A Stochastic Trust Region Algorithm

Frank E. Curtis, Katya Scheinberg, and Rui Shi

Department of Industrial and Systems Engineering, Lehigh University

COR@L Technical Report 17T-17

A Stochastic Trust Region Algorithm

Frank E. Curtis*[1], Katya Scheinberg†[1], and Rui Shi‡[1]

[1]Department of Industrial and Systems Engineering, Lehigh University

December 29, 2017

Abstract

An algorithm is proposed for solving stochastic and finite sum minimization problems. Based on a trust region methodology, the algorithm employs normalized steps, at least as long as the norms of the stochastic gradient estimates are within a user-defined interval. The complete algorithm—which dynamically chooses whether or not to employ normalized steps—is proved to have convergence guarantees that are similar to those possessed by a traditional stochastic gradient approach under various sets of conditions related to the accuracy of the stochastic gradient estimates and choice of stepsize sequence. The results of numerical experiments are presented when the method is employed to minimize convex and nonconvex machine learning test problems, illustrating that the method can outperform a traditional stochastic gradient approach.

1 Introduction

The stochastic gradient (SG) method is the signature strategy for solving stochastic and finite-sum minimization problems. In this iterative approach, each step to update the solution estimate is obtained by taking a negative multiple of an unbiased gradient estimate. With careful choices for the stepsize sequence, the SG method possesses convergence guarantees and has been employed to great success for solving various types of problems, such as those arising in machine learning.

One disadvantage of the SG method is that stochastic gradients, like the gradients that they approximate, possess *no natural scaling*. By this, we mean that in order to guarantee convergence, the algorithm needs to choose stepsizes in a problem-dependent manner; e.g., common theoretical guarantees require that the stepsize is proportional to $1/L$, where L is a Lipschitz constant for the gradient of the objective function. This is in contrast to Newton's method for minimization, for which one can obtain (local) convergence guarantees with a stepsize of 1. Admittedly, Newton's method is not generally guaranteed to converge from remote starting points with unit stepsizes, but these observations do highlight a shortcoming of first-order methods, namely, that for convergence guarantees the stepsizes need always be chosen in a problem-dependent manner.

The purpose of this paper is to propose a new algorithm for stochastic and finite-sum minimization. Our proposed approach can be viewed as a modification of the SG method. The approach does not completely overcome the issue of requiring problem-dependent stepsizes, but we contend that our approach does, for practical purposes, reduce somewhat this dependence. This is achieved by employing, under certain conditions, *normalized* steps. We motivate our proposed approach by illustrating how it can be derived from a trust region methodology.

The use of normalized steps has previously been proposed in the context of (stochastic) gradient methods for solving minimization problems. For example, in a method that is similar to ours, [4] proposes an approach

*E-mail: frank.e.curtis@lehigh.edu

†E-mail: katyas@lehigh.edu

‡E-mail: rus415@lehigh.edu

that employs normalized steps in every iteration. They show that, if the objective function is *M-bounded* and *strictly-locally-quasi-convex*, the stochastic gradients are sufficiently accurate with respect to the true gradients (specifically, when mini-batch sizes are $\Omega(\epsilon^{-2})$), and a sufficiently large number of iterations are run (specifically, $\Omega(\epsilon^{-2})$), then their method will, with high probability, yield a solution estimate that is ϵ-optimal. By contrast, our approach, by employing a modified update that does not always involve the use of a normalized step, enjoys convergence guarantees in more general settings. Indeed, we argue in this paper that employing normalized steps in all iterations cannot lead to general convergence guarantees, which perhaps explains the various assumptions required for convergence by [4].

It is also worthwhile to mention the broader literature, such as the work in [6] and [3]. The approaches proposed in these papers, which are based on the use of randomized models of the objective function constructed during each iteration, are quite distinct from our proposed method. For example, these approaches follow a traditional trust region strategy of accepting or rejecting each step based on the magnitude of an (approximate) *actual-to-predicted reduction ratio*. Our method, on the other hand, is closer to the SG method in that it accepts the computed step in every iteration. Another distinction is that these other approaches rely on the use of so-called *fully linear* models of the objective function to obtain their convergence guarantees. Our convergence guarantees are obtained under new and different sets of conditions.

The paper is organized as follows. Our algorithm and motivation for our specific iterate updating scheme are the subject of §2. In §3, we prove convergence guarantees for the algorithm under various types of assumptions on the stochastic gradient estimates and stepsize choices. The results of numerical experiments on two test problems—one convex and one nonconvex—are given in §4. Concluding remarks are given in §5.

2 Algorithm

Our problem of interest is a stochastic optimization problem in which the goal is to minimize over a vector of decision variables, indicated by $x \in \mathbb{R}^n$, a function $f : \mathbb{R}^n \to \mathbb{R}$ defined by the expectation of another function $F : \mathbb{R}^n \times \Xi \to \mathbb{R}$ that depends on a random variable ξ, i.e.,

$$\min_{x\in\mathbb{R}^n} \; f(x), \quad \text{where} \quad f(x) = \mathbb{E}_\xi[F(x,\xi)], \tag{2.1}$$

where $\mathbb{E}_\xi[\cdot]$ denotes expectation with respect to the distribution of ξ. Our algorithm is also applicable for finite-sum minimization where the objective takes the form

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x). \tag{2.2}$$

Such objectives often arise in sample average approximations of (2.1); e.g., see [8].

2.1 Algorithm Description

Our algorithm is stated below as `TRish`, a trust-region-*ish* algorithm for stochastic optimization. Each iteration involves taking a step along the negative of a stochastic gradient direction. In the context of problem (2.1), this stochastic gradient can be viewed as $g_k = \nabla_x F(x_k, \xi_k)$, where x_k is the current iterate and ξ_k is a realization of the random variable ξ. In the context of problem (2.2), it can be viewed as $g_k = \nabla_x f_{i_k}(x_k)$ where i_k has been chosen randomly as an index in $\{1,\dots,N\}$. In addition, in either case, g_k could represent an average of such quantities, i.e., over a set of independently generated realizations $\{\xi_{k,j}\}_j$ or over independently generated indices $\{i_{k,j}\}_j$ (leading to a so-called *mini-batch* approach). In the algorithm, we simply write $g_k \approx \nabla f(x_k)$ to cover all of these situations, since in any case g_k represents a stochastic gradient estimate for f at x_k.

The scaling of the stochastic gradient employed in TRish can be motivated in the following manner. Given a stochastic gradient g_k and a stepsize α_k, consider the trust region subproblem

$$\min_{d\in\mathbb{R}^n} \; f(x_k) + g_k^T d \quad \text{s.t.} \quad \|d\|_2 \le \alpha_k. \tag{2.3}$$

Algorithm TRish (Trust-region-ish algorithm)

1: choose an initial iterate x_1, positive stepsizes $\{\alpha_k\}$, and positive constants $\gamma_1 > \gamma_2 > 0$
2: **for** $k \in \mathbb{N} := \{1, 2, \dots\}$ **do**
3: generate a stochastic gradient $g_k \approx \nabla f(x_k)$
4: set

$$x_{k+1} \leftarrow x_k - \begin{cases} \gamma_1 \alpha_k g_k & \text{if } \|g_k\|_2 \in [0, \frac{1}{\gamma_1}) \\ \alpha_k g_k / \|g_k\|_2 & \text{if } \|g_k\|_2 \in [\frac{1}{\gamma_1}, \frac{1}{\gamma_2}] \\ \gamma_2 \alpha_k g_k & \text{if } \|g_k\|_2 \in (\frac{1}{\gamma_2}, \infty) \end{cases}$$

5: **end for**

The solution of this subproblem, namely, $d_k = -\alpha_k g_k / \|g_k\|_2$, represents a step toward minimizing the first-order model $f(x_k) + g_k^T d$ of the objective function f at x_k subject to d having norm less than or equal to α_k. This is the prototypical strategy in a trust region methodology. When the norm of g_k falls within the interval $[\frac{1}{\gamma_1}, \frac{1}{\gamma_2}]$, Algorithm TRish takes the step d_k. However, if this were to be done no matter the norm of g_k, then the resulting algorithm might fail to make progress in expectation. This is illustrated in the following example.

Example 2.1. *Suppose that, at a point $x_k \in \mathbb{R}$, one has $\nabla f(x_k) = 1$ and obtains*

$$g_k = \begin{cases} 6 & \text{with probability } \frac{1}{3} \\ -\frac{3}{2} & \text{with probability } \frac{2}{3}. \end{cases}$$

Then, $\mathbb{E}_k[g_k] = 1 = \nabla f(x_k)$, where $\mathbb{E}_k$ denotes expectation given that an algorithm has reached x_k as the kth iterate. However, this means that the normalized stochastic gradient satisfies

$$\frac{g_k}{\|g_k\|_2} = \begin{cases} 1 & \text{with probability } \frac{1}{3} \\ -1 & \text{with probability } \frac{2}{3}, \end{cases}$$

from which it follows that $d_k = -\alpha_k g_k / \|g_k\|_2$ is twice more likely to be a direction of ascent for f at x_k than it is to be a direction of descent for f at x_k.

One can argue from this example that, without potentially restrictive assumptions on the objective function f and/or the manner in which the stochastic gradient is computed, one cannot expect to be able to prove convergence guarantees for an algorithm that solely computes steps based on solving the trust region subproblem (2.3). In particular, the existence of any point (let alone more than one) at which the expectation is to follow an ascent direction foils the typical convergence theory for a stochastic gradient approach; e.g., see [1].

In TRish, we overcome the issue highlighted in Example 2.1 by only choosing the trust region step when the norm of the gradient is within a user-defined interval; otherwise, we compute a stochastic gradient step with a stepsize that is a multiple of α_k. It is for this reason that we refer to the algorithm as a trust-region-*ish* approach. Overall, as a function of the norm of the stochastic gradient, the norm of the step taken by the algorithm is illustrated in Figure 1. Note that care has been taken to make sure that the norm of the step is a continuous function of the norm of the stochastic gradient estimate. The plot in Figure 1 illustrates the relationship for moderate values of (γ_1, γ_2), but notice that for more extreme values (i.e., $\gamma_1 \gg 0$ and $\gamma_2 \approx 0$) the function would essentially be flat (except for stochastic gradients that are very small in norm), meaning that the stepsize would typically be scaled so that the step norm is approximately α_k for all $k \in \mathbb{N}$.

3 Convergence Analysis

Our goal in this section is to prove convergence guarantees for TRish that are similar to fundamental guarantees for a straightforward stochastic gradient method; see [1].

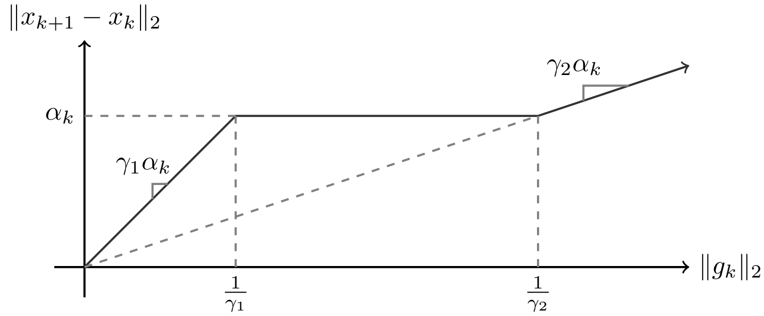

Figure 1: Relationship between $\|g_k\|_2$ and $\|x_{k+1} - x_k\|_2$ in Algorithm TRish.

Throughout our analysis, we make the following assumption about the objective function.

Assumption 3.1. *The objective $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and bounded below by $f_* = \inf_{x\in\mathbb{R}^n} f(x) \in \mathbb{R}$. In addition, at any $x \in \mathbb{R}^n$, the objective is bounded above by a first-order Taylor series approximation of f at x plus a quadratic term with constant $L \in (0, \infty)$, i.e.,*

$$f(x) \leq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) + \tfrac{1}{2}L\|x - \bar{x}\|_2^2 \ \text{ for all } \ (x, \bar{x}) \in \mathbb{R}^n \times \mathbb{R}^n. \tag{3.1}$$

We also make the following assumption about the stochastic gradients computed in TRish.

Assumption 3.2. *For all $k \in \mathbb{N}$, the stochastic gradient g_k is an unbiased estimator of $\nabla f(x_k)$ in the sense that $\mathbb{E}_k[g_k] = \nabla f(x_k)$. In addition, there exists a pair $(M_1, M_2) \in (0, \infty) \times (0, \infty)$ (independent of k) such that, for all $k \in \mathbb{N}$, the squared norm of g_k satisfies*

$$\mathbb{E}_k[\|g_k\|_2^2] \leq M_1 + M_2\|\nabla f(x_k)\|_2^2. \tag{3.2}$$

Under these assumptions, we prove the following lemma providing fundamental inequalities satisfied by TRish. For ease of reference in this result and throughout the remainder of our analysis, we define the following cases based on those indicated in Step 4 of TRish:

$$\text{“case 1”}: \quad \|g_k\|_2 \in [0, \tfrac{1}{\gamma_1}); \quad \text{“case 2”}: \quad \|g_k\|_2 \in [\tfrac{1}{\gamma_1}, \tfrac{1}{\gamma_2}]; \quad \text{“case 3”}: \quad \|g_k\|_2 \in (\tfrac{1}{\gamma_2}, \infty).$$

The following lemma reveals an upper bound for the expected decrease in f for all $k \in \mathbb{N}$.

Lemma 3.1. *Under Assumptions 3.1 and 3.2, the iterates of TRish satisfy, for all $k \in \mathbb{N}$,*

$$\begin{aligned} &\mathbb{E}_k[f(x_{k+1})] - f(x_k) \\ &\leq \begin{cases} -\gamma_1\alpha_k(1 - \frac{1}{2}\gamma_1 L M_2 \alpha_k)\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\gamma_1^2 L M_1 \alpha_k^2 & \text{if } \|g_k\|_2 \in [0, \frac{1}{\gamma_1}) \\ -\gamma_1\alpha_k\|\nabla f(x_k)\|_2^2 + (\gamma_1 - \gamma_2)\alpha_k \mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] + \frac{1}{2}L\alpha_k^2 & \text{if } \|g_k\|_2 \in [\frac{1}{\gamma_1}, \frac{1}{\gamma_2}] \\ -\gamma_2\alpha_k(1 - \frac{1}{2}\gamma_2 L M_2 \alpha_k)\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\gamma_2^2 L M_1 \alpha_k^2 & \text{if } \|g_k\|_2 \in (\frac{1}{\gamma_2}, \infty), \end{cases} \end{aligned} \tag{3.3}$$

where E_k is the event that $\nabla f(x_k)^T g_k \geq 0$ and $\mathbb{P}_k[E_k]$ is the probability of this event.

Proof. Proof. In case 1, one finds $\|g_k\|_2 \in [0, \frac{1}{\gamma_1})$ and $x_{k+1} \leftarrow x_k - \gamma_1\alpha_k g_k$. Therefore, from (3.1),

$$f(x_{k+1}) = f(x_k - \gamma_1\alpha_k g_k) \leq f(x_k) - \gamma_1\alpha_k\nabla f(x_k)^T g_k + \tfrac{1}{2}\gamma_1^2 L\alpha_k^2\|g_k\|_2^2.$$

Taking expectations on both sides, one finds with (3.2) that

$$\begin{aligned} \mathbb{E}_k[f(x_{k+1})] &\leq f(x_k) - \gamma_1\alpha_k\nabla f(x_k)^T\mathbb{E}_k[g_k] + \tfrac{1}{2}\gamma_1^2 L\alpha_k^2\mathbb{E}_k[\|g_k\|_2^2] \\ &\leq f(x_k) - \gamma_1\alpha_k\|\nabla f(x_k)\|_2^2 + \tfrac{1}{2}\gamma_1^2 L\alpha_k^2(M_1 + M_2\|\nabla f(x_k)\|_2^2) \\ &= f(x_k) - \gamma_1\alpha_k(1 - \tfrac{1}{2}\gamma_1 L M_2\alpha_k)\|\nabla f(x_k)\|_2^2 + \tfrac{1}{2}\gamma_1^2 L M_1\alpha_k^2, \end{aligned}$$

as desired. The proof for case 3, when one finds $\|g_k\|_2 \in (\frac{1}{\gamma_2}, \infty)$ and $x_{k+1} \leftarrow x_k - \gamma_2 \alpha_k g_k$, follows similarly with γ_2 in place of γ_1. Finally, consider case 2 when $\|g_k\|_2 \in [\frac{1}{\gamma_1}, \frac{1}{\gamma_2}]$ and the iterate update has the form $x_{k+1} \leftarrow x_k - \alpha_k g_k / \|g_k\|_2$. From (3.1), it follows that

$$f(x_{k+1}) = f\left(x_k - \frac{\alpha_k}{\|g_k\|_2} g_k\right) \leq f(x_k) - \frac{\alpha_k \nabla f(x_k)^T g_k}{\|g_k\|_2} + \tfrac{1}{2} L \alpha_k^2. \tag{3.4}$$

In the event that $\nabla f(x_k)^T g_k \geq 0$, it follows from (3.4) and the fact that $\|g_k\|_2 \leq \frac{1}{\gamma_2}$ that

$$f(x_{k+1}) \leq f(x_k) - \gamma_2 \alpha_k \nabla f(x_k)^T g_k + \tfrac{1}{2} L \alpha_k^2,$$

which, after taking expectations on both sides, implies with (3.2) that

$$\mathbb{E}_k[f(x_{k+1}) | E_k] \leq f(x_k) - \gamma_2 \alpha_k \mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] + \tfrac{1}{2} L \alpha_k^2. \tag{3.5}$$

On the other hand, if $\nabla f(x_k)^T g_k < 0$, then, from (3.4) and $\|g_k\|_2 \geq \frac{1}{\gamma_1}$, it follows similarly that

$$\mathbb{E}_k[f(x_{k+1}) | \overline{E}_k] \leq f(x_k) - \gamma_1 \alpha_k \mathbb{E}_k[\nabla f(x_k)^T g_k | \overline{E}_k] + \tfrac{1}{2} L \alpha_k^2, \tag{3.6}$$

where $\overline{E}_k$ denotes the complement of E_k. From the Law of Total Expectation, it follows that

$$\begin{aligned} \|\nabla f(x_k)\|_2^2 &= \mathbb{E}_k[\nabla f(x_k)^T g_k] \\ &= \mathbb{P}_k[E_k] \mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] + (1 - \mathbb{P}_k[E_k]) \mathbb{E}_k[\nabla f(x_k)^T g_k | \overline{E}_k]. \end{aligned} \tag{3.7}$$

Hence, combining (3.5), (3.6), and (3.7), it follows in this case that

$$\begin{aligned} &\mathbb{E}_k[f(x_{k+1})] \\ &= \mathbb{P}_k[E_k] \mathbb{E}_k[f(x_{k+1}) | E_k] + (1 - \mathbb{P}_k[E_k]) \mathbb{E}_k[f(x_{k+1}) | \overline{E}_k] \\ &\leq f(x_k) - \mathbb{P}_k[E_k] \gamma_2 \alpha_k \mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] - (1 - \mathbb{P}_k[E_k]) \gamma_1 \alpha_k \mathbb{E}_k[\nabla f(x_k)^T g_k | \overline{E}_{x_k}] + \tfrac{1}{2} L \alpha_k^2 \\ &= f(x_k) - \mathbb{P}_k[E_k] \gamma_2 \alpha_k \mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] - \gamma_1 \alpha_k (\|\nabla f(x_k)\|_2^2 - \mathbb{P}_k[E_k] \mathbb{E}_k[\nabla f(x_k)^T g_k | E_k]) + \tfrac{1}{2} L \alpha_k^2 \\ &= f(x_k) - \gamma_1 \alpha_k \|\nabla f(x_k)\|_2^2 + (\gamma_1 - \gamma_2) \alpha_k \mathbb{P}_k[E_k] \mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] + \tfrac{1}{2} L \alpha_k^2, \end{aligned}$$

as desired. □ □

For some of our convergence guarantees, we also make the following assumption.

Assumption 3.3. *At any $x \in \mathbb{R}^n$, the Polyak-Lojasiewicz condition holds with $c \in (0, \infty)$, i.e.,*

$$2c(f(x) - f_*) \leq \|\nabla f(x)\|_2^2 \quad \textit{for all} \quad x \in \mathbb{R}^n. \tag{3.8}$$

Assumptions 3.1 and 3.3 do not ensure that a stationary point for f exists, though, when combined, they do guarantee that any stationary point for f is a global minimizer of f. Assumption 3.3 holds when f is c-strongly convex, but it is also satisfied for other functions that are not convex. We direct the interested reader to [5] for a discussion on the relationship between the Polyak-Łojasiewicz condition and the related *error bounds*, *essential strong convexity*, *weak strong convexity*, *restricted secant inequality*, and *quadratic growth* conditions. In short, when f has a Lipschitz continuous gradient, the Polyak-Lojasiewicz is the weakest of these except for the quadratic growth condition, though these two are equivalent when f is convex.

We now proceed to prove convergence guarantees for TRish in various cases depending on whether or not the Polyak-Lojasiewicz condition (hereafter referred to as the P-L condition) holds and based on different sets of properties of the sequence of stepsizes and stochastic gradient estimates. Our analysis spans various types of convex and nonconvex objective functions.

3.1 P-L Condition, Bounded Variance with a Fixed Stepsize

Let us first prove a convergence result for TRish when the P-L condition holds, a sufficiently small fixed stepsize is employed (see Theorem 3.1 later on), and the inner products between the gradients and stochastic gradients satisfy the following assumption.

Assumption 3.4. *There exist constants $h_1 \in (0,\infty)$ and $h_2 \in (1,\infty)$ such that, for all $k \in \mathbb{N}$,*

$$\mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] \leq h_1 + h_2 \|\nabla f(x_k)\|_2^2. \tag{3.9}$$

In order to motivate this being a reasonable assumption in certain settings, consider the following two examples wherein we show that if the stochastic gradient follows a normal distribution centered at the true gradient with bounded variance, then (3.9) holds for certain h_1 and h_2.

Example 3.1. *Suppose $f : \mathbb{R} \to \mathbb{R}$ and x_k are given such that $\nabla f(x_k) = \mu_k \in \mathbb{R}$, where without loss of generality one can assume that $\mu_k \geq 0$. In addition, suppose that g_k follows a normal distribution with mean μ_k and variance σ_k^2. Then,*

$$\begin{aligned}\mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] &= \mu_k \int_0^\infty g \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-(g-\mu_k)^2}{2\sigma_k^2}} dg \\ &= \mu_k \int_0^{\mu_k} g \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-(g-\mu_k)^2}{2\sigma_k^2}} dg + \mu_k \int_{\mu_k}^\infty g \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-(g-\mu_k)^2}{2\sigma_k^2}} dg.\end{aligned}$$

Let us separately investigate these two terms on the right-hand side. First, one finds that

$$\mu_k \int_0^{\mu_k} g \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-(g-\mu_k)^2}{2\sigma_k^2}} dg \leq \mu_k^2 \int_0^{\mu_k} \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-(g-\mu_k)^2}{2\sigma_k^2}} dg \leq \mu_k^2 \int_{-\infty}^{\mu_k} \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-(g-\mu_k)^2}{2\sigma_k^2}} dg = \tfrac{1}{2}\mu_k^2.$$

Second, one finds that

$$\begin{aligned}\mu_k \int_{\mu_k}^\infty g \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-(g-\mu_k)^2}{2\sigma_k^2}} dg &= \mu_k \int_0^\infty (t+\mu_k) \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-t^2}{2\sigma_k^2}} dt \\ &\leq \mu_k \int_0^\infty t \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-t^2}{2\sigma_k^2}} dt + \mu_k^2 \int_0^\infty \tfrac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-t^2}{2\sigma_k^2}} dt = \mu_k \tfrac{\sigma_k}{\sqrt{2\pi}} + \tfrac{1}{2}\mu_k^2.\end{aligned}$$

Thus, combining the bounds above, one finds that

$$\mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] \leq \mu_k \frac{\sigma_k}{\sqrt{2\pi}} + \mu_k^2 \leq \left(\frac{\mu_k^2 + 1}{2}\right) \frac{\sigma_k}{\sqrt{2\pi}} + \mu_k^2 = \frac{\sigma_k}{2\sqrt{2\pi}} + \left(1 + \frac{\sigma_k}{2\sqrt{2\pi}}\right)\mu_k^2.$$

Overall, if $\sigma_k \leq \sigma$ for some positive $\sigma \in \mathbb{R}$ for all $k \in \mathbb{N}$, then Assumption 3.4 holds with

$$h_1 = \frac{\sigma}{2\sqrt{2\pi}} \quad \text{and} \quad h_2 = 1 + \frac{\sigma}{2\sqrt{2\pi}}. \tag{3.10}$$

Example 3.2. *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ and x_k are given such that $\nabla f(x_k) \in \mu_k \in \mathbb{R}^n$. In addition, suppose that g_k follows a normal distribution with mean μ_k and covariance matrix Σ_k. Then, by Theorem 3.3.3 in [9], the inner product $\nabla f(x_k)^T g_k$ follows a normal distribution with mean $\|\mu_k\|_2^2$ and variance $\mu_k^T \Sigma_k \mu_k$. Hence, following the analysis in Example 3.1, if $\sqrt{\mu_k^T \Sigma_k \mu_k} \leq \sigma$ for some positive $\sigma \in \mathbb{R}$ for all $k \in \mathbb{N}$, then Assumption 3.4 holds with h_1 and h_2 from (3.10).*

We now prove our first theorem on the behavior of TRish.

Theorem 3.1. *Under Assumptions 3.1, 3.2, 3.3, and 3.4, suppose that TRish is run with* $\gamma_1 > \gamma_2 > 0$ *such that* $\frac{\gamma_1}{\gamma_2} < \frac{h_2}{h_2 - 1}$ *(meaning* $\gamma_1 - h_2(\gamma_1 - \gamma_2) > 0$*) and with* $\alpha_k = \alpha$ *for all* $k \in \mathbb{N}$ *such that*

$$0 < \alpha \leq \min\left\{\frac{1}{2c\theta_1}, \frac{1}{\gamma_1 L M_2}\right\},$$

where

$$\theta_1 = \tfrac{1}{2}\min\{\gamma_2, \gamma_1 - h_2(\gamma_1 - \gamma_2)\} > 0. \tag{3.11}$$

Then, for all $k \in \mathbb{N}$*, the expected optimality gap satisfies*

$$\mathbb{E}[f(x_{k+1})] - f_* \leq \frac{\theta_2}{2c\alpha\theta_1} + (1 - 2c\alpha\theta_1)^{k-1}\left(f(x_1) - f_* - \frac{\theta_2}{2c\alpha\theta_1}\right) \xrightarrow{k\to\infty} \frac{\theta_2}{2c\alpha\theta_1}, \tag{3.12}$$

where

$$\theta_2 = \max\{\tfrac{1}{2}\gamma_1^2 L M_1 \alpha^2, h_1(\gamma_1 - \gamma_2)\alpha + \tfrac{1}{2}L\alpha^2\} > 0. \tag{3.13}$$

Proof. Proof. Combining the result of Lemma 3.1 and (3.9), it follows that, for all $k \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}_k[f(x_{k+1})] - f(x_k) \\
&\leq \begin{cases} -\gamma_1\alpha(1 - \frac{1}{2}\gamma_1 L M_2 \alpha)\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\gamma_1^2 L M_1 \alpha^2 & \text{if } \|g_k\|_2 \in [0, \frac{1}{\gamma_1}) \\ -\gamma_1\alpha\|\nabla f(x_k)\|_2^2 + (\gamma_1 - \gamma_2)\alpha(h_1 + h_2\|\nabla f(x_k)\|_2^2) + \frac{1}{2}L\alpha^2 & \text{if } \|g_k\|_2 \in [\frac{1}{\gamma_1}, \frac{1}{\gamma_2}] \\ -\gamma_2\alpha(1 - \frac{1}{2}\gamma_2 L M_2 \alpha)\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\gamma_2^2 L M_1 \alpha^2 & \text{if } \|g_k\|_2 \in (\frac{1}{\gamma_2}, \infty). \end{cases}
\end{aligned} \tag{3.14}$$

Therefore, with (θ_1, θ_2) defined in (3.11)/(3.13), it follows with (3.8) that, for all $k \in \mathbb{N}$,

$$\begin{aligned}
\mathbb{E}_k[f(x_{k+1})] - f(x_k) &\leq -\alpha\theta_1\|\nabla f(x_k)\|_2^2 + \theta_2 \\
&\leq -2c\alpha\theta_1(f(x_k) - f_*) + \theta_2.
\end{aligned}$$

Adding and subtracting f_* on the left-hand side, taking total expectations, and rearranging yields

$$\begin{aligned}
\mathbb{E}[f(x_{k+1})] - f_* &\leq (1 - 2c\alpha\theta_1)(\mathbb{E}[f(x_k)] - f_*) + \theta_2 \\
&= \frac{\theta_2}{2c\alpha\theta_1} + (1 - 2c\alpha\theta_1)(\mathbb{E}[f(x_k)] - f_*) + \theta_2 - \frac{\theta_2}{2c\alpha\theta_1} \\
&= \frac{\theta_2}{2c\alpha\theta_1} + (1 - 2c\alpha\theta_1)\left(\mathbb{E}[f(x_k)] - f_* - \frac{\theta_2}{2c\alpha\theta_1}\right).
\end{aligned}$$

Since $1 - 2c\alpha\theta_1 \in (0, 1)$, this represents a contraction inequality. Applying the result repeatedly through iteration $k \in \mathbb{N}$, one obtains the desired result. □ □

It is worthwhile to compare the result of Theorem 3.1 with a corresponding result known to hold for a straightforward stochastic gradient method. For example, from [1, Thm. 4.6] with our notation, it is known that for a stochastic gradient method with fixed stepsize $\alpha = \frac{1}{LM_2}$ an upper bound for the expected optimality gap converges to $\frac{\alpha L M_1}{2c} = \frac{M_1}{2cM_2}$. On the other hand, the analysis in Theorem 3.1 shows that if $c \approx 0$ and $\|g_k\|_2 \in [\frac{1}{\gamma_1}, \frac{1}{\gamma_2}]$ for all $k \in \mathbb{N}$, then TRish with $\alpha = \frac{1}{\gamma_1 L M_2}$ yields an upper bound for the expected optimality gap that converges to

$$\frac{h_1(\gamma_1 - \gamma_2) + \frac{1}{2}L\alpha}{c(\gamma_1 - h_2(\gamma_1 - \gamma_2))} = \frac{h_1(\gamma_1 - \gamma_2)}{c(\gamma_1 - h_2(\gamma_1 - \gamma_2))} + \frac{1}{2cM_2\gamma_1(\gamma_1 - h_2(\gamma_1 - \gamma_2))}. \tag{3.15}$$

We can now make a couple observations. On one hand, if $h_1 \approx M_1$ and $h_2 \approx M_2 \approx 1$, then the condition that $\frac{\gamma_1}{\gamma_2} < \frac{h_2}{h_2 - 1}$ essentially does not restrict (γ_1, γ_2), in which case (3.15) is approximately

$$\frac{M_1(\gamma_1 - \gamma_2)}{c\gamma_2} + \frac{1}{2c\gamma_1\gamma_2}.$$

This quantity is less than $\frac{M_1}{2c}$, i.e., the approximate bound for the stochastic gradient method, if $\gamma_1 = \frac{5}{4}\gamma_2$ with $\gamma_2 \geq \sqrt{\frac{8}{5M_1}}$. On the other hand, if $h_1 \approx M_1 \approx h_2 \approx M_2 \gg 1$, then the condition that $\frac{\gamma_1}{\gamma_2} < \frac{h_2}{h_2 - 1}$ essentially requires that $\gamma_1 \approx \gamma_2$. In particular, supposing $\gamma_1 - \gamma_2 \approx 1/h_1 \approx 1/h_2$, the bound (3.15) is approximately

$$\frac{1}{c(\gamma_1 - 1)} + \frac{1}{2cM_2\gamma_1(\gamma_1 - 1)}.$$

This quantity is less than $\frac{1}{2c}$, i.e., the approximate bound for the stochastic gradient method in this case, if one has $\gamma_1 \in \left[\frac{3M_2 + \sqrt{9M_2^2 + 4M_2}}{2M_2}, \infty\right) \approx [3, \infty)$. Overall, we have shown in both of these cases that TRish possesses an asymptotic bound on the expected optimality gap that is at least on par with a straightforward stochastic gradient method. Of course, in this discussion, we have assumed that $\|g_k\|_2 \in [\frac{1}{\gamma_1}, \frac{1}{\gamma_2}]$ for all $k \in \mathbb{N}$, which is perhaps unlikely to hold in general. However, since TRish behaves as a stochastic gradient method otherwise, one may still conclude that, overall, the theoretical behavior of TRish is at least on par with that of a stochastic gradient method when fixed stepsizes are employed.

3.2 P-L Condition, Sublinearly Diminishing Variance and Stepsizes

Let us now consider the behavior of TRish when the P-L condition holds, diminishing stepsizes are employed (see Theorem 3.2 later on), and the inner products between the gradients and stochastic gradients satisfy the following assumption.

Assumption 3.5. *There exist constants $h_3 \in (0, \infty)$ and $h_4 \in (1, \infty)$ such that, for all $k \in \mathbb{N}$,*

$$\mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] \leq h_3\alpha_k + h_4\|\nabla f(x_k)\|_2^2. \tag{3.16}$$

We motivate this assumption by providing the following examples. In this case, we show that it can be satisfied if the stochastic gradient estimates become increasingly accurate.

Example 3.3. *Consider the scenario in Example 3.1. Then, if $\sigma_k \leq \alpha_k$ for all $k \in \mathbb{N}$ with $\alpha_k \leq \alpha$ for some $\alpha \in (0, \infty)$ for all $k \in \mathbb{N}$, it follows that Assumption 3.5 holds with*

$$h_3 = \frac{1}{2\sqrt{2\pi}} \quad \text{and} \quad h_4 = 1 + \frac{\alpha}{2\sqrt{2\pi}}. \tag{3.17}$$

Example 3.4. *Consider the scenario in Example 3.2. Then, if $\sqrt{\mu_k^T \Sigma_k \mu_k} \leq \alpha_k$ for all $k \in \mathbb{N}$ with $\alpha_k \leq \alpha$ for some $\alpha \in (0, \infty)$ for all $k \in \mathbb{N}$, Assumption 3.5 holds with h_3 and h_4 from (3.17).*

Our second theorem on the behavior of TRish is now proved as the following.

Theorem 3.2. *Under Assumptions 3.1, 3.2, 3.3, and 3.5, suppose that TRish is run with $\gamma_1 > \gamma_2 > 0$ such that $\frac{\gamma_1}{\gamma_2} < \frac{h_4}{h_4 - 1}$ (meaning $\gamma_1 - h_4(\gamma_1 - \gamma_2) > 0$), and with, for all $k \in \mathbb{N}$,*

$$\alpha_k = \frac{a}{b + k} \quad \text{for some} \quad a \in \left(\frac{1}{2c\beta_1}, \frac{b+1}{2c\beta_1}\right) \quad \text{and} \quad b > 0 \quad \text{such that} \quad \alpha_1 \in \left(0, \frac{1}{\gamma_1 L M_2}\right],$$

where

$$\beta_1 = \tfrac{1}{2}\min\{\gamma_2, \gamma_1 - h_4(\gamma_1 - \gamma_2)\} > 0. \tag{3.18}$$

Then, for all $k \in \mathbb{N}$, the expected optimality gap satisfies

$$\mathbb{E}[f(x_k)] - f_* \leq \frac{\nu}{b + k}, \tag{3.19}$$

where

$$\nu = \max\left\{\frac{a^2\beta_2}{2ac\beta_1 - 1}, (b+1)(f(x_1) - f_*)\right\} > 0 \tag{3.20}$$

$$\text{and} \quad \beta_2 = \max\{h_3(\gamma_1 - \gamma_2) + \tfrac{1}{2}L, \tfrac{1}{2}\gamma_1^2 L M_1\} > 0. \tag{3.21}$$

Proof. Proof. Combining the result of Lemma 3.1 and (3.16), it follows that, for all $k \in \mathbb{N}$,

$$\begin{aligned}&\mathbb{E}_k[f(x_{k+1})] - f(x_k)\\ &\leq \begin{cases} -\gamma_1\alpha_k(1-\frac{1}{2}\gamma_1 LM_2\alpha_k)\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\gamma_1^2 LM_1\alpha_k^2 & \text{if } \|g_k\|_2 \in [0, \frac{1}{\gamma_1}) \\ -\gamma_1\alpha_k\|\nabla f(x_k)\|_2^2 + (\gamma_1-\gamma_2)\alpha_k(h_3\alpha_k + h_4\|\nabla f(x_k)\|_2^2) + \frac{1}{2}L\alpha_k^2 & \text{if } \|g_k\|_2 \in [\frac{1}{\gamma_1}, \frac{1}{\gamma_2}] \\ -\gamma_2\alpha_k(1-\frac{1}{2}\gamma_2 LM_2\alpha_k)\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\gamma_2^2 LM_1\alpha_k^2 & \text{if } \|g_k\|_2 \in (\frac{1}{\gamma_2}, \infty). \end{cases}\end{aligned}$$

Therefore, with (β_1, β_2) defined in (3.18)/(3.21), it follows with (3.8) that, for all $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}_k[f(x_{k+1})] - f(x_k) &\leq -\beta_1\alpha_k\|\nabla f(x_k)\|^2 + \beta_2\alpha_k^2\\ &\leq -2c\beta_1\alpha_k(f(x_k) - f_*) + \beta_2\alpha_k^2.\end{aligned}$$

Adding and subtracting f_* on the left-hand side, taking total expectations, and rearranging yields

$$\mathbb{E}[f(x_{k+1})] - f_* \leq (1 - 2c\beta_1\alpha_k)(\mathbb{E}[f(x_k)] - f_*) + \beta_2\alpha_k^2.$$

Let us now prove (3.19) by induction. First, for $k = 1$, the inequality holds by the definition of ν. Now suppose that (3.19) holds up to $k \in \mathbb{N}$; then, for $k+1$, one finds from above that

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] - f_* &\leq (1 - 2c\beta_1\alpha_k)(\mathbb{E}[f(x_k)] - f_*) + \beta_2\alpha_k^2\\ &= \left(1 - \frac{2ac\beta_1}{b+k}\right)(\mathbb{E}[f(x_k)] - f_*) + \frac{a^2\beta_2}{(b+k)^2}\\ &\leq \left(1 - \frac{2ac\beta_1}{b+k}\right)\frac{\nu}{b+k} + \frac{a^2\beta_2}{(b+k)^2}\\ &= \frac{(b+k)\nu}{(b+k)^2} - \frac{2ac\beta_1\nu}{(b+k)^2} + \frac{a^2\beta_2}{(b+k)^2}\\ &= \frac{(b+k-1)\nu}{(b+k)^2} - \frac{(2ac\beta_1 - 1)\nu}{(b+k)^2} + \frac{a^2\beta_2}{(b+k)^2}\\ &\leq \frac{(b+k-1)\nu}{(b+k)^2} \leq \frac{\nu}{b+k+1},\end{aligned}$$

where the last two inequalities follow from the definition of ν and since $(b+k-1)(b+k+1) \leq (b+k)^2$, respectively. The desired conclusion now follows from this inductive argument. □ □

3.3 P-L Condition, Linearly Diminishing Variance with a Fixed Stepsize

Let us now prove a convergence result for TRish when the P-L condition holds, a sufficiently small fixed stepsize is used (see Theorem 3.3 later on), and the stochastic gradients satisfy the following.

Assumption 3.6. *There exist constants $M_3 \in (0,\infty)$ and $\zeta \in (0,1)$ such that*

$$\mathbb{E}_k[\|g_k\|_2^2] \leq M_3\zeta^{k-1} + \|\nabla f(x_k)\|_2^2. \tag{3.22}$$

In addition, there exist constants $h_5 \in (0,\infty)$, $h_6 \in (1,\infty)$, and $\lambda \in (0,1)$ such that

$$\mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] \leq h_5\lambda^{k-1} + h_6\|\nabla f(x_k)\|_2^2. \tag{3.23}$$

Example 3.5. *Consider the scenario in Example 3.1. Then, since (3.22) implies that $\sigma_k^2 \leq M_3\zeta^{k-1}$ for all $k \in \mathbb{N}$, it follows along with the fact that $\zeta \in (0,1)$ that*

$$\begin{aligned}\mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] &\leq \frac{\sigma_k}{2\sqrt{2\pi}} + \left(1 + \frac{\sigma_k}{2\sqrt{2\pi}}\right)\mu_k^2\\ &\leq \frac{\sqrt{M_3}}{2\sqrt{2\pi}}(\sqrt{\zeta})^{k-1} + \left(1 + \frac{\sqrt{M_3}}{2\sqrt{2\pi}}\right)\mu_k^2.\end{aligned}$$

Hence, it follows that (3.23) holds with

$$h_5 = \frac{\sqrt{M_3}}{2\sqrt{2\pi}}, \quad h_6 = 1 + \frac{\sqrt{M_3}}{2\sqrt{2\pi}}, \quad \text{and} \quad \lambda = \sqrt{\zeta}. \tag{3.24}$$

Example 3.6. *Consider the scenario in Example 3.2. Then, with $\sqrt{\mu_k^T \Sigma_k \mu_k} \leq M_3 \zeta^{k-1}$ for all $k \in \mathbb{N}$, it follows that (3.23) holds with h_5, h_6, and λ from (3.24).*

Our next theorem on the behavior of TRish is now proved as the following.

Theorem 3.3. *Under Assumptions 3.1, 3.2, 3.3, and 3.6, suppose that TRish is run with $\gamma_1 > \gamma_2 > 0$ such that $\frac{\gamma_1}{\gamma_2} < \frac{h_6}{h_6 - 1}$ (meaning $\gamma_1 - h_6(\gamma_1 - \gamma_2) > 0$), and with $\alpha_k = \alpha$ for all $k \in \mathbb{N}$ such that*

$$0 < \alpha \leq \min\left\{\frac{\gamma_1 - h_6(\gamma_1 - \gamma_2)}{\gamma_1^2 L}, \frac{1}{\gamma_1 L}, \frac{1}{c\kappa_1}\right\}, \tag{3.25}$$

where

$$\kappa_1 := \tfrac{1}{2} \min\{\gamma_2, \gamma_1 - h_6(\gamma_1 - \gamma_2)\} > 0.$$

Then, for all $k \in \mathbb{N}$, the expected optimality gap satisfies

$$\mathbb{E}[f(x_k)] - f_* \leq \omega \rho^{k-1}, \tag{3.26}$$

where

$$\begin{aligned} \kappa_2 &:= h_5(\gamma_1 - \gamma_2) + \tfrac{1}{2}\gamma_1^2 \alpha L M_3 > 0, \\ \omega &:= \max\{f(x_1) - f_*, \tfrac{\kappa_2}{c\kappa_1}\} > 0, \\ \text{and} \quad \rho &:= \max\{1 - \alpha c \kappa_1, \lambda, \zeta\} \in (0,1). \end{aligned}$$

Proof. Proof. For any $k \in \mathbb{N}$, as in the proof of Lemma 3.1, one finds in case 1 with (3.22) and (3.25) that

$$\begin{aligned} \mathbb{E}_k[f(x_{k+1})] &\leq f(x_k) - \gamma_1 \alpha_k \nabla f(x_k)^T \mathbb{E}_k[g_k] + \tfrac{1}{2}\gamma_1^2 L \alpha_k^2 \mathbb{E}_k[\|g_k\|_2^2] \\ &\leq f(x_k) - \gamma_1 \alpha_k \|\nabla f(x_k)\|_2^2 + \tfrac{1}{2}\gamma_1^2 L \alpha_k^2 (M_3 \zeta^{k-1} + \|\nabla f(x_k)\|_2^2) \\ &= f(x_k) - \alpha\gamma_1(1 - \tfrac{1}{2}\gamma_1 \alpha L)\|\nabla f(x_k)\|_2^2 + \tfrac{1}{2}\gamma_1^2 \alpha^2 L M_3 \zeta^{k-1} \\ &\leq f(x_k) - \tfrac{1}{2}\alpha\gamma_1 \|\nabla f(x_k)\|_2^2 + \tfrac{1}{2}\gamma_1^2 \alpha^2 L M_3 \zeta^{k-1}. \end{aligned}$$

Similarly, one finds in case 3 that

$$\mathbb{E}_k[f(x_{k+1})] \leq f(x_k) - \tfrac{1}{2}\alpha\gamma_2 \|\nabla f(x_k)\|_2^2 + \tfrac{1}{2}\gamma_2^2 \alpha^2 L M_3 \zeta^{k-1}.$$

Finally, in case 2, in which one has $\gamma_1 \|g_k\|_2 \geq 1$, one finds that

$$\begin{aligned} f(x_{k+1}) = f\left(x_k - \frac{\alpha_k}{\|g_k\|_2} g_k\right) &\leq f(x_k) - \frac{\alpha_k \nabla f(x_k)^T g_k}{\|g_k\|_2} + \tfrac{1}{2} L \alpha_k^2 \\ &\leq f(x_k) - \frac{\alpha_k \nabla f(x_k)^T g_k}{\|g_k\|_2} + \tfrac{1}{2}\gamma_1^2 \alpha^2 L \|g_k\|_2^2, \end{aligned}$$

which, as in the proof of Lemma 3.1 and with (3.22), (3.23), and (3.25), yields

$$\begin{aligned} &\mathbb{E}_k[f(x_{k+1})] \\ &= \mathbb{P}_k[E_k]\mathbb{E}_k[f(x_{k+1})|E_k] + (1 - \mathbb{P}_k[E_k])\mathbb{E}_k[f(x_{k+1})|\overline{E}_k] \\ &\leq f(x_k) - \mathbb{P}_k[E_k]\gamma_2 \alpha \mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] - (1 - \mathbb{P}_k[E_k])\gamma_1 \alpha \mathbb{E}_k[\nabla f(x_k)^T g_k | \overline{E}_{x_k}] + \tfrac{1}{2}\gamma_1^2 \alpha^2 L \mathbb{E}_k[\|g_k\|_2^2] \\ &= f(x_k) - \alpha\gamma_1 \|\nabla f(x_k)\|_2^2 + (\gamma_1 - \gamma_2)\alpha \mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] + \tfrac{1}{2}\gamma_1^2 \alpha^2 L \mathbb{E}_k[\|g_k\|_2^2] \\ &= f(x_k) - \alpha(\gamma_1 - h_6(\gamma_1 - \gamma_2) - \tfrac{1}{2}\gamma_1^2 \alpha L)\|\nabla f(x_k)\|_2^2 + (\gamma_1 - \gamma_2)\alpha h_5 \lambda^{k-1} + \tfrac{1}{2}\gamma_1^2 \alpha^2 L M_3 \zeta^{k-1} \\ &\leq f(x_k) - \tfrac{1}{2}\alpha(\gamma_1 - h_6(\gamma_1 - \gamma_2))\|\nabla f(x_k)\|_2^2 + (\gamma_1 - \gamma_2)\alpha h_5 \lambda^{k-1} + \tfrac{1}{2}\gamma_1^2 \alpha^2 L M_3 \zeta^{k-1}. \end{aligned}$$

Combining all cases, it follows that

$$\begin{aligned}\mathbb{E}_k[f(x_{k+1})] &\leq f(x_k) - \alpha\kappa_1\|\nabla f(x_k)\|^2 + \alpha\kappa_2 \max\{\lambda,\zeta\}^{k-1} \\ &\leq f(x_k) - 2\alpha c\kappa_1(f(x_k) - f_*) + \alpha\kappa_2 \max\{\lambda,\zeta\}^{k-1},\end{aligned}$$

from which it follows that

$$\mathbb{E}[f(x_{k+1})] - f_* \leq (1 - 2\alpha c\kappa_1)(\mathbb{E}[f(x_k)] - f_*) + \alpha\kappa_2 \max\{\lambda,\zeta\}^{k-1}.$$

Let us now prove (3.26) by induction. First, for $k = 1$, the inequality follows by the definition of ω. Then, assuming the inequality holds true for $k \in \mathbb{N}$, one finds that

$$\begin{aligned}\mathbb{E}[f(x_{k+1})] - f_* &\leq (1 - 2\alpha c\kappa_1)(\mathbb{E}[f(x_k)] - f_*) + \alpha\kappa_2 \max\{\lambda,\zeta\}^{k-1} \\ &\leq (1 - 2\alpha c\kappa_1)\omega\rho^{k-1} + \alpha\kappa_2 \max\{\lambda,\zeta\}^{k-1} \\ &\leq \omega\rho^{k-1}\left(1 - 2\alpha c\kappa_1 + \frac{\alpha\kappa_2}{\omega}\left(\frac{\max\{\lambda,\zeta\}}{\rho}\right)^{k-1}\right) \\ &\leq \omega\rho^{k-1}\left(1 - 2\alpha c\kappa_1 + \frac{\alpha\kappa_2}{\omega}\right) \\ &\leq \omega\rho^{k-1}(1 - \alpha c\kappa_1) \\ &\leq \omega\rho^k,\end{aligned}$$

which proves that the conclusion holds for $k+1$, as desired. □ □

3.4 No P-L Condition, Bounded Variance with a Fixed Stepsize

Let us now consider the behavior of TRish when the P-L condition does not hold. Our first such result involves the use of a sufficiently small fixed positive stepsize.

Theorem 3.4. *Under Assumptions 3.1, 3.2, and 3.4, suppose that TRish is run with $\gamma_1 > \gamma_2 > 0$ such that $\frac{\gamma_1}{\gamma_2} < \frac{h_2}{h_2-1}$ (meaning $\gamma_1 - h_2(\gamma_1 - \gamma_2) > 0$) and with $\alpha_k = \alpha$ for all $k \in \mathbb{N}$ such that*

$$0 < \alpha \leq \frac{1}{\gamma_1 L M_2}.$$

Then, with (θ_1, θ_2) defined in (3.11)/(3.13), it follows that, for all $K \in \mathbb{N}$,

$$\mathbb{E}\left[\sum_{k=1}^{K}\|\nabla f(x_k)\|_2^2\right] \leq \frac{K\theta_2}{\alpha\theta_1} + \frac{f(x_1) - f_*}{\alpha\theta_1} \tag{3.27a}$$

$$\text{and}\quad \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla f(x_k)\|_2^2\right] \leq \frac{\theta_2}{\alpha\theta_1} + \frac{f(x_1) - f_*}{K\alpha\theta_2} \xrightarrow{K\to\infty} \frac{\theta_2}{\alpha\theta_1}. \tag{3.27b}$$

Proof. Proof. As in the proof of Theorem 3.1, combining the result of Lemma 3.1 and (3.9), it follows that the inequality (3.14) holds for all $k \in \mathbb{N}$. Taking total expectations, it follows that, for all $k \in \mathbb{N}$,

$$\mathbb{E}[f(x_{k+1})] - \mathbb{E}[f(x_k)] \leq -\alpha\theta_1\mathbb{E}[\|\nabla f(x_k)\|_2^2] + \theta_2.$$

Summing both sides for $k \in \{1, \ldots, K\}$ yields

$$f_* - f(x_1) \leq \mathbb{E}[f(x_{K+1})] - f(x_1) \leq -\alpha\theta_1\sum_{k=1}^{K}\mathbb{E}[\|\nabla f(x_k)\|_2^2] + K\theta_2.$$

Rearranging yields (3.27a), then dividing by K yields (3.27b). □ □

As in the case of [1, Thm. 4.8], this result shows that while one cannot bound the expected optimality gap as when the P-L condition holds, one can bound the average norm of the gradients of the objective that are observed during the optimization process.

3.5 No P-L Condition, Sublinearly Diminishing Variance and Stepsizes

Finally, let us consider the behavior of TRish when the P-L condition does not hold and diminishing stepsizes are employed.

Theorem 3.5. *Under Assumptions 3.1, 3.2, and 3.5, suppose that TRish is run with* $\gamma_1 > \gamma_2 > 0$ *such that* $\frac{\gamma_1}{\gamma_2} < \frac{h_4}{h_4 - 1}$ *(meaning* $\gamma_1 - h_4(\gamma_1 - \gamma_2) > 0$*), and with* $\{\alpha_k\}$ *satisfying*

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \textit{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty. \tag{3.28}$$

Then, with $A_K := \sum_{k=1}^{K} \alpha_k$*, it follows that*

$$\lim_{K\to\infty} \mathbb{E}\left[\sum_{k=1}^{K} \alpha_k \|\nabla f(x_k)\|_2^2\right] < \infty \tag{3.29a}$$

$$\textit{and} \quad \mathbb{E}\left[\frac{1}{A_K}\sum_{k=1}^{K} \alpha_k \|\nabla f(x_k)\|_2^2\right] \xrightarrow{K\to\infty} 0. \tag{3.29b}$$

Proof. Proof. By (3.28), it follows that $\alpha_k \leq 1/(LM_1)$ for all sufficiently large $k \in \mathbb{N}$. Hence, without loss of generality, one may assume that $\alpha_k \leq 1/(LM_1)$ for all $k \in \mathbb{N}$. Then, as in the proof of Theorem 3.2, it follows by Lemma 3.1, (3.16), and taking total expectations that, with (β_1, β_2) defined in (3.18)/(3.21), one has, for all $k \in \mathbb{N}$,

$$\mathbb{E}[f(x_{k+1})] - \mathbb{E}[f(x_k)] \leq -\beta_1 \alpha_k \mathbb{E}[\|\nabla f(x_k)\|^2] + \beta_2 \alpha_k^2.$$

Summing both sides for $k \in \{1, \dots, K\}$ yields

$$f_* - f(x_1) \leq \mathbb{E}[f(x_{K+1})] - f(x_1) \leq -\beta_1 \sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla f(x_k)\|_2^2] + \beta_2 \sum_{k=1}^{K} \alpha_k^2,$$

which after rearrangement gives

$$\sum_{k=1}^{K} \alpha_k \mathbb{E}[\|\nabla f(x_k)\|_2^2] \leq \frac{f(x_1) - f_*}{\beta_1} + \frac{\beta_2}{\beta_1} \sum_{k=1}^{K} \alpha_k^2.$$

From (3.28), it follows that the right-hand side converges to a finite limit as $K \to \infty$, giving (3.29a). Then, the limit (3.29b) follows since (3.28) ensures that $\{A_K\} \to \infty$ as $K \to \infty$. □ □

A consequence of this theorem is the straightforward fact that

$$\liminf_{k\to\infty} \mathbb{E}[\|\nabla f(x_k)\|_2^2] = 0.$$

That is, under the conditions of the theorem, the expected squared norms of the gradients at the iterates of the algorithm cannot stay bounded away from zero.

4 Numerical Experiments

In this section, we provide the results of numerical experiments to demonstrate the performance of TRish compared to a traditional stochastic gradient (SG) approach. Through solving two machine learning test problems involving objective functions of the form (2.2)—one convex and one nonconvex—we demonstrate that TRish can outperform SG in terms of achieving better training loss, training accuracy, testing loss, and testing accuracy with comparable computational effort.

4.1 Logistic Regression

As a first test case, we consider a problem of binary classification through logistic regression involving the `rcv1.binary` dataset available in the well-known LIBSVM repository; see [2]. In particular, with training feature vectors $\{z_i\}_{i=1}^N \in \mathbb{R}^n \times \cdots \times \mathbb{R}^n$ and training labels $\{y_i\}_{i=1}^N \in \{-1,1\} \times \cdots \times \{-1,1\}$, the objective of this problem has the form

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + e^{-y_i(x^T z_i)}). \tag{4.1}$$

Also available is a testing dataset $\{(\bar{z}_i, \bar{y}_i)\}_{i=1}^{\overline{N}}$. For this problem, the feature vectors have length $n = 47236$, the number of points in the training set is $N = 20242$, and the number of points in the testing set is $\overline{N} = 677399$. In our experiments, we ran implementations of TRish and SG and compare performance by comparing training and testing losses (i.e., the objective function (4.1) evaluated with the training and testing data, respectively) as well as training and testing accuracy (i.e., for a given approximate solution, what fraction of the training and testing set, respectively, is classified correctly) for iterates throughout the optimization process.

For our purposes, we ran each algorithm for a single epoch (i.e., until N training pairs have been accessed) with a fixed stepsize. To tune parameters for each algorithm, we searched over various mini-batch sizes for the stochastic gradient computations, stepsizes, and, for TRish, values for the parameters γ_1 and γ_2. For each algorithm and each combination of these parameters, we ran the algorithm five times from the same starting point, keeping track of the parameter settings that lead to be best average testing accuracy over the five runs.

For SG, the parameters that we tuned are the mini-batch size b, for which we considered the values in the set $\{5, 10, 20\}$, and the fixed stepsize α, for which we considered the values $\{1, 10, 100\}$. The values that led to the best average testing accuracy were $(b, \alpha) = (10, 10)$. For TRish, the same sets for b and α were considered, and in addition we considered values for γ_1 from the set $\{7, 11, 15\}$, where for each case we set $\gamma_2 \leftarrow 0.4\gamma$. The values that led to the best average testing accuracy for TRish were $(b, \alpha, \gamma_1, \gamma_2) = (10, 5, 15, 6)$.

Using these tuned parameter choices, the average training losses, training accuracies, testing losses, and testing accuracies obtained by SG and TRish after one epoch are provided in Table 1. We also provide, in Figures 2 and 3, these quantities after various fractions of the first epoch. These plots show that while the two algorithms appear to be converging to similar final values, TRish yields lower values throughout the first epoch.

	Average Training Loss	Average Training Accuracy	Average Testing Loss	Average Testing Accuracy
SG	0.1014	0.9749	0.1345	0.9569
TRish	0.0521	0.9847	0.1174	0.9584

Table 1: Performance measures after 1 epoch for SG and TRish employed to minimize the logistic regression function (4.1) using the `rcv1.binary` dataset.

4.2 Neural Network Training

As a second test case, we consider the problem to train a convolutional neural network to classify handwritten digits in the well-known `mnist` dataset; see [7]. Implemented using `tensorflow`, the neural network that we consider is composed of two convolutional layers followed by two fully connected layers. The total number of trainable parameters is 535040. The training dataset involves 60000 feature/label pairs while the testing set involves 10000 pairs.

For the purposes of this test problem, we used the same experimental set-up as in the previous section, except that different sets of mini-batch sizes, stepsizes, and TRish parameters were explored. In particular,

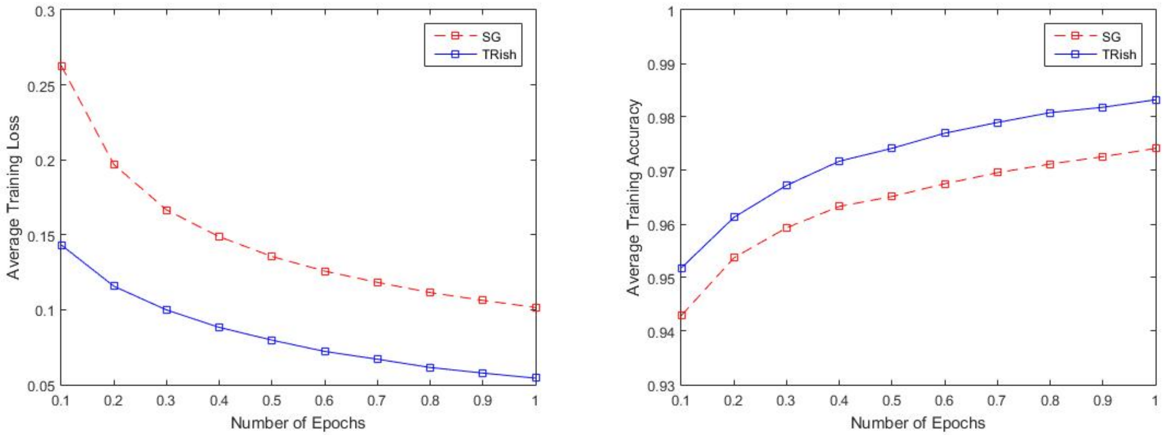

Figure 2: Average training loss and accuracy during the first epoch when SG and TRish employed to minimize the logistic regression function (4.1) using the `rcv1.binary` dataset.

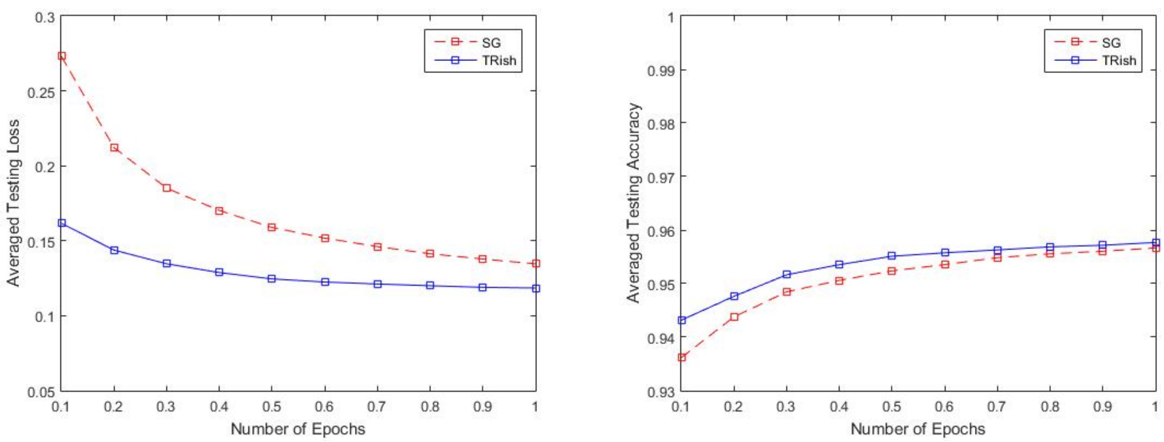

Figure 3: Average testing loss and accuracy during the first epoch when SG and TRish employed to minimize the logistic regression function (4.1) using the `rcv1.binary` dataset.

we considered $b \in \{10, 20, 40, 80\}$, $\alpha \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$, $\gamma_1 \in \{2, 4, 6, 8\}$, and $\gamma_2 \in \{0.5, 0.25, 0.125\}$. After tuning, the values that led to the best average testing accuracy for SG were $(b, \alpha) = (40, 10^{-2})$ and those that led to the best average testing accuracy for TRish were $(b, \alpha, \gamma_1, \gamma_2) = (40, 10^{-1}, 8, 0.5)$. The average training losses, training accuracies, testing losses, and testing accuracies obtained by the methods after one epoch are provided in Table 2 and these quantities after various fractions of the first epoch are plotted in Figures 4 and 5. Again, these results show that TRish offers better performance throughout the first epoch, even more noticably than for the convex problem in the previous subsection.

5 Conclusion

An algorithm inspired by a trust region methodology has been proposed, analyzed, and tested for solving stochastic and finite sum minimization problems. Our proved theoretical guarantees show that our method, deemed TRish, has convergence properties that are similar to a traditional stochastic gradient method. Our numerical results, on the other hand, show that TRish can outperform a traditional SG approach. We

	Average Training Loss	Average Training Accuracy	Average Testing Loss	Average Testing Accuracy
SG	0.1479	0.9557	0.1415	0.9581
TRish	0.0860	0.9737	0.0793	0.9751

Table 2: Performance measures after 1 epoch for SG and TRish employed to train a convolutional neural network using the mnist dataset.

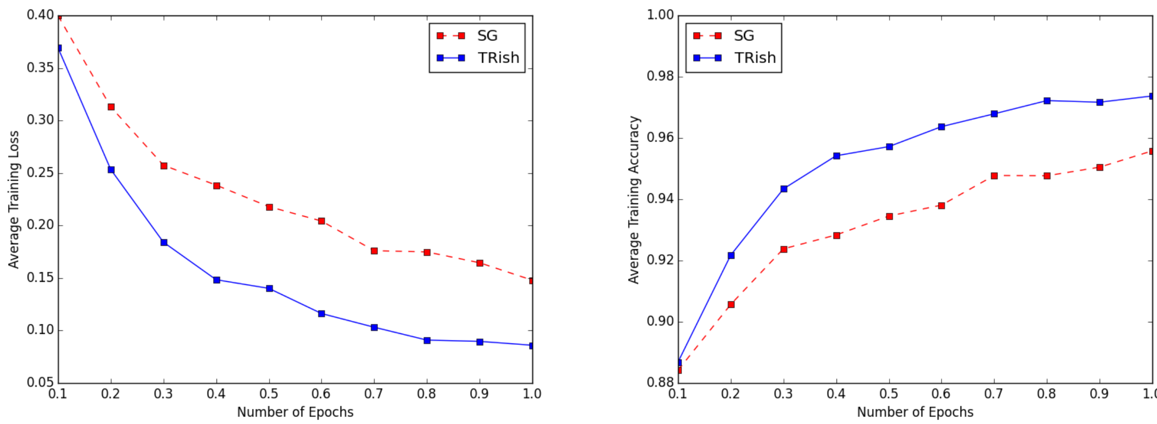

Figure 4: Average training loss and accuracy during the first epoch when SG and TRish employed to train a convolutional neural network using the mnist dataset.

attribute this better behavior to the algorithm's use of normalized steps, which one can argue lessens its dependence on problem-specific quantities.

While not considered in this paper, we believe it would be interesting to explore the incorporation within TRish of various enhancements, such as the use of second-derivative (i.e., Hessian) approximations, acceleration, and/or momentum. These might further improve the practical performance of the framework set forth in this paper.

Acknowledgment

All authors were supported by the U.S. National Science Foundation's Division of Computing and Communication Foundations and Division of Mathematical Sciences under award numbers CCF–1618717 and DMS–1319356. The first author was also supported by the U.S. Department of Energy, Office of Science, Applied Mathematics, Early Career Research Program under Award Number DE–SC0010615.

References

[1] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. arXiv 1606.04838, 2016.

[2] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

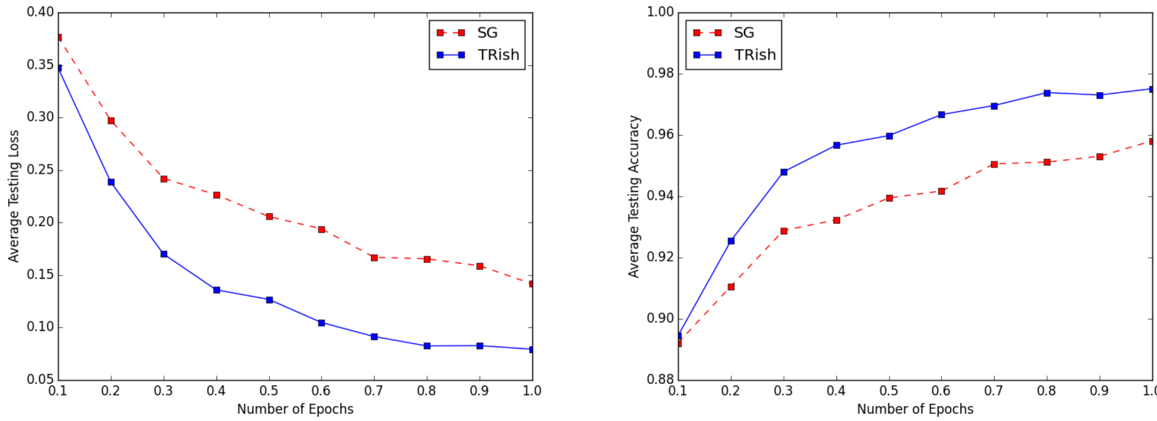

Figure 5: Average testing loss and accuracy during the first epoch when SG and TRish employed to train a convolutional neural network using the `mnist` dataset.

[3] Ruobing Chen, Matt Menickelly, and Katya Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, pages 1–41, 2015.

[4] Elad Hazan, Kfir Levy, and Shai Shalev-Shwartz. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pages 1594–1602, 2015.

[5] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[6] Jeffrey Larson and Stephen C Billups. Stochastic derivative-free optimization using a trust region framework. *Computational Optimization and Applications*, 64(3):619–645, 2016.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE, 86(11)*, pages 2278–2324, 1009.

[8] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[9] Yung Liang Tong. *The multivariate normal distribution*. Springer Science & Business Media, 2012.