search methods. In [6] expected complexity of line search method is obtained for the case when the gradient (and Hessian) information may be stochastic but the function values are computed exactly. This work is thus an extension of the line search analysis in [6]. As we will see the analysis of the fully stochastic case is significantly more complex.

Our contribution. In this work we propose the first stochastic backtracking line search method which has rigorous convergence guarantees and requires only *knowable quantities* for implementation. While traditional line search methods rely on exact computations of the gradient and function values, our method assumes that these values are available up to some dynamically adjusted accuracy which holds with some sufficiently large, but fixed, probability. Moreover the step sizes are chosen adaptively. We show that the expected number of iterations to reach an approximate-stationary point matches the worst-case efficiency of typical first-order methods $O(\varepsilon^{-2})$, while for convex and strongly convex objectives, it achieves rates of deterministic GD in function values, $O(\varepsilon^{-1})$ and $O(\log(\varepsilon^{-1}))$, respectively. Our analysis does not require unbiased estimators of either $f(x)$ or $\nabla f(x)$.

Background. There are many types of (deterministic) line search methods (see [16, Chapter 3]), but all share a common philosophy. First, at each iteration, the methods compute a search direction d_k by, e.g., the gradient or (quasi) Newton directions. Next, they determine how far to move in this direction through the univariate function, $\phi(\alpha) = f(x_k + \alpha d_k)$. Typical line searches try out a sequences of potential values for the step size, accepting α once some verifiable criterion becomes satisfied. One popular line search criterion specifies an acceptable step length which gives *sufficient decrease* in the objective function f:

$$\text{(1.3)} \qquad \text{(Armijo condition [1])} \qquad f(x_k + \alpha d_k) \le f(x_k) - \theta\alpha \left\|\nabla f(x_k)\right\|^2,$$

where the constant $\theta \in (0, 1)$ is chosen by the user and $d_k = -\nabla f(x_k)$. Larger step sizes imply larger gains towards optimality and lead to fewer overall iterations. When step sizes get too small no progress is made and the algorithm stagnates. A popular way to systematically search the domain of α while simultaneously preventing small step sizes is backtracking. Backtracking starts with an overestimate of α and decreases it until (1.3) becomes true. Our exposition is on a stochastic version of backtracking using the stochastic gradient estimate as a search direction and stochastic function estimates in (1.3).

1.1. Notation.

The notation we follow is standard. Throughout, we consider an Euclidean space, denoted by $\mathbb{R}^n$, with an inner product and an induced norm $\|\cdot\|$. All stochastic quantities defined hereafter live on a probability space, denoted by $(\mathbf{Pr}, \Omega, \mathcal{F})$, with the probability measure $\mathbf{Pr}$ and σ-algebra $\mathcal{F}$ containing subsets of Ω. A random variable (vector) is a measurable map from Ω to $\mathbb{R}$ ($\mathbb{R}^n$), respectively. As is standard in probability theory, we never explicitly define the space Ω but implicitly specify it through the random variables. In the remainder of the paper, all random quantities will be denoted by capitalized letters and their respective realizations by corresponding lower case letters. For instance, a *realization* of the random variable $X : \Omega \to \mathbb{R}$ is given by $X(\omega) =: x$ for some fixed $\omega \in \Omega$. Given a subset A of Ω, we call the set A an *event* and denote the indicator of the event A by

$$1_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{otherwise.} \end{cases}$$

For any random variable $X : \Omega \to \mathbb{R}$ and any constant $c \in \mathbb{R}$, we use the following notation for an event generated from a random variable:

$$\{X \leq c\} := \{\omega \,:\, X(\omega) \leq c\}.$$

2. Stochastic backtracking line search method.

We present here our main algorithm for GD with backtracking line search. We impose a standard assumption on the objective function.

Assumption 2.1. The gradient of f is L-Lipschitz continuous for all $x \in \mathbb{R}^n$ and there exists a lower bound $f_{\min}$, i.e.,

$$f_{\min} \leq f(x) \qquad \text{for all } x \in \mathbb{R}^n.$$

For the convergence analysis to hold, we will impose additional assumptions on how certain steps in Algorithm 1 are performed. For the moment, we present the algorithm as a general framework, without particularizing the details of each step, and we introduce the assumptions and examples of how each step can be satisfied later in this section.

2.1. Outline of the method.

At each iteration, our scheme computes a random direction g_k via, e.g., a minibatch stochastic gradient estimate or sampling the function $f(x)$ itself and using finite differences. Then, we compute stochastic function estimates at the current iterate and at the prospective new iterate, respectively, f_k^0 and f_k^s. Given these stochastic estimates, we check the Armijo condition [1]:

$$\text{(Stochastic) Armijo} \qquad f_k^s \leq f_k^0 - \theta\alpha_k \left\|g_k\right\|^2. \tag{2.1}$$

If (2.1) holds, the next iterate becomes $x_{k+1} = x_k - \alpha_k g_k$ and the step size α_k increases; otherwise $x_{k+1} = x_k$ and α_k decreases, as is typical in (deterministic) backtracking line searches.

Algorithm 1 describes our method.[1] Unlike classical backtracking line search, the gradient estimate is recomputed every time, even if the iterate is not changed. This is necessary since each particular g_k may not be a descent direction. Also there is an additional control, δ_k, which serves as a guess of the increase in the true function at the point x_k and controls the accuracy of the function estimates. We discuss this further next.

Challenges with randomized line search. Due to the stochasticity in the gradient and/or function values, two major challenges result:

- a series of erroneous unsuccessful steps cause α_k to become arbitrarily small;
- steps may satisfy (2.1) but, in fact, $f(x_k - \alpha_k g_k) > f(x_k)$ leading to the objective value at the next iteration larger than the current iterate.

Convergence proofs for line searches in the deterministic setting rely on the fact that neither of the above problems arises. To handle the first difficulty, our approach controls the probability that the random gradients and function values are representative of their true counterparts. When this probability is large and the step size α_k is sufficiently small, the method tends to make successful steps. Intuitively, the step sizes α_k behave like a random walk with an upward drift; thus they tend to stay away from 0.

Yet, even when the probability of good gradients/function estimates is near 1, it is not guaranteed that $f(x_{k+1}) < f(x_k)$ holds, even in expectation, at each iteration

[1] We state the algorithm using the lower case notation to represent a realization of the algorithm.

```
Algorithm 1: Line search method.
  Initialization: Choose constants $\gamma > 1$, $\theta \in (0,1)$, and $\alpha_{\max} > 0$. Pick
    initial point $x_0$, $\alpha_0 = \gamma^{j_0}\alpha_{\max}$ for some $j_0 \leq 0$ and $\delta_0 > 0$.
  Repeat for $k = 0, 1, \ldots$
      1. Compute a gradient estimate Based on $x_k$ compute a gradient
         estimate $g_k$ satisfying Assumption 2.4. Set the step $s_k = -\alpha_k g_k$.
      2. Compute function estimates Based on $\delta_k$, $g_k$, and $x_k$ obtain
         estimates of $f_k^0$ and $f_k^s$ of $f(x_k)$ and $f(x_k+s_k)$, respectively, satisfying
         Assumption 2.4.
      3. Check sufficient decrease
         Check if
         (2.2)    $f_k^s \leq f_k^0 - \alpha_k \theta \|g_k\|^2$.
      4. Successful step
         If (2.2) set $x_{k+1} = x_k + s_k$ and $\alpha_{k+1} = \min\{\alpha_{\max}, \gamma\alpha_k\}$.
           • Reliable step: If $\alpha_k \|g_k\|^2 \geq \delta_k^2$, then increase $\delta_{k+1}^2 = \gamma\delta_k^2$.
           • Unreliable step: If $\alpha_k \|g_k\|^2 < \delta_k^2$, then decrease $\delta_{k+1}^2 = \gamma^{-1}\delta_k^2$.
      5. Unsuccessful step
         Otherwise, set $x_{k+1} = x_k$, $\alpha_{k+1} = \gamma^{-1}\alpha_k$, and $\delta_{k+1}^2 = \gamma^{-1}\delta_k^2$.
```

due to the second issue. When such an increase in the objective function occurs, it can be arbitrarily large; in fact, the objective function can increase by at most $\alpha_k^2 \|g_k\|^2$. To control this increase in the objective, we introduce the quantity δ_k^2, which, on the one hand, is meant to predict the largest possible increase in the objective function and, on the other hand, changes conservatively from one iteration to the next. When the predicted increase is small relative to the decrease in the function estimates given through the sufficient decrease condition (2.2), namely, $\delta_k^2 \leq \alpha_k \|g_k\|^2$, then the step, $-\alpha_k g_k$, will likely decrease the objective function. We call such a step *reliable* and increase the parameter δ_k^2 for the next iteration. Otherwise, our predicted increase is larger than the decrease in the function estimates so the step, $-\alpha_k g_k$, may increase the objective function. In this case, we call the step *unreliable* and decrease the parameter δ_k^2.

Unlike the typical stochastic convergence rate analysis, which bounds expected improvement in either $\mathbf{E}(\|\nabla f(x)\|)$ or $\mathbf{E}(f(x) - f_{\min})$ after a given number of iteration, our analysis bounds the expected complexity, which is the total expected number of steps that the algorithm takes before either $\|\nabla f(x)\| \leq \varepsilon$ or $f(x) - f_{\min} \leq \varepsilon$ is reached. Like in this paper, in the works [19, 17], it is assumed that at each iteration the stochastic gradient and/or the stochastic Hessian approximate their true values with sufficiently high accuracy and this holds with some probability p. However in [19, 17] unlike our work, the accuracy of the stochastic gradient and/or the stochastic Hessian is not chosen dynamically but instead depends on the final desired optimization accuracy ϵ. They choose, at each iteration, the approximate stochastic gradient and/or Hessian to be essentially the true gradient and/or Hessian with a small error and their choices of the stochastic gradient/Hessian hold with probability p that depends on the optimization accuracy ϵ. As such, the analysis of their methods reduces to the deterministic setting and yields expected complexity bounds that hold with high probability. First, they provide no complexity analysis when the stochastic gradient/Hessian fails to be essentially the true gradient/Hessian. Secondly, the probability p relies heavily on ϵ. In contrast, we derive our expected complexity bounds using stochastic gradient estimates that, at each iteration, dynamically change and the stochastic estimates hold for a *fixed probability*, independent of the optimization accuracy. Our results rely on a stochastic process framework introduced and analyzed in [3] for a stochastic trust region method.

2.2. Random gradient and function estimates.

Overview. At each iteration, we compute a stochastic gradient and stochastic function values. With probability p_g, the stochastic gradient g_k is "close" to the true gradient, which means that the error between g_k and the true gradient at the current iterate is bounded using the current step length.This procedure *naturally adapts* the required accuracy of gradient estimates as the algorithm progresses. As the steps get shorter (i.e., either the gradient gets smaller or the step size parameter does), we require the accuracy to increase, but the probability p_g of encountering a good gradient g_k at any iteration is the same.

A similar procedure applies to function estimates, f_k^0 and f_k^s. The error between the function estimates and the true function values at the points x_k and $x_k + s_k$ ($s_k = -\alpha_k g_k$) are tied to the size of the step, $\alpha_k \|g_k\|$. At each iteration, there is a probability p_f of obtaining good function estimates. By choosing the probabilities of good gradient and estimates, we show Algorithm 1 converges. To formalize this procedure, we introduce the following.

Notation and definitions. Algorithm 1 generates a random process given by the sequence $\{G_k, X_k, \mathcal{A}_k, \Delta_k, S_k, F_k^0, F_k^s\}$. In what follows we will denote all random quantities by capital letters and their realization by small letters. For instance, the random gradient estimate is denoted by G_k and its realization by $g_k = G_k(\omega)$. Similarly, let the quantities $x_k = X_k(\omega)$ (iterates), $\alpha_k = \mathcal{A}_k(\omega)$ (step size), control size $\Delta_k(\omega) = \delta_k$, and $s_k = S_k(\omega)$ (step) denote the respective realizations of the random quantities. Similarly, we let $\{F_k^0, F_k^s\}$ denote estimates of $f(X_k)$ and $f(X_k + S_k)$, with their realizations denoted by $f_k^0 = F_k^0(\omega)$ and $f_k^s = F_k^s(\omega)$. Our goal is to show that under some assumptions on G_k and $\{F_k^0, F_k^s\}$ the resulting stochastic process converges with probability one and at an appropriate rate. In particular, we assume that the estimates G_k and F_k^0 and F_k^s are sufficiently accurate with sufficiently high probability, conditioned on the past.

To formalize the conditioning on the past, let $\mathcal{F}_{k-1}^{G\cdot F}$ denote the σ-algebra generated by the random variables $G_0, G_1, \ldots, G_{k-1}$ and $F_0^0, F_0^s, F_1^0, F_1^s, \ldots, F_{k-1}^0, F_{k-1}^s$, and let $\mathcal{F}_{k-1/2}^{G\cdot F}$ denote the σ-algebra generated by the random variables $G_0, G_1, \ldots, G_k$ and $F_0^0, F_0^s, F_1^0, F_1^s, \ldots, F_{k-1}^0, F_{k-1}^s$. For completeness, we set $\mathcal{F}_{-1}^{G\cdot F} = \sigma(x_0)$. As a result, we have that $\mathcal{F}_k^{G\cdot F}$ for $k \geq -1$ is a filtration. By construction of the random variables X_k and $\mathcal{A}_k$ in Algorithm 1, we see $\mathbf{E}[X_k | \mathcal{F}_{k-1}^{G\cdot F}] = X_k$ and $\mathbf{E}[\mathcal{A}_k | \mathcal{F}_{k-1}^{G\cdot F}] = \mathcal{A}_k$ for all $k \geq 0$.

We measure accuracy of the gradient estimates G_k and function estimates F_k^0 and F_k^s using the following definitions.

DEFINITION 2.2. *We say that a sequence of random directions $\{G_k\}$ is (p_g)-probabilistically κ_g-sufficiently accurate for Algorithm* 1 *for the corresponding sequence $\{\mathcal{A}_k, X_k\}$ if there exists a constant $\kappa_g > 0$, such that the event*

$$I_k = \{\|G_k - \nabla f(X_k)\| \leq \kappa_g \mathcal{A}_k \|G_k\|\}$$

satisfies the conditions

$$\mathbf{Pr}(I_k | \mathcal{F}_{k-1}^{G\cdot F}) = \mathbf{E}[1_{I_k} | \mathcal{F}_{k-1}^{G\cdot F}] \geq p_g.$$

In addition to sufficiently accurate gradients, we require estimates on the function values $f(x_k)$ and $f(x_k + s_k)$ to also be sufficiently accurate.

Definition 2.3. *A sequence of random estimates* $\{F_k^0, F_k^s\}$ *is said to be* p_f*-probabilistically* ε_f*-accurate with respect to the corresponding sequence* $\{X_k, \mathcal{A}_k, S_k\}$ *if the event*

$$J_k = \{|F_k^0 - f(x_k)| \le \varepsilon_f \mathcal{A}_k^2 \|G_k\|^2 \quad \textit{and} \quad |F_k^s - f(x_k + s_k)| \le \varepsilon_f \mathcal{A}_k^2 \|G_k\|^2\}$$

satisfies the condition

$$\mathbf{Pr}(J_k | \mathcal{F}_{k-1/2}^{G\cdot F}) = \mathbf{E}[1_{J_k} | \mathcal{F}_{k-1/2}^{G\cdot F}] \ge p_f.$$

We note here that the filtration $\mathcal{F}_{k-1/2}^{G\cdot F}$ determines the random quantities $\mathcal{A}_k$ and G_k; hence the accuracy of the estimates is measured with respect to fixed quantities. Next, we state the key assumption on the nature of the stochastic information in Algorithm 1.

Assumption 2.4. The following hold for the quantities in the algorithm:

(i) The random gradients G_k generated by Algorithm 1 are p_g-probabilistically κ_g-sufficiently accurate for some sufficiently large $p_g \in (0, 1]$.

(ii) The estimates $\{F_k^0, F_k^s\}$ generated by Algorithm 1 are p_f-probabilistically ε_f-accurate estimates for some $\varepsilon_f \le \frac{\theta}{4\alpha_{\max}}$ and sufficiently large $p_f \in (0, 1]$.

(iii) The sequence of estimates $\{F_k^0, F_k^s\}$ generated by Algorithm 1 satisfies a κ_f-variance condition for all $k \ge 0$.[2]

(2.3)
$$\begin{aligned} &\mathbf{E}[|F_k^s - f(X_k + S_k)|^2 | \mathcal{F}_{k-1/2}^{G\cdot F}] \le \max\{\kappa_f^2 \mathcal{A}_k^2 \|\nabla f(X_k)\|^4, \theta^2 \Delta_k^4\} \\ \text{and} \quad &\mathbf{E}[|F_k^0 - f(X_k)|^2 | \mathcal{F}_{k-1/2}^{G\cdot F}] \le \max\{\kappa_f^2 \mathcal{A}_k^2 \|\nabla f(X_k)\|^4, \theta^2 \Delta_k^4\}. \end{aligned}$$

A simple calculation shows that under Assumption 2.4 the following hold:

$$\mathbf{E}[1_{I_k \cap J_k} | \mathcal{F}_{k-1}^{G\cdot F}] \ge p_g p_f, \quad \mathbf{E}[1_{I_k^c \cap J_k} | \mathcal{F}_{k-1}^{G\cdot F}] \le 1 - p_g, \quad \text{and} \quad \mathbf{E}[1_{J_k^c} | \mathcal{F}_{k-1}^{G\cdot F}] \le 1 - p_f.$$

Remark 1. We are interested in deriving convergence results for the case when κ_g may be large. For the rest of the exposition, without loss of generality $\kappa_g \ge 2$. It is clear that if κ_g happens to be smaller, somewhat better bounds than the ones we derive here will result since the gradients give tighter approximations of the true gradient. Equation (2.3) includes the maximum of two terms—one of the terms $\|\nabla f(X_k)\|$ is unknown. When one possesses external knowledge of $\|\nabla f(X_k)\|$, one could use this value. This is particularly useful when $\|\nabla f(X_k)\|$ is big since it allows large variance in the function estimates; for example, assumption that $\|\nabla f(X_k)\| \ge \varepsilon$ implies that this variance does not have to be driven to zero before the algorithm reaches a desired accuracy. Since a useful lower bound on $\|\nabla f(X_k)\|$ may be unknown, we include the parameter Δ_k as a way to adaptively control the variance. As such κ_f should be small; in fact, it can be set equal to 0. The analysis can be performed for any other values of the above constants—the choices here are for simplicity and convenience.

Assumption 2.4 on the accuracy of the gradient and function estimates is key in our complexity analysis. We derive specific bounds on p_g and p_f under which our results hold. We note here that if $p_f = 1$, then Assumption 2.4(iii) is not needed,

[2] We implicitly assume $|F_k^s - f(X_k + S_k)|^2$ and $|F_k^0 - f(X_k)|^2$ are integrable for all k; thus it is straightforward to deduce $|F_k^s - f(X_k + S_k)|$ and $|F_k^0 - f(X_k)|$ are integrable for all k.

and condition $p_g > 1/2$ is sufficient for the convergence results. This case can be considered as an extension of results in [6]. Before concluding this section, we state a result showing the relationship between the variance assumption on the function values and the probability of inaccurate estimates.

LEMMA 2.5. *Let Assumption 2.4 hold. Suppose the random process generated by Algorithm 1 is $\{G_k, X_k, \mathcal{A}_k, \Delta_k, S_k, F_k^0, F_k^s\}$ and $\{F_k^0, F_k^s\}$ are p_f-probabilistically accurate estimates. Then for every $k \geq 0$ we have*

$$\mathbf{E}[1_{J_k^c}|F_k^s - f(X_k+S_k)|\ |\mathcal{F}_{k-1/2}^{G\cdot F}] \leq (1-p_f)^{1/2} \max\{\kappa_f \mathcal{A}_k \|\nabla f(X_k)\|^2, \theta\Delta_k^2\}$$

$$\textit{and}\quad \mathbf{E}[1_{J_k^c}|F_k^0 - f(X_k)|\ |\mathcal{F}_{k-1/2}^{G\cdot F}] \leq (1-p_f)^{1/2} \max\{\kappa_f \mathcal{A}_k \|\nabla f(X_k)\|^2, \theta\Delta_k^2\}.$$

Proof. We show the result for $F_k^0 - f(X_k)$ and note the proof for $F_k^s - f(X_k+S_k)$ is the same. Using Hölder's inequality for conditional expectations, we deduce

$$\mathbf{E}\left[\frac{1_{J_k^c}|F_k^0 - f(X_k)|}{\max\{\kappa_f \mathcal{A}_k \|\nabla f(X_k)\|^2, \theta\Delta_k^2\}} |\mathcal{F}_{k-1/2}^{G\cdot F}\right]$$
$$\leq \left(\mathbf{E}[1_{J_k^c}|\mathcal{F}_{k-1/2}^{G\cdot F}]\right)^{1/2} \left(\mathbf{E}\left[\frac{|F_k^0 - f(X_k)|^2}{\max\{\kappa_f^2 \mathcal{A}_k^2 \|\nabla f(X_k)\|^4, \theta^2\Delta_k^4\}} |\mathcal{F}_{k-1/2}^{G\cdot F}\right]\right)^{1/2}.$$

The result follows after noting by (2.3)

$$\left(\mathbf{E}\left[\frac{|F_k^0 - f(X_k)|^2}{\max\{\kappa_f^2 \mathcal{A}_k^2 \|\nabla f(X_k)\|^4, \theta^2\Delta_k^4\}} |\mathcal{F}_{k-1/2}^{G\cdot F}\right]\right)^{1/2} \leq 1. \qquad \square$$

2.3. Computing G_k, F_k^0, and F_k^s to satisfy Assumption 2.4. In this section, we discuss one approach for computing stochastic gradients and stochastic function estimates that satisfy Assumption 2.4 when minimizing an expected loss function

$$\min_x \left\{f(x) = \mathbf{E}_{\xi\sim P}[\tilde{f}(x;\xi)]\right\}.$$

Recall that $f(x)$ and $\nabla f(x)$ can be approximately computed using random realizations $\tilde{f}(x,\xi)$ and $\nabla \tilde{f}(x;\xi)$. One approach for computing g_k (resp., f_k^0 and f_k^s) such that it satisfies Assumption 2.4 is as follows—sample ξ from the probability distribution P a total of $|\mathcal{S}_k^g|$ (resp., $|\mathcal{S}_k^f|$) times and then average the estimates $\nabla\tilde{f}(x_k,\xi_i)$ (resp., $\tilde{f}(x_k,\xi)$ and $\tilde{f}(x_k - \alpha_k g_k, \xi)$). By choosing specific values for the number of samples, $|\mathcal{S}_k^g|$ and $|\mathcal{S}_k^f|$, the averaged random realizations satisfy Assumption 2.4. For many machine learning problems, one thinks of ξ as a data point. We describe this procedure below.

We impose, only for this section, an assumption on the boundedness of the variances of the random function and gradient realizations

$$\mathbf{E}(\|\nabla\tilde{f}(x,\xi) - \nabla f(x)\|^2) \leq V_g \text{ and } \mathbf{E}(|\tilde{f}(x,\xi) - f(x)|^2) \leq V_f.$$

At each iteration k, sample ξ from the probability distribution P a total of $\mathcal{S}_k^g$ number of times. Then compute $\nabla f(x_k,\xi_i)$ for every $i \in \mathcal{S}_k^g$, and set the stochastic gradient estimate as the average $g_k := \frac{1}{|\mathcal{S}_k^g|}\sum_{i\in\mathcal{S}_k^g} \nabla\tilde{f}(x_k,\xi_i)$. Using results (see, e.g., in [19, 20]), we choose the number of samples from the probability distribution such that

$$|\mathcal{S}_k^g| \geq \tilde{O}\left(\frac{V_g}{\kappa_g^2\alpha_k^2\|g_k\|^2}\right) \tag{2.4}$$

(where $\tilde{O}$ hides the log factor of $1/(1-p_g)$). This choice ensures that Assumption 2.4(i) is satisfied for the stochastic gradient. While g_k is not known when $|\mathcal{S}_k^g|$ is chosen, one can design a simple loop by guessing the value of $\|g_k\|$ and increasing the number of samples until (2.4) is satisfied; this procedure is discussed in [6]. Next, with the stochastic gradient g_k chosen, we use a similar procedure to generate f_k^0 and f_k^0 by sampling ξ a total of $\mathcal{S}_k^f$ number of times. To satisfy Assumption 2.4(ii), it suffices to compute $f_k^0 = \frac{1}{|\mathcal{S}_k^f|}\sum_{i\in\mathcal{S}_k^f}\tilde{f}(x_k,\xi_i)$ with

$$|\mathcal{S}_k^f| \geq \tilde{O}\left(\frac{V_f}{\kappa_f^2\alpha_k^2\|g_k\|^4}\right)$$

(where $\tilde{O}$ hides the log factor of $1/(1-p_f)$). We obtain f_k^s analogously. Finally if the number of samples satisfies $|\mathcal{S}_k^f| \geq \frac{V_f}{\theta^2\delta_k^4}$, then Assumption 2.4(iii) holds for the function estimates. Chebyshev inequality, a standard probability inequality, directly proves this result. To summarize, Assumption 2.4 holds provided we choose the number of samples larger than

$$|\mathcal{S}_k^g| \geq \tilde{O}\left(\frac{V_g}{\kappa_g^2\alpha_k^2\|g_k\|^2}\right) \quad \text{and} \quad |\mathcal{S}_k^f| \geq \max\left\{\tilde{O}\left(\frac{V_f}{\kappa_f^2\alpha_k^2\|g_k\|^4}\right), \frac{V_f}{\theta^2\delta_k^4}\right\}.$$

We observe the following.

- Unlike [5, 10], the number of samples for gradient and function estimation does not increase at any predefined rate but is closely related to the progress of the algorithm. In particular if $\alpha_k\|g_k\|$ and δ_k increase, then the sample sizes can decrease.
- Also, unlike [19] where the number of samples is simply chosen large enough a priori for all k so that the right-hand side in Assumption 2.4(i) is bounded by a predefined accuracy $O(\varepsilon)$, our algorithm can be applied without an a priori choice of ε but with a choice of a total computational budget, for instance.
- Finally, unlike [4] where theoretical results require that $|\mathcal{S}_k^g|$ and $|\mathcal{S}_k^f|$ depend on $\|\nabla f(x_k)\|$, which is unknown, our bounds on the sample sizes can be computed using knowable values, such as bounds on the variances and quantities determined by prior iterates in the algorithm.

We also point out κ_g can be arbitrarily big and that p_g, as we will show later, depends only on the backtracking factor γ and is not close to 1; hence the number of samples to satisfy Assumption 2.4(i) is moderate. On the other hand, p_f will have to depend on κ_g; hence a looser control of the gradient estimates results in tighter control, i.e., larger sample sets, for function estimates.

Our last comment is that G_k does not have to be an unbiased estimate of $\nabla f(X_k)$ and does not need to be computed via gradient samples. Instead it can be computed via stochastic finite differences, as is discussed, for example, in [7].

3. Renewal-reward process.

In this section, we define a general random process introduced in [3] and its stopping time T that serve as a general framework for analyzing the behaviors of a stochastic trust region method in [3] and our stochastic line search method. We will then show that our stochastic line search method satisfies the properties of this random process, with the stopping time defined by the time of reaching desired accuracy. We will show how this framework applies to the nonconvex, convex, and strongly convex cases, which will allow us to derive the bound on the expected complexity for our method in each of these cases.

Here we state the relevant definitions, assumptions, and theorems and refer the reader to the proofs in [3].

Definition 3.1. *Given a discrete time stochastic process $\{X_k\}$, a random variable T is a stopping time for $\{X_k\}$ if the event $\{T = k\} \in \sigma(X_0, \ldots, X_k)$.*

Let $\{\Phi_k, \mathcal{A}_k\}$ be a random process such that $\Phi_k \in [0, \infty)$ and $\mathcal{A}_k \in [0, \infty)$ for all $k \geq 0$. We, also, introduce a biased random walk process, $\{W_k\}_{k=1}^{\infty}$, defined on the same probability space as $\{\Phi_k, \mathcal{A}_k\}$. We denote $\mathcal{F}_k$ the σ-algebra generated by $\{\Phi_0, \mathcal{A}_0, W_0, \ldots, \Phi_k, \mathcal{A}_k, W_k\}$, where $W_0 = 1$. In addition, W_k obeys the following dynamics for some constant $\frac{1}{2} < p < 1$:

$$\mathbf{Pr}(W_{k+1} = 1 | \mathcal{F}_k) = p \quad \text{and} \quad \mathbf{Pr}(W_{k+1} = -1 | \mathcal{F}_k) = (1 - p). \tag{3.1}$$

We define T_ε to be a family of stopping times parameterized by ε. In [3] a bound on $\mathbf{E}(T_\varepsilon)$ is derived under the following assumption on the process $\{\Phi_k, \mathcal{A}_k\}$.

Assumption 3.2. The following hold for the process $\{\Phi_k, \mathcal{A}_k, W_k\}$.

(i) The random variable, $\mathcal{A}_0$, is a constant. Fix a constant $\alpha_{\max} > 0$. There exists a constant $\lambda \in (0, \infty)$ such that $\alpha_{\max} = \mathcal{A}_0 e^{\lambda j_{\max}}$ for some $j_{\max} \in \mathbb{Z}$ and the random variables satisfy $\mathcal{A}_k \leq \alpha_{\max}$ for all $k \geq 0$.

(ii) There exists a constant $\bar{\mathcal{A}} = \mathcal{A}_0 e^{\lambda \bar{j}}$ for some $\bar{j} \in \mathbb{Z}$ with $\bar{j} \leq 0$ such that the following holds for all $k \geq 0$:

$$1_{\{T_\epsilon > k\}} \mathcal{A}_{k+1} \geq 1_{\{T_\epsilon > k\}} \min\left\{ \mathcal{A}_k e^{\lambda W_{k+1}}, \bar{\mathcal{A}} \right\},$$

where W_{k+1} satisfies (3.1) with $p > \frac{1}{2}$.

(iii) There exist a nondecreasing function $h : [0, \infty) \to (0, \infty)$ and a constant $\Theta > 0$ such that

$$1_{\{T_\epsilon > k\}} \mathbf{E}\left[\Phi_{k+1} | \mathcal{F}_k\right] \leq 1_{\{T_\epsilon > k\}} (\Phi_k - \Theta h(\mathcal{A}_k)).$$

Assumption 3.2(iii) states that conditioned on the event $T_\varepsilon > k$ and the past, the random variable Φ_k decreases by $\Theta h(\mathcal{A}_k)$ at each iteration. Whereas Assumption 3.2(ii) says that once $\mathcal{A}_k$ falls below the fixed constant $\bar{\mathcal{A}}$, the sequence has a tendency to increase. Assumption 3.2(i) and (ii) together also ensure that $\bar{\mathcal{A}}$ belongs to the sequence of values taken by the sequence $\mathcal{A}_k$. As we will see this is a simple technical assumption that can be satisfied without loss of generality.

Remark 2. Computational complexity (in deterministic methods) measures the number of iterations until an event such as $\|\nabla f(x)\|$ is small or $f(x_k) - f^*$ is small or, equivalently, the rate at which the gradient/function values decreases as a function of the iteration counter k. For randomized or stochastic methods, previous works tended to focus on the second definition, i.e., showing the expected size of the gradient or function values decreases like $1/k$. Instead, here we bound the expected number of iterations until the size of the gradient or function values is small, which is the same as bounding the stopping times $T_\varepsilon = \inf\{k \geq 0 : \|\nabla f(X_k)\| < \varepsilon\}$ and $T_\varepsilon = \inf\{k \geq 0 : f(X_k) - f^* \leq \varepsilon\}$ for a fixed $\varepsilon > 0$.

Remark 3. In the context of deterministic line search when the step size α_k falls below the constant $1/L$, where L is the Lipschitz constant of $\nabla f(x)$, the iterate $x_k + s_k$ always satisfies the sufficient decrease condition, namely, $f(x_k + s_k) \leq f(x_k) - \theta \alpha_k \|\nabla f(x_k)\|^2$. Thus α_k never falls much below $1/L$. To match the dynamics behind deterministic line search, we expect $\Phi_{k+1} - \Phi_k \approx f(X_{k+1}) - f(X_k)$ with $\Theta h(\mathcal{A}_k) \approx \mathcal{A}_k \|\nabla f(X_k)\|^2$ and the constant $\bar{\mathcal{A}} \approx 1/L$. However, in the stochastic setting there is a positive probability of $\mathcal{A}_k$ being arbitrarily small. Theorem 3.3, below, is derived by observing that on average $\mathcal{A}_k \geq \bar{\mathcal{A}}$ occurs frequently due to the upward drift in the random walk process. Consequently, $\mathbf{E}[\Phi_{k+1} - \Phi_k]$ can be bounded by a negative fixed value (dependent on ε) frequently; thus we can derive a bound on $\mathbf{E}[T_\varepsilon]$.

The following theorem (Theorem 2.2 in [3]) bounds $\mathbf{E}[T_\varepsilon]$ in terms of $h(\bar{\mathcal{A}})$ and Φ_0.

THEOREM 3.3. *Under Assumption* 3.2*,*

$$\mathbf{E}[T_\varepsilon] \leq \frac{p}{2p-1} \cdot \frac{\Phi_0}{\Theta h(\bar{\mathcal{A}})} + 1.$$

4. Convergence of stochastic line search.

4.1. Useful results. Before delving into the main analysis, we state some auxiliary lemmas similar to those derived in [6, 2, 7].

LEMMA 4.1 (accurate gradients ⇒ lower bound on $\|g_k\|$). *Suppose g_k is κ_g-sufficiently accurate. Then*

$$\frac{\|\nabla f(x_k)\|}{(\kappa_g \alpha_{\max} + 1)} \leq \|g_k\|.$$

Proof. The fact that g_k is κ_g-sufficiently accurate together with the triangle inequality implies

$$\|\nabla f(x_k)\| \leq (\kappa_g \alpha_k + 1)\|g_k\| \leq (\kappa_g \alpha_{\max} + 1)\|g_k\|. \qquad \square$$

LEMMA 4.2 (accurate gradient and function estimates and small step size ⇒ successful step). *Suppose g_k is κ_g-sufficiently accurate and $\{f_k^0, f_k^s\}$ are ε_f-accurate estimates. If*

$$\alpha_k \leq \frac{1-\theta}{\kappa_g + \frac{L}{2} + 2\varepsilon_f},$$

then the kth step is successful. In particular, this means $f_k^s \leq f_k^0 - \theta\alpha_k \|g_k\|^2$.

Proof. The L-smoothness of f and the κ_g-sufficiently accurate gradient immediately yield

$$\begin{aligned} f(x_k + s_k) &\leq f(x_k) - \alpha_k(\nabla f(x_k) - g_k)^T g_k - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2 \\ &\leq f(x_k) + \kappa_g \alpha_k^2 \|g_k\|^2 - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2. \end{aligned}$$

Since the estimates are ε_f-accurate, we obtain

$$\begin{aligned} f_k^s - \varepsilon_f \alpha_k^2 \|g_k\|^2 &\leq f(x_k + s_k) - f_k^s + f_k^s \\ &\leq f(x_k) - f_k^0 + f_k^0 + \kappa_g \alpha_k^2 \|g_k\|^2 - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2 \\ &\leq f_k^0 + \varepsilon_f \alpha_k^2 \|g_k\|^2 + \kappa_g \alpha_k^2 \|g_k\|^2 - \alpha_k \|g_k\|^2 + \frac{L\alpha_k^2}{2}\|g_k\|^2. \end{aligned}$$

The result follows by noting $f_k^s \leq f_k^0 - \alpha_k \|g_k\|^2 \left(1 - \alpha_k\left(\kappa_g + \frac{L}{2} + 2\varepsilon_f\right)\right)$. $\square$

Lemma 4.3 (accurate function estimates and successful step ⇒ decrease in function). *Suppose* $\varepsilon_f < \frac{\theta}{4\alpha_{\max}}$ *and* $\{f_k^s, f_k^0\}$ *are* ε_f*-accurate estimates. If the step is successful, then the improvement in function value is*

$$f(x_{k+1}) \le f(x_k) - \frac{\theta\alpha_k}{2}\left\|g_k\right\|^2. \tag{4.1}$$

If, in addition, the step is reliable, then the improvement in function value is

$$f(x_{k+1}) \le f(x_k) - \frac{\theta\alpha_k}{4}\left\|g_k\right\|^2 - \frac{\theta}{4}\delta_k^2. \tag{4.2}$$

Proof. The step is successful and the estimates are ε_f-accurate, so we conclude

$$\begin{aligned} f(x_k + s_k) &\le f(x_k+s_k) - f_k^s + f_k^0 - f(x_k) + f(x_k) - \alpha_k\theta\left\|g_k\right\|^2 \\ &\le f(x_k) + 2\varepsilon_f\alpha_k^2\left\|g_k\right\|^2 - \alpha_k\theta\left\|g_k\right\|^2 \\ &\le f(x_k) - \alpha_k\left\|g_k\right\|^2(\theta - 2\varepsilon_f\alpha_{\max}), \end{aligned}$$

where the last inequality follows because $\alpha_k \le \alpha_{\max}$. The condition $\varepsilon_f < \frac{\theta}{4\alpha_{\max}}$ immediately implies (4.1). By noticing $\frac{\theta\alpha_k}{2}\left\|g_k\right\|^2 \ge \frac{\theta\alpha_k}{4}\left\|g_k\right\|^2 + \frac{\theta\delta_k^2}{4}$ holds for reliable steps, we deduce (4.2). □

Lemma 4.4 (bound on gradient change on successful iterations). *Suppose the kth step is successful. Then*

$$\left\|\nabla f(x_{k+1})\right\|^2 \le 2\left(L^2\alpha_k^2\left\|g_k\right\|^2 + \left\|\nabla f(x_k)\right\|^2\right).$$

In particular, the following inequality holds:

$$\frac{1}{L^2}\left(\alpha_{k+1}\left\|\nabla f(x_{k+1})\right\|^2 - \alpha_k\left\|\nabla f(x_k)\right\|^2\right) \le 2\gamma\alpha_k\left(\alpha_{\max}^2\left\|g_k\right\|^2 + \frac{1}{L^2}\left\|\nabla f(x_k)\right\|^2\right).$$

Proof. An immediate consequence of the L-smoothness of f is $\|\nabla f(x_{k+1})\| \le L\alpha_k\|g_k\| + \|\nabla f(x_k)\|$. The result follows from squaring both sides and applying the bound, $(a+b)^2 \le 2(a^2+b^2)$. To obtain the second inequality, we note that in the case the iteration is successful, $\alpha_{k+1} = \gamma\alpha_k$. □

Lemma 4.5 (accurate gradients and function estimates and successful step ⇒ decrease in function). *Suppose* g_k *is* κ_g*-sufficiently accurate and* $\{f_k^0, f_k^s\}$ *are* ε_f*-accurate estimates where* $\varepsilon_f \le \frac{\theta}{4\alpha_{\max}}$*. If the step is successful, then*

$$f(x_{k+1}) - f(x_k) \le -\frac{\theta\alpha_k}{4}\left\|g_k\right\|^2 - \frac{\theta\alpha_k}{4(\kappa_g\alpha_{\max}+1)^2}\left\|\nabla f(x_k)\right\|^2. \tag{4.3}$$

In addition, if the step is reliable, then

$$f(x_{k+1}) - f(x_k) \le -\frac{\theta\alpha_k}{8}\left\|g_k\right\|^2 - \frac{\theta}{8}\delta_k^2 - \frac{\theta\alpha_k}{4(\kappa_g\alpha_{\max}+1)^2}\left\|\nabla f(x_k)\right\|^2. \tag{4.4}$$

Proof. Lemma 4.1 implies

$$-\frac{\theta}{2}\alpha_k\left\|g_k\right\|^2 \le -\frac{\theta}{4}\alpha_k\left\|g_k\right\|^2 - \frac{\theta}{4(\kappa_g\alpha_{\max}+1)^2}\alpha_k\left\|\nabla f(x_k)\right\|^2. \tag{4.5}$$

We combine this result with Lemma 4.3 to conclude the first result. For the second result, since the step is reliable, (4.5) improves to

$$-\frac{\theta}{2}\alpha_k \|g_k\|^2 \leq -\frac{\theta}{8}\alpha_k \|g_k\|^2 - \frac{\theta}{8}\delta_k^2 - \frac{\theta}{4(\kappa_g \alpha_{\max}+1)^2}\alpha_k \|\nabla f(x_k)\|^2 ,$$

and again the result follows from Lemma 4.3. □

4.2. Definition and analysis of $\{\Phi_k, \mathcal{A}_k, W_k\}$ process for Algorithm 1. We base our proof of convergence on properties of the random function

$$\Phi_k = \nu(f(X_k) - f_{\min}) + (1-\nu)\frac{1}{L^2}\mathcal{A}_k \|\nabla f(X_k)\|^2 + (1-\nu)\theta\Delta_k^2 \tag{4.6}$$

for some (deterministic) $\nu \in (0,1)$ and $f_{\min} \leq f(x)$ for all x. The goal is to show that $\{\Phi_k, \mathcal{A}_k\}$ satisfies Assumption 3.2, in particular, that Φ_k is expected to decrease on each iteration. Due to inaccuracy in function estimates and gradients, the algorithm may take a step that increases the objective and thus Φ_k. We will show that such increases are bounded by a value proportional to $\|\nabla f(x)\|^2$. On the other hand, as we will show, on successful steps with accurate function estimates, the objective decreases proportionally to $\|\nabla f(x)\|^2$, while on unsuccessful steps, Φ_k always decreases because both $\mathcal{A}_k$ and Δ_k are decreased. The function Φ is chosen to balance the potential increases and decreases in the objective with changes inflicted by unsuccessful steps.

THEOREM 4.6. *Let Assumptions* 2.1 *and* 2.4 *hold. Suppose the random process generated by Algorithm* 1 *is* $\{G_k, X_k, \mathcal{A}_k, \Delta_k, S_k, F_k^0, F_k^s\}$. *Then there exist probabilities* $p_g, p_f > 1/2$ *and a constant* $\nu \in (0,1)$ *such that the expected decrease in* Φ_k *is*

$$\mathbf{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{G \cdot F}] \leq -\frac{p_g p_f (1-\nu)(1-\gamma^{-1})}{4}\left(\frac{\mathcal{A}_k}{L^2}\|\nabla f(X_k)\|^2 + \theta \Delta_k^2\right). \tag{4.7}$$

In particular, the constant ν *and probabilities* $p_g, p_f > 1/2$ *satisfy*

$$\frac{\nu}{1-\nu} \geq \max\left\{\frac{32\gamma\alpha_{\max}^2}{\theta}, 16(\gamma-1), \frac{16\gamma(\kappa_g\alpha_{\max}+1)^2}{\theta}\right\}, \tag{4.8}$$

$$p_g \geq \frac{2\gamma}{1/2(1-\gamma^{-1})+2\gamma}, \tag{4.9}$$

$$\textit{and} \qquad \frac{p_g p_f}{\sqrt{1-p_f}} \geq \max\left\{\frac{8L^2\nu\kappa_f + 16\gamma(1-\nu)}{(1-\nu)(1-\gamma^{-1})}, \frac{8\nu}{(1-\nu)(1-\gamma^{-1})}\right\}. \tag{4.10}$$

Proof of Theorem 4.6. Our proof considers three separate cases: good gradients/ good estimates, bad gradients/good estimates, and lastly bad estimates. Each of these cases will be broken down into whether a successful/unsuccessful step is reliable/unreliable. To simplify notation, we introduce three sets:

$$\text{Succ} := \{\text{at iteration } k \text{ the step is successful, namely, sufficient decrease occurs}\},$$

$$\text{R} := \{\text{iteration } k \text{ the step is reliable, i.e., } \mathcal{A}_k \|G_k\|^2 \geq \Delta_k^2\},$$

$$\text{and} \qquad \text{U} := \{\text{iteration } k \text{ the step is unreliable, i.e., } \mathcal{A}_k \|G_k\|^2 < \Delta_k^2\}.$$

First, we decompose the difference in Φ_k into three disjoint sets

$$\mathbf{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{G \cdot F}] = \mathbf{E}[(1_{I_k \cap J_k} + 1_{I_k^c \cap J_k} + 1_{J_k^c})(\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}^{G \cdot F}].$$

TABLE 1

We summarize the proof of Theorem 4.6 *by displaying the expected upper bound on* $\Phi_{k+1} - \Phi_k$ *up to constants. The proof considers cases: accurate gradients/functions estimates, bad gradients/accurate functions estimates, and bad function estimates. Each of these is further broken into whether the step was successful/unsuccessful.*

	Upper bound on $\mathbf{E}[\Phi_{k+1} - \Phi_k]$		
	Accurate gradients Accurate functions w/ prob. $p_g p_f$	Bad gradients Accurate functions w/ prob. $(1-p_g)p_f$	Bad functions w/ prob. $1-p_f$
Success	$-\frac{\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2 - \Delta_k^2$ decrease	$\frac{\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2$ increase	$\frac{\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2 + \Delta_k^2$ increase
Unsuccess	$-\frac{\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2 - \Delta_k^2$ decrease	$-\frac{\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2 - \Delta_k^2$ decrease	$-\frac{\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2 - \Delta_k^2$ decrease
Overall worst case improv.	decrease	increase	increase

For each case we will derive a bound on the expected decrease (increase) in Φ_k. These bounds are derived in the proof below and are summarized in Table 1.

Case 1 (accurate gradients and estimates, $\mathbf{1}_{I_k \cap J_k} = 1$). We will show that the Φ_k decreases no matter what type of step occurs and that the smallest decrease happens on the unsuccessful step. Thus this case dominates the other two and overall we will conclude that

$$\begin{aligned}\mathbf{E}[1_{I_k\cap J_k}(\Phi_{k+1} - \Phi_k)|\mathcal{F}_{k-1}^{G\cdot F}] \\ \le -p_g p_f(1-\nu)(1-\gamma^{-1})\left(\frac{\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2 + \theta\Delta_k^2\right).\end{aligned} \tag{4.11}$$

(i) *Successful and reliable step* ($1_{Succ}1_R = 1$). The step is successful, and both the gradient and function estimates are accurate so a decrease in the true objective occurs; specifically, (4.4) from Lemma 4.5 applies:

$$\begin{aligned}&1_{I_k\cap J_k}1_{\mathrm{Succ}}1_{\mathrm{R}}\nu(f(X_{k+1}) - f(X_k)) \\ &\quad \le -\nu 1_{I_k\cap J_k}1_{\mathrm{Succ}}1_{\mathrm{R}}\left(\frac{\theta\mathcal{A}_k}{8}\left\|G_k\right\|^2 + \frac{\theta}{8}\Delta_k^2 + \frac{\theta\mathcal{A}_k}{4(\kappa_g\alpha_{\max}+1)^2}\left\|\nabla f(X_k)\right\|^2\right).\end{aligned} \tag{4.12}$$

As the step is successful, the difference in $\mathcal{A}_k\left\|\nabla f(X_k)\right\|^2$ may increase, but its change is bounded due to Lemma 4.4:

$$\begin{aligned}&1_{I_k\cap J_k}1_{\mathrm{Succ}}1_{\mathrm{R}}(1-\nu)\frac{1}{L^2}\left(\mathcal{A}_{k+1}\left\|\nabla f(X_{k+1})\right\|^2 - \mathcal{A}_k\left\|\nabla f(X_k)\right\|^2\right) \\ &\qquad \le 1_{I_k\cap J_k}1_{\mathrm{Succ}}1_{\mathrm{R}}(1-\nu)2\gamma\mathcal{A}_k\left(\alpha_{\max}^2\left\|G_k\right\|^2 + \tfrac{1}{L^2}\left\|\nabla f(X_k)\right\|^2\right).\end{aligned} \tag{4.13}$$

Lastly because we have a reliable step, $\Delta_{k+1}^2 = \gamma\Delta_k^2$. Consequently, we deduce that

$$1_{I_k\cap J_k}1_{\mathrm{Succ}}1_{\mathrm{R}}(1-\nu)\theta(\Delta_{k+1}^2 - \Delta_k^2) = 1_{I_k\cap J_k}1_{\mathrm{Succ}}1_{\mathrm{R}}(1-\nu)\theta(\gamma-1)\Delta_k^2. \tag{4.14}$$

Without loss of generality, suppose $L^2 \geq 1$. We choose ν sufficiently large so that the term on the right-hand side of (4.12) dominates the sum of the right-hand sides of (4.13), and (4.14), specifically,

$$-\frac{\nu\theta\mathcal{A}_k}{8}\left\|G_k\right\|^2 + (1-\nu)2\gamma\mathcal{A}_k\alpha_{\max}^2\left\|G_k\right\|^2 \leq -\frac{\nu\theta\mathcal{A}_k}{16}\left\|G_k\right\|^2,$$
$$-\frac{\nu\theta\mathcal{A}_k}{4L^2(\kappa_g\alpha_{\max}+1)^2}\left\|\nabla f(X_k)\right\|^2 + (1-\nu)\frac{2\gamma\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2 \leq -\frac{\nu\theta\mathcal{A}_k}{8L^2(\kappa_g\alpha_{\max}+1)^2}\left\|\nabla f(X_k)\right\|^2,$$
$$\text{and} \qquad -\frac{\nu\theta}{8}\Delta_k^2 + (1-\nu)(\gamma-1)\theta\Delta_k^2 \leq -\frac{\nu\theta}{16}\Delta_k^2, \tag{4.15}$$

which is satisfied if (4.8) holds. We combine (4.12), (4.13), and (4.14) to conclude

$$\begin{aligned}&1_{I_k\cap J_k}1_{\text{Succ}}1_{\text{R}}(\Phi_{k+1}-\Phi_k)\\ &\qquad\leq -1_{I_k\cap J_k}1_{\text{Succ}}1_{\text{R}}\left(\frac{\nu\theta\mathcal{A}_k}{8L^2(\kappa_g\alpha_{\max}+1)^2}\left\|\nabla f(X_k)\right\|^2 + \frac{\nu\theta}{16}\Delta_k^2\right).\end{aligned}\tag{4.16}$$

(ii) *Successful and unreliable step ($1_{Succ}1_U = 1$).* Because the step is successful and our gradient/estimates are accurate, we again apply Lemma 4.5 to bound $f(X_{k+1}) - f(X_k)$ but this time using (4.3) which holds for unreliable steps. The possible increase from the change in the $\left\|\nabla f(X_k)\right\|^2$ term is the same as (4.13) where we replace 1_{R} with 1_{U} since Lemma 4.4 still applies. Lastly with an unreliable step, the change in Δ_k^2 is

$$\begin{aligned}&1_{I_k\cap J_k}1_{\text{Succ}}1_{\text{U}}(1-\nu)\theta(\Delta_{k+1}^2-\Delta_k^2)\\ &\qquad\leq -1_{I_k\cap J_k}1_{\text{Succ}}1_{\text{U}}(1-\nu)(1-\gamma^{-1})\theta\Delta_k^2.\end{aligned}\tag{4.17}$$

Therefore by choosing ν such that (4.15) holds, we have that

$$\begin{aligned}&1_{I_k\cap J_k}1_{\text{Succ}}1_{\text{U}}(\Phi_{k+1}-\Phi_k)\\ &\qquad\leq -1_{I_k\cap J_k}1_{\text{Succ}}1_{\text{U}}\left(\frac{\nu\theta\mathcal{A}_k\left\|\nabla f(X_k)\right\|^2}{8L^2(\kappa_g\alpha_{\max}+1)^2} + (1-\nu)(1-\gamma^{-1})\theta\Delta_k^2\right).\end{aligned}\tag{4.18}$$

(iii) *Unsuccessful step ($1_{Succ^c} = 1$).* Because the step is unsuccessful, the change in the function values is 0 and the constants $\mathcal{A}_k$ and Δ_k^2 decrease. Consequently, we deduce that

$$\begin{aligned}&1_{I_k\cap J_k}1_{\text{Succ}^c}(\Phi_{k+1}-\Phi_k)\\ &\qquad\leq -1_{I_k\cap J_k}1_{\text{Succ}^c}(1-\nu)(1-\gamma^{-1})\left(\frac{\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2 + \theta\Delta_k^2\right).\end{aligned}\tag{4.19}$$

We chose ν sufficiently large to ensure that the third case (iii), unsuccessful step (4.19), provides the worst case decrease when compared to (4.16) and (4.18). Specifically the constant ν in (4.8) was chosen so that

$$\begin{aligned}&\frac{-\nu\theta\mathcal{A}_k}{8L^2(\kappa_g\alpha_{\max}+1)^2}\left\|\nabla f(X_k)\right\|^2 \leq -(1-\nu)(1-\gamma^{-1})\frac{\mathcal{A}_k}{L^2}\left\|\nabla f(X_k)\right\|^2\\ &\text{and} \qquad \frac{-\nu\theta}{16}\Delta_k^2 \leq -(1-\nu)(1-\gamma^{-1})\theta\Delta_k^2.\end{aligned}\tag{4.20}$$

As such, we bounded the change in Φ_k in the case of accurate gradients and estimates by

$$1_{I_k \cap J_k}(\Phi_{k+1} - \Phi_k) \leq -1_{I_k \cap J_k}(1-\nu)(1-\gamma^{-1}) \left(\frac{\mathcal{A}_k}{L^2} \|\nabla f(X_k)\|^2 + \theta \Delta_k^2 \right). \tag{4.21}$$

We take conditional expectations with respect to $\mathcal{F}_{k-1}^{G \cdot F}$, and using Assumption 2.4, (4.11) holds.

Case 2 (bad gradients and accurate estimates, $\mathbf{1}_{I_k^c \cap J_k} = \mathbf{1}$). Unlike the previous case, the difference in the Φ_k may increase, since the step along an inaccurate probabilistic gradients may not provide enough decrease to cancel the increase from the $\|\nabla f(X_k)\|^2$. Precisely, the successful and unreliable case dominates the worst case increase in the difference of the Φ_k:

$$\mathbf{E}[1_{I_k^c \cap J_k}(\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}^{G \cdot F}] \leq (1-p_g)(1-\nu) \frac{2\gamma \mathcal{A}_k}{L^2} \|\nabla f(X_k)\|^2. \tag{4.22}$$

As before, we consider three separate cases.

(i) *Successful and reliable step* ($1_{Succ} 1_R = 1$). A successful, reliable step with accurate function estimates but bad gradients has functional improvement (see Lemma 4.3, (4.2)):

$$1_{I_k^c \cap J_k} 1_{\text{Succ}} 1_{\text{R}} \nu (f(X_{k+1}) - f(X_k)) \leq -1_{I_k^c \cap J_k} 1_{\text{Succ}} 1_{\text{R}} \nu \left(\frac{\mathcal{A}_k \theta \|G_k\|^2}{4} + \frac{\theta}{4} \Delta_k^2 \right).$$

In contrast to (4.12), we lose the $\|\nabla f(X_k)\|^2$ term. A reliable, successful step increases both constants $\mathcal{A}_{k+1}$ and Δ_{k+1}^2, leading to (4.13) and (4.14) with $1_{I_k \cap J_k}$ replaced by $1_{I_k^c \cap J_k}$. Hence by choosing ν to satisfy (4.8), the dominant term in the difference in the Φ_k is

$$\begin{aligned} &1_{I_k^c \cap J_k} 1_{\text{Succ}} 1_{\text{R}} (\Phi_{k+1} - \Phi_k) \\ &\quad \leq 1_{I_k^c \cap J_k} 1_{\text{Succ}} 1_{\text{R}} \left(-\frac{\nu \theta \mathcal{A}_k}{16} \|G_k\|^2 - \frac{\nu \theta}{16} \Delta_k^2 + \frac{2\gamma(1-\nu)}{L^2} \mathcal{A}_k \|\nabla f(X_k)\|^2 \right). \end{aligned} \tag{4.23}$$

(ii) *Successful and unreliable step* ($1_{Succ} 1_U = 1$). Lemma 4.3 holds, but this time (4.1) for unreliable steps applies. Moreover, (4.13) and (4.17) that bound the change in the last two terms of Φ_k also apply. Again by choosing ν to satisfy (4.8), we deduce

$$\begin{aligned} &1_{I_k^c \cap J_k} 1_{\text{Succ}} 1_{\text{U}} (\Phi_{k+1} - \Phi_k) \\ &\quad \leq 1_{I_k^c \cap J_k} 1_{\text{Succ}} 1_{\text{U}} \left(-\frac{\nu \theta \mathcal{A}_k \|G_k\|^2}{16} - (1-\nu)(1-\gamma^{-1}) \theta \Delta_k^2 \right. \\ &\qquad \left. + \frac{2\gamma(1-\nu) \mathcal{A}_k \|\nabla f(X_k)\|^2}{L^2} \right). \end{aligned} \tag{4.24}$$

(iii) *Unsuccessful* ($1_{Succ^c} = 1$). As in the previous case, (4.19) holds.

The right-hand sides of (4.23), (4.24), and (4.19) are trivially upper bounded by the positive term $\mathcal{A}_k \|\nabla f(X_k)\|^2$. Hence, we conclude that

$$1_{I_k^c \cap J_k}(\Phi_{k+1} - \Phi_k) \leq 1_{I_k^c \cap J_k} \frac{2\gamma(1-\nu)}{L^2} \mathcal{A}_k \|\nabla f(X_k)\|^2. \tag{4.25}$$

Inequality (4.22) follows by taking expectations with respect to $\mathcal{F}_{k-1}^{G\cdot F}$ and noting that $\mathbf{E}[1_{I_k^c\cap J_k}|\mathcal{F}_{k-1}^{M\cdot F}] \le 1-p_g$ as in Assumption 2.4.

Case 3 (bad estimates, $\mathbf{1}_{J_k^c} = 1$). Inaccurate estimates can cause the algorithm to accept a step which can lead to an increase in f, $\mathcal{A}$, and Δ and hence in the difference of the Φ_k. We control this increase in the difference of the Φ_k by bounding the variance in the function estimates, as in (2.3), which is the key reason for Assumption 2.4(iii). By choosing the probability of J_k^c to be sufficiently small, we can ensure that, in expectation, the difference in the Φ_k is sufficiently reduced. Precisely, we will show

$$\begin{aligned}\mathbf{E}[1_{J_k^c}(\Phi_{k+1}-\Phi_k)|\mathcal{F}_{k-1}^{G\cdot F}] &\le 2\nu(\sqrt{1-p_f})\max\left\{\kappa_f\mathcal{A}_k\left\|\nabla f(X_k)\right\|^2,\theta\Delta_k^2\right\}\\ &\quad + (1-p_f)\frac{(1-\nu)2\gamma}{L^2}\mathcal{A}_k\left\|\nabla f(X_k)\right\|^2.\end{aligned}\tag{4.26}$$

A successful step leads to the following bound:

$$\begin{aligned}&1_{J_k^c}1_{\mathrm{Succ}}\nu\left(f(X_{k+1})-f(X_k)\right)\\ &\qquad\le 1_{J_k^c}1_{\mathrm{Succ}}\nu\left((F_k^s-F_k^0)+|f(X_{k+1})-F_k^s|+|F_k^0-f(X_k)|\right)\\ &\qquad\le 1_{J_k^c}1_{\mathrm{Succ}}\nu\left(-\theta\mathcal{A}_k\left\|G_k\right\|^2+|f(X_{k+1})-F_k^s|+|F_k^0-f(X_k)|\right),\end{aligned}\tag{4.27}$$

where the last inequality is due to the sufficient decrease condition. As before, we consider three separate cases.

(i) *Successful and reliable step* ($1_{Succ}1_R = 1$). With a reliable step we have $-\mathcal{A}_k\left\|G_k\right\|^2 \le -\Delta_k^2$; thus (4.27) implies

$$\begin{aligned}&1_{J_k^c}1_{\mathrm{Succ}}1_{\mathrm{R}}\nu(f(X_{k+1})-f(X_k))\\ &\qquad\le 1_{J_k^c}1_{\mathrm{Succ}}1_{\mathrm{R}}\nu\left(-\frac{1}{2}\theta\mathcal{A}_k\left\|G_k\right\|^2-\frac{\theta}{2}\Delta_k^2+|f(X_{k+1})-F_k^s|+|F_k^0-f(X_k)|\right).\end{aligned}$$

We note that $\Phi_{k+1}-\Phi_k$ is upper bounded by the sum of the right-hand side of the above inequality and the right-hand sides of (4.13) and (4.14). As before, by choosing ν as in (4.8) we ensure $\frac{-\nu\theta}{2}\mathcal{A}_k\left\|G_k\right\|^2+(1-\nu)2\gamma\mathcal{A}_k\alpha_{\max}^2\left\|G_k\right\|^2 \le 0$ and $\frac{-\nu\theta}{2}\Delta_k^2+(1-\nu)(\gamma-1)\theta\Delta_k^2 \le 0$. It follows that

$$\begin{aligned}&1_{J_k^c}1_{\mathrm{Succ}}1_{\mathrm{R}}(\Phi_{k+1}-\Phi_k)\\ &\qquad\le 1_{J_k^c}\left(\nu|f(X_{k+1})-F_k^s|+\nu|F_k^0-f(X_k)|+(1-\nu)\frac{2\gamma\mathcal{A}_k\left\|\nabla f(X_k)\right\|^2}{L^2}\right).\end{aligned}\tag{4.28}$$

(ii) *Successful and unreliable step* ($1_{Succ}1_U = 1$). Since on unreliable steps, Δ_{k+1}^2 is decreased, then the increase in the difference of the Φ_k is always smaller than the worst-case increase we just derived for the successful and reliable step. Thus (4.28) holds with 1_{R} replaced by 1_{U}.

(iii) *Unsuccessful* ($1_{Succ^c} = 1$). As we decrease both Δ and $\mathcal{A}$, and $X_{k+1} = X_k$, we conclude that (4.19) holds.

The equation (4.28) dominates (4.19); thus in all three cases (4.28) holds. We take expectations of (4.28) and apply Lemma 2.5 to conclude that

$$\begin{aligned}\mathbf{E}[1_{J_k^c}(\Phi_{k+1}-\Phi_k)|\mathcal{F}_{k-1}^{G\cdot F}] &\le 2\nu(1-p_f)^{1/2}\max\left\{\kappa_f\mathcal{A}_k\left\|\nabla f(X_k)\right\|^2,\theta\Delta_k^2\right\}\\ &\quad + (1-p_f)(1-\nu)\frac{2\gamma}{L^2}\mathcal{A}_k\left\|\nabla f(X_k)\right\|^2.\end{aligned}\tag{4.29}$$

Now we combine the expectations (4.11), (4.22), and (4.26) to obtain

$$\begin{aligned}\mathbf{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{G\cdot F}] &= \mathbf{E}[(1_{I_k \cap J_k} + 1_{I_k^c \cap J_k} + 1_{J_k^c})(\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}^{G\cdot F}] \\ &\leq -p_g p_f (1-\nu)(1-\gamma^{-1}) \left(\frac{\mathcal{A}_k \|\nabla f(X_k)\|^2}{L^2} + \theta \Delta_k^2 \right) \\ &\quad + p_f(1-p_g) \frac{2\gamma(1-\nu)\mathcal{A}_k \|\nabla f(X_k)\|^2}{L^2} \\ &\quad + 2\nu(1-p_f)^{1/2} \left(\kappa_f \mathcal{A}_k \|\nabla f(X_k)\|^2 + \theta \Delta_k^2 \right) \\ &\quad + (1-p_f)^{1/2} \cdot \frac{4\gamma(1-\nu)\mathcal{A}_k}{L^2} \|\nabla f(X_k)\|^2 ,\end{aligned}$$

where the inequality follows from $1 - p_f \leq (1-p_f)^{1/2}$ and $1 - p_g = p_f(1-p_g) + (1-p_f)(1-p_g) \leq p_f(1-p_g) + (1-p_f)^{1/2}$. Let us choose $p_g \in (0,1]$ as in (4.9) which implies

$$\begin{aligned}&\left(-p_g p_f \frac{(1-\nu)(1-\gamma^{-1})\mathcal{A}_k}{L^2} + p_f(1-p_g) \frac{2\gamma(1-\nu)\mathcal{A}_k}{L^2} \right) \|\nabla f(X_k)\|^2 \\ &\qquad \leq -p_g p_f \frac{(1-\nu)(1-\gamma^{-1})\mathcal{A}_k}{2L^2} \|\nabla f(X_k)\|^2 .\end{aligned}$$

We have now reduced the number of terms in the conditional expectation:

$$\begin{aligned}\mathbf{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{G\cdot F}] &\leq -p_g p_f \frac{1}{2} (1-\nu)(1-\gamma^{-1}) \left(\frac{\mathcal{A}_k}{L^2} \|\nabla f(X_k)\|^2 + \theta \Delta_k^2 \right) \\ &\quad + 2\nu(1-p_f)^{1/2} \left(\kappa_f \mathcal{A}_k \|\nabla f(X_k)\|^2 + \theta \Delta_k^2 \right) \\ &\quad + (1-p_f)^{1/2} \cdot \frac{4\gamma(1-\nu)\mathcal{A}_k}{L^2} \|\nabla f(X_k)\|^2 .\end{aligned}$$

We choose $p_f \in (0,1]$ large enough so that $\frac{p_g p_f}{\sqrt{1-p_f}}$ satisfies (4.10) which implies

$$\begin{aligned}&\left(-\frac{p_g p_f (1-\nu)(1-\gamma^{-1})}{2L^2} + (1-p_f)^{1/2} \left(2\nu\kappa_f + \frac{4\gamma(1-\nu)}{L^2} \right) \right) \mathcal{A}_k \|\nabla f(X_k)\|^2 \\ &\qquad \leq -\frac{p_g p_f (1-\nu)(1-\gamma^{-1})\mathcal{A}_k}{4L^2} \|\nabla f(X_k)\|^2\end{aligned}$$

and $\frac{-p_g p_f}{2}(1-\nu)(1-\gamma^{-1})\theta\Delta_k^2 + 2\nu(1-p_f)^{1/2}\theta\Delta_k^2 \leq \frac{-p_g p_f}{4}(1-\nu)(1-\gamma^{-1})\theta\Delta_k^2$. The proof is complete. ∎

Remark 4. To simplify the expression for the constants we will assume that $\theta = 1/2$ and $\gamma = 2$ which are typical values for these constants. We also assume that without loss of generality $\kappa_g \geq 2$ and $\nu \geq 1/2$. The analysis can be performed for any other values of the above constants—the choices here are for simplicity and illustration. The conditions on p_g and p_f under the above choice of constants will be shown below.

THEOREM 4.7. *Let Assumptions* 2.1 *and* 2.4 *hold, and chose constants as in Remark* 4. *Suppose* $\{G_k, X_k, \mathcal{A}_k, \Delta_k, S_k, F_k^0, F_k^s\}$ *is the random process generated by Algorithm* 1. *Then there exist probabilities* p_g, p_f *and a constant* $\nu \geq 1/2$ *such that the expected decrease in* Φ_k *is*

$$\mathbf{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{G\cdot F}] \leq -\frac{1}{2048(\kappa_g \alpha_{\max} + 1)^2} \left(\frac{\mathcal{A}_k}{L^2} \|\nabla f(X_k)\|^2 + \frac{1}{2}\Delta_k^2 \right). \tag{4.30}$$

In particular, the constant ν and probabilities p_g, p_f must satisfy

$$\frac{\nu}{1-\nu} = 64(\kappa_g \alpha_{\max} + 1)^2, \tag{4.31}$$

$$p_g \geq \frac{16}{17} \quad \textit{and} \quad p_f > \frac{1}{2}, \tag{4.32}$$

$$\textit{and} \quad \frac{p_g p_f}{\sqrt{1-p_f}} \geq \max\left\{1024\kappa_f L^2(\kappa_g\alpha_{\max}+1)^2 + 64, 1024(\kappa_g\alpha_{\max}+1)^2\right\}. \tag{4.33}$$

Proof. We plug in the values for γ and θ and use the fact that $\kappa_g \geq 2$ to obtain the expression for $\nu/(1-\nu)$ and p_g. In order to deduce the expression for $p_g p_f/(1-p_f)^{1/2}$, we assume that $\nu/(1-\nu) = 64(\kappa_g\alpha_{\max}+1)^2$. Lastly, we suppose $\nu > 1/2$, $p_g p_f \geq 1/2$, and $\frac{\nu}{64(\kappa_g\alpha_{\max}+1)2} = (1-\nu)$. Therefore, we have

$$\frac{-p_g p_f(1-\nu)(1-\gamma^{-1})}{4} \leq \frac{-(1-\nu)}{16} \leq \frac{-\nu}{1024(\kappa_g\alpha_{\max}+1)^2} \leq \frac{1}{2048(\kappa_g\alpha_{\max}+1)^2}.$$

The result is shown. ∎

4.3. Convergence rates for the nonconvex case. Our primary goal in this paper is to bound the expected number of steps that the algorithm takes until $\|\nabla f(X_k)\| \leq \varepsilon$. Define the stopping time

$$T_\varepsilon = \inf\{k \geq 0 \,:\, \|\nabla f(X_k)\| < \varepsilon\}.$$

We show in this section, under the simplified assumptions on the constant from Theorem 4.7,

$$\mathbf{E}[T_\varepsilon] \leq \mathcal{O}(1) \cdot \frac{p_g p_f}{2p_g p_f - 1} \cdot \frac{L^3(\kappa_g\alpha_{\max}+1)^2\Phi_0}{\varepsilon^2} + 1.$$

Here $\mathcal{O}(1)$ hides universal constants and dependencies on θ, γ, $\alpha_{\max}$. We derive this result from Theorem 3.3; therefore, the remainder of this section is devoted to showing Assumption 3.2 holds. Given Theorem 4.6, it follows immediately that the random variable Φ_k defined as in (4.6) satisfies Assumption 3.2(iii) by multiplying both sides by the indicator, $1_{\{T_\varepsilon > k\}}$. In particular, we define the function $h(\mathcal{A}_k) = \mathcal{A}_k\varepsilon^2$ to obtain from Theorem 4.7

$$\mathbf{E}[1_{\{T_\varepsilon>k\}}(\Phi_{k+1} - \Phi_k)|\mathcal{F}_{k-1}^{G\cdot F}] \leq -\Theta h(\mathcal{A}_k) 1_{\{T_\varepsilon > k\}},$$

where $\Theta = \frac{1}{2048L^2(\kappa_g\alpha_{\max}+1)^2}$. It remain to show Assumption 3.2(ii) holds.

Lemma 4.8. *Let p_g and p_f be such that $p_g p_f \geq 1/2$; then Assumption* 3.2(ii) *is satisfied for $W_k = 2(1_{I_k \cap J_k} - 1/2)$, $\lambda = \log(\gamma)$, $p = p_g p_f$, and*

$$\bar{\mathcal{A}} = \frac{1-\theta}{\kappa_g + \frac{L}{2} + 2\varepsilon_f}.$$

Proof. We can shrink $\bar{\mathcal{A}}$, without loss of generality, so that $\bar{\mathcal{A}} = \mathcal{A}_0 e^{\lambda\bar{j}}$ for some $\bar{j} \in \mathbb{Z}$ and $\bar{j} \leq 0$. It remains to show that

$$1_{\{T_\varepsilon>k\}}\mathcal{A}_{k+1} \geq 1_{\{T_\varepsilon>k\}} \min\left\{\bar{\mathcal{A}}, \min\{\alpha_{\max}, \gamma\mathcal{A}_k\} I_k J_k + \gamma^{-1}\mathcal{A}_k(1 - 1_{I_k\cap J_k})\right\}.$$

Suppose $\mathcal{A}_k > \bar{\mathcal{A}}$. Then $\mathcal{A}_k \geq \gamma\bar{\mathcal{A}}$, and hence $\mathcal{A}_{k+1} \geq \bar{\mathcal{A}}$. Now, assume that $\mathcal{A}_k \leq \bar{\mathcal{A}}$. If we have $1_{I_k} = 1$ and $1_{J_k} = 1$, it follows from Lemma 4.2 that the kth step is successful, i.e., $x_{k+1} = x_k + s_k$ and $\alpha_{k+1} = \max\{\alpha_{\max}, \gamma\alpha_k\}$. If $I_k J_k = 0$, then $\alpha_{k+1} \geq \gamma^{-1}\alpha_k$. ∎

Finally substituting the expressions for h, $\bar{\mathcal{A}}$, and Φ_k into the bound on $\mathbf{E}[T_\varepsilon]$ from Theorem 3.3 we obtain the following complexity result.

Theorem 4.9. *Under the assumptions in Theorem* 4.7*, suppose the probabilities* p_g, p_f *satisfy*

$$p_g \geq \frac{16}{17}, \qquad p_f > \frac{1}{2}, \qquad \textit{and}$$

$$\frac{p_g p_f}{\sqrt{1-p_f}} \geq \max\left\{1024\kappa_f L^2(\kappa_g \alpha_{\max}+1)^2 + 64, 1024(\kappa_g\alpha_{\max}+1)^2\right\}$$

with $\frac{\nu}{1-\nu} = 64(\kappa_g\alpha_{\max}+1)^2$*. Then the expected number of iterations that Algorithm* 1 *takes until* $\|\nabla f(X_k)\|^2 \leq \varepsilon$ *occurs is bounded as follows:*

$$\mathbf{E}[T_\varepsilon] \leq \frac{p_g p_f}{2p_g p_f - 1} \cdot \frac{L^2(\kappa_g + L/2 + 2\varepsilon_f)(\kappa_g \alpha_{\max}+1)^2}{C\varepsilon^2}\Phi_0 + 1,$$

where $C = 1/4096$ *and* $\Phi_0 = \nu(f(X_0) - f_{\min}) + (1-\nu)(1/L^2\mathcal{A}_0 \|\nabla f(X_0)\|^2 + 1/2\Delta_0^2)$*.*

As a simple corollary to the complexity results we have the liminf-type a.s. convergence result.

Theorem 4.10. *Let the assumptions of Theorem* 4.6 *(or Theorem* 4.7*) hold. Then the sequence of random iterates generated by Algorithm* 1*,* X_k*, almost surely satisfy*

$$\liminf_{k\to\infty} \|\nabla f(X_k)\| = 0.$$

4.4. Convex case. We now analyze line search (Algorithm 1) under the setting that the objective function is convex.

Assumption 4.11. Suppose, in addition to Assumption 2.1, the function f in (1.1) is convex. We also assume there exists a constant $D > 0$ such that

$$\|x - x^*\| \leq D \quad \text{for all } x \in \mathcal{U},$$

where x^* is some global minimizer of f and the set $\mathcal{U}$ contains all iteration realizations. Moreover, we assume there exists a $L_f > 0$ such that $\|\nabla f(x)\| \leq L_f$ for all $x \in \mathcal{U}$.

Remark 5. In deterministic optimization, it is common to assume that the function f has bounded level sets and that all the iterates remain within the bounded set defined by $f(x) \leq f(x_0)$. For the stochastic case, it is not guaranteed that all the iterates remain in the bounded level set because it is possible to take steps that increase the function value. Clearly iterates remain in a (large enough) bounded set with high probability. Alternatively, if it is known that the optimal solution lies within some bounded set, Algorithm 1 can be simply modified to project iterates onto that set. This modified version for the convex case can be analyzed in an almost identical way as is done in Theorem 4.6. However, for simplicity of the presentation, for the convex case, we simply impose Assumption 4.11.

In the convex setting, the goal is to bound the expected number of iterations T_ε of Algorithm 1 until one reaches a nearly optimal value,

$$f(x_k) - f^* < \varepsilon.$$

In the convex and deterministic setting, the complexity bound is derived by showing that $1/(f(x_k) - f^*)$ has a constant increase until an ε-accurate functional decrease is reached. For the randomized line search we follow the same idea, replacing $f(x_k) - f^*$ in Φ_k (modified by substituting $f_{\min}$ in (4.6) by f^*) and defining the function

$$\Psi_k = \frac{1}{\varepsilon} - \frac{1}{\Phi_k + \varepsilon}, \tag{4.34}$$

where the constant $\varepsilon > 0$ is the same level of optimality as T_ε. To simplify the argument, we impose an upper bound on Δ_k.

Assumption 4.12. Suppose there exists a constant $\delta_{\max}$ such that the random variable $\Delta_k \leq \delta_{\max}$.

First, with a simple modification to Algorithm 1, we can impose this assumption. Second, the dynamics of the algorithm suggest Δ_k eventually decreases until it is smaller than any $\varepsilon > 0$.

We show the random process $\{\Psi_k, \mathcal{A}_k\}$ satisfies Assumption 3.2 for all $k \geq 0$. The dynamics of the random variables $\mathcal{A}_k$ behave the same as in the nonconvex setting; hence Assumption 3.2(i) and (ii) follow from Lemma 4.8. We ensure boundedness of the random process $\{\Psi_k\}$ by incorporating the optimality level directly into the definition of Ψ; hence the dependency on ε for complexity bounds is built directly into the function Ψ. The main component of this section is proving Assumption 3.2(iii) holds for this Ψ_k, i.e., an expected improvement occurs.

THEOREM 4.13. *Let Assumptions* 2.1, 2.4, 4.11, *and* 4.12 *hold. Suppose the random process generated by Algorithm* 1 *is* $\{G_k, X_k, \mathcal{A}_k, \Delta_k, S_k, F_k^0, F_k^s\}$. *Then there exist probabilities* p_g *and* p_f *and a constant* $\nu \in (0,1)$ *such that*

$$1_{\{T_\varepsilon > k\}} \cdot \mathbf{E}[\Psi_{k+1} - \Psi_k | \mathcal{F}_{k-1}^{G\cdot F}] \leq \frac{-p_g p_f (1-\nu)(1-\gamma^{-1})}{8((\nu+1)DL + \frac{(1-\nu)\alpha_{\max} L_f}{L} + (1-\nu)\sqrt{\theta}\delta_{\max})^2} \cdot \mathcal{A}_k 1_{\{T_\varepsilon > k\}},$$

where Ψ_k *is defined in* (4.34). *In particular, the probabilities* p_g *and* p_f *and constant* ν *in* (4.8), (4.9), *and* (4.10) *from Theorem* 4.6 *suffice.*

Proof. First, by convexity, we have that

$$\begin{aligned}
&1_{\{T_\varepsilon > k\}} \cdot (\Phi_k + \varepsilon) \\
&\leq 1_{\{T_\varepsilon > k\}} \cdot \left((\nu+1)(f(X_k) - f^*) + (1-\nu)\frac{\mathcal{A}_k \|\nabla f(X_k)\|^2}{L^2} + (1-\nu)\theta\Delta_k^2 \right) \\
&\leq 1_{\{T_\varepsilon > k\}} \cdot \left((\nu+1)\langle \nabla f(X_k), X_k - x^* \rangle + (1-\nu)\alpha_{\max}\frac{\|\nabla f(X_k)\|^2}{L^2} + (1-\nu)\theta\delta_{\max}\Delta_k \right) \\
&\leq 1_{\{T_\varepsilon > k\}} \cdot \left((\nu+1)DL + \frac{(1-\nu)\alpha_{\max}L_f}{L} + (1-\nu)\sqrt{\theta}\delta_{\max} \right) \left(\frac{\|\nabla f(X_k)\|}{L} + \sqrt{\theta}\Delta_k \right),
\end{aligned}$$

where we used $\|\nabla f(X_k)\| < L_f$. Without loss of generality, we assume $\alpha_{\max} \leq 1$; one may prove the same result with any step size, but for the sake of simplicity we will defer to the standard case when $\alpha_{\max} \leq 1$. By squaring both sides, we conclude

$$\begin{aligned}
\frac{1_{\{T_\varepsilon > k\}} \cdot \mathcal{A}_k(\Phi_k + \varepsilon)^2}{\tilde{C}} &:= \frac{1_{\{T_\varepsilon > k\}} \cdot \mathcal{A}_k(\Phi_k + \varepsilon)^2}{2((\nu+1)DL + \frac{(1-\nu)\alpha_{\max}L_f}{L} + (1-\nu)\sqrt{\theta}\delta_{\max})^2} \\
&\leq 1_{\{T_\varepsilon > k\}} \cdot \left(\mathcal{A}_k \frac{\|\nabla f(X_k)\|^2}{L^2} + \theta\Delta_k^2 \right),
\end{aligned} \tag{4.35}$$

where we used the inequality $(a+b)^2 \leq 2(a^2+b^2)$. From the above inequality combined with (4.7) we have

$$\mathbf{E}[1_{\{T_\varepsilon > k\}} \cdot (\Phi_{k+1} - \Phi_k) | \mathcal{F}_{k-1}^{G \cdot F}] \leq \frac{-p_g p_f (1-\nu)(1-\gamma^{-1})}{4\tilde{C}} \cdot 1_{\{T_\varepsilon > k\}} \cdot \mathcal{A}_k (\Phi_k + \varepsilon)^2.$$

We can then use Jensen's inequality applied to the function $x \mapsto \frac{1}{x}$ to derive the following bound:

$$\begin{aligned}
1_{\{T_\varepsilon > k\}} \cdot \mathbf{E}\left[\frac{1}{\Phi_k + \varepsilon} - \frac{1}{\Phi_{k+1} + \varepsilon} \Big| \mathcal{F}_{k-1}^{G \cdot F}\right] &\leq 1_{\{T_\varepsilon > k\}} \cdot \left(\frac{1}{\Phi_k + \varepsilon} - \frac{1}{\mathbf{E}[\Phi_{k+1} + \varepsilon | \mathcal{F}_{k-1}^{G \cdot F}]} \right) \\
&= 1_{\{T_\varepsilon > k\}} \cdot \left(\frac{\mathbf{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{G \cdot F}]}{(\Phi_k + \varepsilon)\mathbf{E}[\Phi_{k+1} + \varepsilon | \mathcal{F}_{k-1}^{G \cdot F}]} \right) \\
&\leq \frac{-p_g p_f (1-\nu)(1-\gamma^{-1})\mathcal{A}_k}{4\tilde{C}} \cdot \frac{(\Phi_k + \varepsilon)^2}{(\Phi_k + \varepsilon)\mathbf{E}[\Phi_{k+1} + \varepsilon | \mathcal{F}_{k-1}^{G \cdot F}]} \cdot 1_{\{T_\varepsilon > k\}} \\
&\leq \frac{-p_g p_f (1-\nu)(1-\gamma^{-1})\mathcal{A}_k}{4\tilde{C}} \cdot 1_{\{T_\varepsilon > k\}},
\end{aligned}$$

where the last inequality follows from $\mathbf{E}[\Phi_{k+1} + \varepsilon | \mathcal{F}_{k-1}^{G \cdot F}] \leq \Phi_k + \varepsilon$. □

The expected improvement in Ψ_k allows us to use Theorem 3.3 with $h(\mathcal{A}) = \mathcal{A}$, $\bar{\mathcal{A}} = \frac{1-\theta}{\kappa_g + L/2 + 2\varepsilon_f}$, $\Theta = \frac{p_g p_f (1-\nu)(1-\gamma^{-1})}{4\tilde{C}}$ where $\tilde{C}$ is defined in (4.35), and $p = p_g p_f$. This directly gives us the complexity bound.

THEOREM 4.14. *Let the assumptions of Theorem* 4.13 *hold with constant ν and probabilities p_f and p_g as in Theorem* 4.6. *Suppose we choose the constants as in Remark* 4. *Then the expected number of iterations that Algorithm* 1 *takes until $f(X_k) - f^* < \varepsilon$ is bounded as follows:*

$$\mathbf{E}[T_\varepsilon] \leq \mathcal{O}(1) \cdot \frac{p_g p_f}{2p_g p_f - 1} \cdot \frac{(\kappa_g \alpha_{\max} + 1)^2 (\kappa_g + L + \varepsilon_f)\left((\nu+1)DL + (1-\nu)\left(\frac{\alpha_{\max} L_f}{L} + \sqrt{\theta}\delta_{\max}\right)\right)^2}{\varepsilon + \Phi_0}.$$

The bound in Theorem 4.14 can be further simplified as follows:

$$\mathbf{E}[T_\varepsilon] \leq \mathcal{O}(1) \cdot \frac{p_g p_f}{2p_g p_f - 1} \left(\frac{L^3 \kappa_g^3 (D^2 + L_f^2 + \delta_{\max}^2)}{\varepsilon} \right).$$

4.5. Strongly convex case. Lastly, we analyze the stochastic line search (Algorithm 1) under the setting that the objective function is strongly convex. As such, we assume the following is now true of the objective function while dropping Assumption 4.11 and the bound on Δ_k.

Assumption 4.15. Suppose that, in addition to Assumption 2.1, the function f is μ-strongly convex, namely, for all $x, y \in \mathbb{R}^n$ the following inequality holds:

$$f(x) \geq f(y) + \nabla f(y)^T (x-y) + \frac{\mu}{2} \|x - y\|^2.$$

Our goal, like the convex setting, is to bound the expected number of iterations T_ε until $f(x) - f^* < \varepsilon$. We show that this bound is of the order of $\log(1/\varepsilon)$, as in the deterministic case. Our proof follows the same technique used in the deterministic setting which relies on showing that $\log(f(x_k) - f^*)$ decreases by a constant at each iteration. Here, instead of tracking the decrease in $\log(f(x_k) - f^*)$, we define the function

$$\Psi_k = \log(\Phi_k + \varepsilon) + \log\left(\frac{1}{\varepsilon}\right), \tag{4.36}$$

where the constant $\varepsilon > 0$ is the same level of optimality as T_ε and Φ_k is defined in (4.6). We show the random process $\{\Psi_k, \mathcal{A}_k\}$ satisfies Assumption 3.2. Again, the dynamics of $\mathcal{A}_k$ do not change and $\Psi \geq 0$ since we incorporated the optimality condition directly into the definition of Ψ. Hence Assumption 3.2(i) and (ii) hold. The next result shows the expected decrease of Ψ_k (Assumption 3.2(iii)).

THEOREM 4.16. *Let Assumptions* 2.1, 2.4, *and* 4.15 *hold. Suppose the random process generated by Algorithm* 1 *is* $\{G_k, X_k, \mathcal{A}_k, \Delta_k, S_k, F_k^0, F_k^s\}$. *The expected improvement is*

$$1_{\{T_\varepsilon > k\}} \cdot \mathbf{E}[\Psi_{k+1} - \Psi_k | \mathcal{F}_{k-1}^{G\cdot F}] \leq -\frac{p_g p_f (1-\nu)(1-\gamma^{-1})}{4(\frac{L^2(\nu+1)}{4\mu} + (1-\nu)\alpha_{\max} + (1-\nu))} \mathcal{A}_k \cdot 1_{\{T_\varepsilon > k\}},$$

where Ψ_k *is defined in* (4.36) *and the probabilities* p_g *and* p_f *and constant* ν *are defined in Theorem* 4.6.

Proof. By strong convexity, for all x, we have $f(x) - f^* \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$; hence we obtain

$$\begin{aligned} 1_{\{T_\varepsilon > k\}} \cdot (\Phi_k + \varepsilon) &\leq 1_{\{T_\varepsilon > k\}} \cdot \left((\nu+1)(f(X_k) - f^*) + (1-\nu)\left(\frac{\mathcal{A}_k \|\nabla f(X_k)\|^2}{L^2} + \theta \Delta_k^2 \right) \right) \\ &\leq 1_{\{T_\varepsilon > k\}} \cdot \left(\left(\frac{(\nu+1)L^2}{2\mu} + (1-\nu)\alpha_{\max} \right) \frac{\|\nabla f(X_k)\|^2}{L^2} + (1-\nu)\theta\Delta_k^2 \right) \\ &\leq 1_{\{T_\varepsilon > k\}} \cdot \left(\left(\frac{(\nu+1)L^2}{2\mu} + (1-\nu)(1+\alpha_{\max}) \right) \left(\frac{\|\nabla f(X_k)\|^2}{L^2} + \theta\Delta_k^2 \right) \right). \end{aligned}$$

For simplicity of notation, we define

$$\tilde{C} := \tfrac{(\nu+1)L^2}{2\mu} + (1-\nu)(1+\alpha_{\max}). \tag{4.37}$$

Also for simplicity and without loss of generality, we assume $\alpha_{\max} \leq 1$; hence, we conclude

$$1_{\{T_\varepsilon > k\}} \cdot \mathcal{A}_k(\Phi_k + \varepsilon) \leq 1_{\{T_\varepsilon > k\}} \cdot \tilde{C} \left(\frac{\mathcal{A}_k \|\nabla f(X_k)\|^2}{L^2} + \theta\Delta_k^2 \right).$$

We have an expected decrease in Φ_k (4.7) from Theorem 4.6. This together with the equality $1_{\{T_\varepsilon > k\}}(\Phi_{k+1} - \Phi_k) = \Phi_{(k+1)\wedge T_\varepsilon} - \Phi_{k \wedge T_\varepsilon}$ gives the following bound:[3]

$$\begin{aligned} \mathbf{E}[\Phi_{(k+1)\wedge T_\varepsilon} - \Phi_{k\wedge T_\varepsilon} | \mathcal{F}_{k-1}^{G\cdot F}] &\leq -\frac{p_g p_f (1-\nu)(1-\gamma^{-1})}{4} \left(\frac{\mathcal{A}_k \|\nabla f(X_k)\|^2}{L^2} + \theta\Delta_k^2 \right) \cdot 1_{\{T_\varepsilon > k\}} \\ &\leq \frac{-p_g p_f (1-\nu)(1-\gamma^{-1})\mathcal{A}_k}{4\tilde{C}} \cdot (\Phi_k + \varepsilon) \cdot 1_{\{T_\varepsilon > k\}} \end{aligned}$$

[3] We use the notation $a \wedge b = \min\{a, b\}$.

(4.38)

$$\Rightarrow \quad \mathbf{E}[\Phi_{(k+1)\wedge T_\varepsilon} + \varepsilon | \mathcal{F}^{G\cdot F}_{k-1}] \leq \left(1 - \frac{p_g p_f (1-\nu)(1-\gamma^{-1})\mathcal{A}_k}{4\tilde{C}} \cdot 1_{\{T_\varepsilon > k\}}\right)(\Phi_{k\wedge T_\varepsilon} + \varepsilon).$$

Consequently, using Jensen's inequality applied to the function $x \mapsto -\log(x)$ with $x > 0$, we have the following:

$$\begin{aligned} \mathbf{E}[\log(\Phi_{(k+1)\wedge T_\varepsilon} + \varepsilon) &- \log(\Phi_{k\wedge T_\varepsilon} + \varepsilon) | \mathcal{F}^{G\cdot F}_{k-1}] \\ &\leq \log\left(\mathbf{E}[\Phi_{(k+1)\wedge T_\varepsilon} + \varepsilon | \mathcal{F}^{G\cdot F}_{k-1}]\right) - \log(\Phi_{k\wedge T_\varepsilon} + \varepsilon) \\ &= \log\left(\frac{\mathbf{E}[\Phi_{(k+1)\wedge T_\varepsilon} + \varepsilon | \mathcal{F}^{G\cdot F}_{k-1}]}{\Phi_{k\wedge T_\varepsilon} + \varepsilon}\right) \\ &\leq \log\left(1 - \frac{p_g p_f (1-\nu)(1-\gamma^{-1})\mathcal{A}_k}{4\tilde{C}} \cdot 1_{\{T_\varepsilon > k\}}\right), \end{aligned}$$

where the last inequality follows by (4.38). Because $\log(1-x) \leq -x$ for $x < 1$, we deduce our result. □

Using the above theorem allows us to use Theorem 3.3 with $h(\mathcal{A}) = \mathcal{A}$, $\bar{\mathcal{A}} = \frac{1-\theta}{\kappa_g + L/2 + 2\varepsilon_f}$, $\Theta = \frac{p_g p_f (1-\nu)(1-\gamma^{-1})}{4\tilde{C}}$ where $\tilde{C}$ is defined in (4.37), and $p = p_g p_f$. After simplifying some constants, we have the following complexity bound.

THEOREM 4.17. *Let the assumptions of Theorem* 4.16 *hold with constant ν and probabilities p_f and p_g as in Theorem* 4.6. *Suppose we choose the constants as in Remark* 4. *Then the expected number of iterations that Algorithm* 1 *takes until $f(X_k) - f^* < \varepsilon$ is bounded as follows:*

$$\mathbf{E}[T_\varepsilon] \leq \mathcal{O}(1) \cdot \frac{p_g p_f}{2p_g p_f - 1}\left((\kappa_g \alpha_{\max})^2 (\kappa_g + L + \varepsilon_f)\left(\frac{L^2}{2\mu} + \alpha_{\max}\right)\right) \log\left(\frac{\Phi_0 + \varepsilon}{\varepsilon}\right) + 1.$$

Simplifying the bound further gives us

$$\mathbf{E}[T_\varepsilon] \leq \mathcal{O}(1) \cdot \frac{p_g p_f}{2p_g p_f - 1}\left(\frac{L^3(\kappa_g \alpha_{\max})^3}{\mu}\right) \log\left(\frac{\Phi_0}{\varepsilon}\right).$$

4.6. General descent, nonconvex case. In this subsection, we extend the analysis of our line search method to the general setting where steps are taken along some direction d_k, and not the negative stochastic gradient estimate $-g_k$. For example, d_k may be computed by applying a subsampled Newton method [17], or it may be a quasi-Newton direction derived using gradient estimates from the past iteration. We will not assume here any specifics about how d_k is derived, but we will simply assume that d_k and g_k make a sufficiently obtuse angle. Algorithm 1 is then modified as follows:

- a step is reliable when $-\alpha_k g_k^T d_k \geq \delta_k^2$ instead of $\alpha_k \|g_k\|^2 \geq \delta_k^2$;
- the step size $s_k = \alpha_k d_k$ (instead of $-\alpha_k g_k$);
- The sufficient decrease (2.2) is replaced with

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \alpha_k \theta d_k^T g_k; \tag{4.39}$$

- d_k satisfies the following standard conditions.

Assumption 4.18. Given a gradient estimate g_k we assume the following hold for the descent direction d_k.

(i) There exists a constant $\beta > 0$ such that d_k is a descent direction, namely,

$$\frac{d_k^T g_k}{\|d_k\| \, \|g_k\|} \leq -\beta \qquad \text{for all } k.$$

(ii) There exist constants $\kappa_1, \kappa_2 > 0$ such that

$$\kappa_1 \|g_k\| \leq \|d_k\| \leq \kappa_2 \|g_k\| \quad \text{for all } k.$$

We note that to satisfy the above assumption one can always modify the step direction d_k by adding a appropriate multiple of g_k to it. For example, such self-correcting technique has been successfully used together with the stochastic L-BFGS method in [8].

We now provide simple variants of the lemmas derived in section 4.1.

Lemma 4.19 (bound on gradient change, variant of Lemma 4.4). *Suppose the kth step is successful and the descent direction d_k satisfies Assumption* 4.18. *Then*

$$\|\nabla f(x_{k+1})\|^2 \leq 2(L^2 \alpha_k^2 \kappa_2^2 \|g_k\|^2 + \|\nabla f(x_k)\|^2).$$

In particular, the following inequality holds:

$$\frac{1}{L^2}\left(\alpha_{k+1} \|\nabla f(x_{k+1})\|^2 - \alpha_k \|\nabla f(x_k)\|^2\right) \leq 2\gamma\alpha_k \left(\alpha_{\max}^2 \kappa_2^2 \|g_k\|^2 + \frac{1}{L^2} \|\nabla f(x_k)\|^2\right).$$

Proof. An immediate consequence of L-smoothness of f is $\|\nabla f(x_{k+1})\| \leq L\alpha_k \|d_k\| + \|\nabla f(x_k)\|$. The result follows from Assumption 4.18(ii) then squaring both sides and applying the bound, $(a+b)^2 \leq 2(a^2+b^2)$. To obtain the second inequality, we note that in the case $x_k + s_k$ is successful, $\alpha_{k+1} = \gamma\alpha_k$. □

The analysis for the steepest descent relies on successful iterations occurring whenever the step size is sufficiently small. We provide a similar result for the general descent case.

Lemma 4.20 (accurate gradients and function estimates and small step size ⇒ successful step, variant of Lemma 4.2). *Suppose g_k is κ_g-sufficiently accurate, the descent direction d_k satisfies Assumption* 4.18, *and $\{f_k^0, f_k^s\}$ are ε_f-accurate estimates. If*

$$\alpha_k \leq \frac{\beta(1-\theta)}{\kappa_g + \frac{L\kappa_2}{2} + \frac{2\varepsilon_f}{\kappa_1}},$$

then the kth step is successful. In particular, this means $f_k^s \leq f_k^0 + \theta\alpha_k g_k^T d_k$.

Proof. The L-smoothness of f and the κ_g-sufficiently accurate gradient immediately yield

$$\begin{aligned} f(x_k + s_k) &\leq f(x_k) + \alpha_k(\nabla f(x_k) - g_k)^T d_k + \alpha_k g_k^T d_k + \frac{L\alpha_k^2}{2} \|d_k\|^2 \\ &\leq f(x_k) + \kappa_g \alpha_k^2 \|d_k\| \, \|g_k\| + \alpha_k g_k^T d_k + \frac{L\alpha_k^2}{2} \|d_k\|^2. \end{aligned}$$

Since the estimates are ε_f-accurate, we obtain

$$\begin{aligned} f_k^s - \varepsilon_f \alpha_k^2 \|g_k\|^2 &\leq f(x_k + s_k) - f_k^s + f_k^s \\ &\leq f(x_k) - f_k^0 + f_k^0 + \kappa_g \alpha_k^2 \|d_k\| \, \|g_k\| + \alpha_k g_k^T d_k + \frac{L\alpha_k^2}{2} \|d_k\|^2 \\ &\leq f_k^0 + \varepsilon_f \alpha_k^2 \|g_k\|^2 + \kappa_g \alpha_k^2 \|d_k\| \, \|g_k\| + \alpha_k g_k^T d_k + \frac{L\alpha_k^2}{2} \|d_k\|^2. \end{aligned}$$

The above inequality with Assumption 4.18 implies

$$\begin{aligned} f_k^s - f_k^0 &\leq \alpha_k^2 \left(\frac{2\varepsilon_f}{\kappa_1} + \kappa_g + \frac{L\kappa_2}{2} \right) \|g_k\| \, \|d_k\| + \alpha_k g_k^T d_k \\ &\leq \frac{-\alpha_k^2}{\beta} \left(\frac{2\varepsilon_f}{\kappa_1} + \kappa_g + \frac{L\kappa_2}{2} \right) g_k^T d_k + \alpha_k g_k^T d_k. \end{aligned}$$

The result follows by noting $f_k^s \leq f_k^0 + \alpha_k g_k^T d_k (1 - \frac{\alpha_k}{\beta}(\kappa_g + \frac{L\kappa_2}{2} + \frac{2\varepsilon_f}{\kappa_1}))$. □

As in the steepest descent case, we use the same function Φ_k as defined in (4.6). Using the sufficient decrease condition (4.39) and Assumption 4.18 on d_k, a successful step yields a decrease of

$$f(x_k + \alpha_k d_k) \leq -\theta \alpha_k \kappa_1 \beta \|g_k\|^2 .$$

Hence, we can derive, as in the steepest descent scenario, an expected decrease in Φ_k.

THEOREM 4.21. *Let Assumptions* 2.1, 2.4, *and* 4.18 *hold. Suppose the random process generated by Algorithm* 1 *is* $\{G_k, D_k, X_k, \mathcal{A}_k, \Delta_k, S_k, F_k^0, F_k^s\}$. *Then there exist probabilities* $p_g, p_f > 1/2$ *and a constant* $\nu \in (0,1)$ *such that the expected decrease in* Φ_k *is*

$$\mathbf{E}[\Phi_{k+1} - \Phi_k | \mathcal{F}_{k-1}^{G\cdot F}] \leq - \frac{p_g p_f (1-\nu)(1-\gamma^{-1})}{4} \left(\frac{\mathcal{A}_k}{L^2} \|\nabla f(X_k)\|^2 + \theta \Delta_k^2 \right). \tag{4.40}$$

In particular, the constant ν *and probabilities* $p_g, p_f > 1/2$ *satisfy*

$$\frac{\nu}{1-\nu} \geq \max \left\{ \frac{32\gamma \alpha_{\max}^2 \kappa_2^2}{\theta \kappa_1 \beta}, 16(\gamma - 1), \frac{16\gamma(\kappa_g \alpha_{\max} + 1)^2}{\theta \kappa_1 \beta} \right\}, \tag{4.41}$$

$$p_g \geq \frac{2\gamma}{1/2(1-\gamma^{-1}) + 2\gamma} \tag{4.42}$$

$$\textit{and} \qquad \frac{p_g p_f}{\sqrt{1-p_f}} \geq \max \left\{ \frac{8L^2 \nu \kappa_f + 16\gamma(1-\nu)}{(1-\nu)(1-\gamma^{-1})}, \frac{8\nu}{(1-\nu)(1-\gamma^{-1})} \right\}. \tag{4.43}$$

Proof. Using Assumption 4.18 on the descent direction d_k when a step is successful, we see

$$1_{\text{Succ}}(f(X_k + \mathcal{A}_k D_k) - f(X_k)) \leq -1_{\text{Succ}} \theta \mathcal{A}_k \kappa_1 \beta \|G_k\|^2 .$$

Hence, we may replace θ in the proof of Theorem 4.6 with $\theta\kappa_1\beta$. The only other change to the proof and the resulting constants lies in the replacement of Lemma 4.4 by Lemma 4.19. This implies a change in the choice of ν in (4.15). In particular, we choose ν to now satisfy

$$\begin{aligned} &-\frac{\nu\theta\kappa_1\beta\mathcal{A}_k}{8} \|G_k\|^2 + (1-\nu) 2\gamma \mathcal{A}_k \alpha_{\max}^2 \kappa_2^2 \|G_k\|^2 \leq -\frac{\nu\theta\kappa_1\beta\mathcal{A}_k}{16} \|G_k\|^2 , \\ &-\frac{\nu\theta\kappa_1\beta\mathcal{A}_k \|\nabla f(X_k)\|^2}{4L^2(\kappa_g\alpha_{\max}+1)^2} + \frac{2(1-\nu)\gamma\mathcal{A}_k \|\nabla f(X_k)\|^2}{L^2} \leq -\frac{\nu\theta\kappa_1\beta\mathcal{A}_k \|\nabla f(X_k)\|^2}{8L^2(\kappa_g\alpha_{\max}+1)^2}, \\ &\text{and} \qquad -\frac{\nu\theta}{8}\Delta_k^2 + (1-\nu)(\gamma-1)\theta\Delta_k^2 \leq -\frac{\nu\theta}{16}\Delta_k^2. \end{aligned} \tag{4.44}$$

□

The dynamics of $\mathcal{A}_k$ mimic those in the GD case, and thus Lemma 4.8 holds by replacing $\bar{\mathcal{A}}$ with

$$\bar{\mathcal{A}} = \frac{\beta(1-\theta)}{\kappa_g + \frac{L\kappa_2}{2} + \frac{2\varepsilon_f}{\kappa_1}}.$$

The proof of Lemma 4.8 relied on Lemma 4.2 which we replace in the general descent case with Lemma 4.20. We derive a complexity bound using Theorem 3.3 for the general descent setting.

Theorem 4.22. *Under the assumptions in Theorem* 4.21*, and constants chosen in Remark* 4*, suppose the probabilities* $p_g, p_f > 1/2$ *satisfy*

$$p_g \geq \frac{16}{17} \quad \textit{and}$$

$$\frac{p_g p_f}{\sqrt{1-p_f}} \geq \max\left\{\frac{1024\kappa_f L^2(\max\{\kappa_g, 2\kappa_2\}\alpha_{\max}+1)^2}{\kappa_1\beta} + 64,\ \frac{1024(\max\{\kappa_g, 2\kappa_2\}\alpha_{\max}+1)^2}{\kappa_1\beta}\right\}$$

with $\frac{\nu}{1-\nu} = \frac{64(\max\{\kappa_g, 2\kappa_2\}\alpha_{\max}+1)^2}{\kappa_1\beta}$*. Then the expected number of iterations that Algorithm* 1 *takes until* $\|\nabla f(X_k)\|^2 \leq \varepsilon$ *occurs is bounded as follows:*

$$\mathbf{E}[T_\varepsilon] \leq \mathcal{O}(1) \cdot \frac{p_g p_f}{2p_g p_f - 1} \cdot \frac{L^3\kappa_g^3\kappa_2^3\Phi_0}{\kappa_1^2\beta^2} \cdot \frac{1}{\varepsilon^2} + 1,$$

where $\Phi_0 = \nu(f(X_0) - f_{\min}) + (1-\nu)(1/L^2\mathcal{A}_0\|\nabla f(X_0)\|^2 + 1/2\Delta_0^2)$.

5. Conclusions.

We have used a general framework based on the analysis of stochastic processes proposed in [3] with the purpose of analyzing expected complexity of stochastic optimization methods. In [3] the framework is used to analyze a stochastic trust region method, while in this paper we were able to use the same framework to develop and analyze a stochastic backtracking line search method. Our method is the first implementable stochastic line search method that has theoretical convergence rate guarantees. In particular, the accuracy of gradient and function estimates is chosen dynamically, and the requirements of this accuracy are all stated in terms of knowable quantities. We establish complexity results for convex, strongly convex, and general nonconvex, smooth stochastic functions.

REFERENCES

[1] L. Armijo, *Minimization of functions having Lipschitz continuous first partial derivatives*, Pacific J. Math., 16 (1966), pp. 1–3.

[2] A. S. Bandeira, K. Scheinberg, and L. N. Vicente, *Convergence of trust-region methods based on probabilistic models*, SIAM J. Optim., 24 (2014), pp. 1238–1264.

[3] J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg, *Convergence Rate Analysis of a Stochastic Trust Region Method via Submartingales*, arXiv:1609.07428 [math.OC], 2018, https://arxiv.org/abs/1609.07428.

[4] R. Bollapragada, R. Byrd, and J. Nocedal, *Adaptive sampling strategies for stochastic optimization*, SIAM J. Optim., 28 (2018), pp. 3312–3343.

[5] R. 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.

[6] C. Cartis and K. Scheinberg, *Global convergence rate analysis of unconstrained optimization methods based on probabilistic models*, Math. Program., 169 (2018), pp. 337–375.

[7] R. Chen, M. Menickelly, and K. Scheinberg, *Stochastic optimization using a trust-region method and random models*, Math. Program., 169 (2018), pp. 447–487.

[8] F. Curtis, *A self-correcting variable-metric algorithm for stochastic optimization*, in Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 632–641.

[9] J. C. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learn. Res., 12 (2011), pp. 2121–2159.

[10] M. P. Friedlander and M. Schmidt, *Hybrid deterministic-stochastic methods for data fitting*, SIAM J. Sci. Comput., 34 (2012), pp. A1380–A1405.

[11] A. P. George and W. B. Powell, *Adaptive stepsizes for recursive estimation with applications in approximate dynamic programming*, Mach. Learn., 65 (2006), pp. 167–198.

[12] F. S. Hashemi, S. Ghosh, and R. Pasupathy, *On adaptive sampling rules for stochastic recursions*, in Proceedings of the Winter Simulation Conference, IEEE, 2014, pp. 3959–3970.

[13] P. Hennig, *Fast probabilistic optimization from noisy gradients*, in Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 62–70.

[14] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in Proceedings of the International Conference on Learning Representations (ICLR), 2015.

[15] M. Mahsereci and P. Hennig, *Probabilistic line searches for stochastic optimization*, J. Mach. Learn. Res., 18 (2017), pp. 1–59.

[16] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, NY, 2006.

[17] M. W. M. Peng Xu, Farbod Roosta-Khorasani, *Newton-Type Methods for Non-Convex Optimization under Inexact Hessian Information*, arXiv:1708.07164 [math.OC], 2017, https://arxiv.org/abs/1708.07164.

[18] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist., 22 (1951), pp. 400–407.

[19] N. Tripuraneni, M. Stern, C. Jin, J. Regier, and M. I. Jordan, *Stochastic Cubic Regularization for Fast Nonconvex Optimization*, arXiv:1711.02838 [cs.LG], 2017, https://arxiv.org/abs/1711.02838.

[20] J. A. Tropp, *An introduction to matrix concentration inequalities*, Found. Trends Mach. Learn., 8 (2015), pp. 1–230.