STOCHASTIC THREE POINTS METHOD FOR UNCONSTRAINED SMOOTH MINIMIZATION

EL HOUCINE BERGOU*, EDUARD GORBUNOV†, AND PETER RICHTÁRIK‡

Abstract. In this paper we consider the unconstrained minimization problem of a smooth function in $\mathbb{R}^n$ in a setting where only function evaluations are possible. We design a novel randomized derivative-free algorithm — the *stochastic three points (STP)* method — and analyze its iteration complexity. At each iteration, `STP` generates a random search direction according to a certain fixed probability law. Our assumptions on this law are very mild: roughly speaking, all laws which do not concentrate all measure on any halfspace passing through the origin will work. For instance, we allow for the uniform distribution on the sphere and also distributions that concentrate all measure on a positive spanning set.

Although our approach is designed to not use explicitly derivatives, it covers some first order methods. For instance if the probability law is chosen to be the Dirac distribution concentrated at the sign of the gradient then `STP` recovers the Signed Gradient Descent method. If the probability law is the uniform distribution on the coordinates of the gradient then `STP` recovers the Coordinate Descent Method.

Given a current iterate x, `STP` compares the objective function at three points: x, $x + \alpha s$ and $x - \alpha s$, where $\alpha > 0$ is a stepsize parameter and s is the random search direction. The best of these three points is the next iterate. We analyze the method `STP` under several stepsize selection schemes (fixed, decreasing, estimated through finite differences, etc).

The complexity of `STP` depends on the probability law via a simple characteristic closely related to the cosine measure which is used in the analysis of deterministic direct search (`DDS`) methods. Unlike in `DDS`, where $O(n)$ (n is the dimension of x) function evaluations must be performed in each iteration in the worst case, our method only requires two new function evaluations per iteration. Consequently, while `DDS` depends quadratically on n, our method depends linearly on n. In particular, in the nonconvex case, `STP` needs $O(n\varepsilon^{-2})$ function evaluations to find a point at which the gradient of the objective function is below ε, in expectation. In the convex case, the complexity is $O(n\varepsilon^{-1})$. In the strongly convex case `STP` converges linearly, meaning that the complexity is $O\left(n \log\left(\varepsilon^{-1}\right)\right)$.

1. Introduction.

In this paper we consider the problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is a given smooth objective function. We assume that we do not have access to the derivatives of f and only have access to a function evaluation oracle. In other words, we assume that we work in the Derivative-Free Optimization (DFO) setting [3]. Optimization problems of this type appear in many industrial applications where usually the objective function is evaluated through a computer simulation process, and therefore derivatives cannot be directly evaluated; e.g., shape optimization in fluid-dynamics problems [1, 10, 16].

Direct search methods of directional type [13, 3] are a popular class of methods for DFO and are among the first algorithms proposed in numerical optimization [15]. These methods are characterized by evaluating the objective function over a number of (typically predetermined and fixed) directions to ensure descent using a sufficiently

*King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France (`elhoucine.bergou@inra.fr`). This author received support from the AgreenSkills+ fellowship programme which has received funding from the EU's Seventh Framework Programme under grant agreement No FP7-609398 (AgreenSkills+ contract).

†Moscow Institute of Physics and Technology (MIPT), Moscow, Russian Federation (`eduard.gorbunov@phystech.edu`).

‡ King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. University of Edinburgh, Edinburgh, United Kingdom. Moscow Institute of Physics and Technology (MIPT), Moscow, Russian Federation (`peter.richtarik@kaust.edu.sa`).

small stepsize. The directions are typically required to form a *positive spanning set* (i.e. a set of vectors whose conic hull is $\mathbb{R}^n$) in order to make sure that each point in $\mathbb{R}^n$ (and hence also the optimal solution) is achievable by a sequence of positive steps from any starting point.

For instance, the *coordinate search* method uses the coordinate (i.e., standard basic) directions, $e_1, e_2, \ldots, e_n$, and their negatives, $-e_1, -e_2, \ldots, -e_n$ as the set of admissible directions. Clearly, $\{\pm e_i, \, : \, i = 1, 2, \ldots, n\}$ forms a positive spanning set.

1.1. Stochastic Three Points method. In this paper, we study a very general *randomized* variant of direct search methods, which we call *Stochastic Three Points* (`STP`).

`STP` depends on two "parameters": a distribution / probability law $\mathcal{D}$ from which we sample directions, and a stepsize selection rule. At iteration k of `STP`, we generate a random direction s_k by sampling from $\mathcal{D}$, and then choose the next iterate via

$$x_{k+1} = \arg\min \left\{ f(x_k + \alpha_k s_k), f(x_k - \alpha_k s_k), f(x_k) \right\},$$

where $\alpha_k > 0$ is an appropriately chosen stepsize. That is, we pick x_{k+1} as the best of the three points $x_k + \alpha_k s_k, x_k - \alpha_k s_k$ and x_k in terms of the function values.

We prove for such a scheme, with several different choices of stepsizes, that the number of iterations sufficient to guarantee that $\min\limits_{k=0,1,\ldots,K} \mathbf{E}\left[\|\nabla f(x_k)\|_{\mathcal{D}}\right] \leq \varepsilon$ is $O(n\varepsilon^{-2})$, where $\|\cdot\|_{\mathcal{D}}$ is a norm dependent on $\mathcal{D}$ which we introduce in Section 3.1 and $\mathbf{E}\left[\cdot\right]$ is the expectation. This complexity is global since no assumption is made on the starting point. If the objective function f is convex, then the number of iterations needed to get x_k such that $\mathbf{E}\left[f(x_k) - f_*\right] \leq \varepsilon$ is $O(n\varepsilon^{-1})$ where f_* is the optimal value of f. If in addition, f is strongly convex, then we have a global linear rate of convergence. This is an improvement on deterministic direct search (`DDS`) where the best known complexity bounds depend quadratically on n and the same way as our scheme in ε [14, 22, 5]. We propose also a parallel version for `STP`.

Despite our approach shares similarities with other randomized algorithmic approaches, the differences are significant. In the sixties a random optimization approach was proposed in [15]. It was proposed to sample a point randomly around the current iterate and move to this new point if it decreases the objective function. This approach was generalized to cover constrained problems in [2]. The theoretical and numerical performances of this approach for nonconvex functions was studied in [7, 20]. More recently, the works in [4] and [9] use random searching directions, and impose a decrease condition to whether accept the step or reject it, like in `DDS`. They update the stepsize by increasing it if the step is accepted and decreasing it otherwise. Our approach is different from these frameworks in the sense that at each iteration we generate a single direction, then we choose the stepsize independently from any decrease condition. In [9], the authors impose to the search direction some probabilistic property. In fact, they assume that at each iteration their random directions are probabilistic descent conditioned to the past. In other words, at a given iteration, independently from the past with a certain probability at least one of the directions is of descent type. The main result of [9] is the complexity bound $O(rn\varepsilon^{-2})$ to drive the gradient norm below ε with high probability, where $r \geq 2$ is the number of the random directions at each iteration. Also [9] do not cover the cases when the objective function is convex or strongly convex. `STP` method gives similar complexity bound for non-convex problems (with $r = 2$).

More related to our work is the method proposed in [11, 12] for convex problems,

where at iteration k the step is updated as follows

$$x_{k+1} = x_k + \alpha_k u,$$

where u is sampled uniformly from the uniform distribution on the unit sphere, and

$$\alpha_k = \arg\min_{\alpha \in \mathbb{R}} f(x_k + \alpha u).$$

The latter method was improved in two ways by [21]. In fact, the proposed method in [21] i) allows approximate line search, i.e., $\alpha_k \approx \arg\min_{\alpha\in\mathbb{R}} f(x_k + \alpha u)$, ii) and allows discrete sampling from $\{\pm e_i, i = 1, \ldots, n\}$ instead of sampling from the unit sphere. Our approach is different from these methods in the sense that it did not perform any line search approximation to compute the stepsizes, and allows different distributions (which include the uniform distribution over the unit sphere and the discrete sampling from the canonical basis of $\mathbb{R}^n$) to sample the directions. The complexity bounds given in these works are worse than those obtained in this paper. Another method related to our work is the method discussed in [19, Section 3.4], a derivative-free approach based on forming an unbiased estimate of the gradient using Gaussian smoothing. The search direction in this method is distributed uniformly over the unit sphere and it is pre-multiplied by an approximation to the directional derivative along the direction itself. More precisely, this method updates the step at iteration k as follows

$$x_{k+1} = x_k - \alpha_k \frac{f(x_k + \mu_k u) - f(x_k)}{\mu_k} u, \tag{1.2}$$

where $\mu_k \in (0, 1)$ is the finite differences parameter, α_k is the stepsize, and u is a random vector distributed uniformly over the unit sphere. In this work, there is no explicit rules for choosing the parameters and there is no analysis of the worst case complexity. The paper [18] proposes other variants of this method by changing the way of approximating the directional derivative of f along u. Moreover, it gives the worst case complexity analysis of the method (1.2). The complexity bounds in [18] are similar to those of our `STP` approach. Our approach is different from the method (1.2) and its variants porposed in [18], in our approach the search direction can follows a different distribution from the uniform distribution over the unit sphere. For instance, we allow a distribution that has all its mass concentrated on a discrete set of vectors – which makes a direct connection with the (deterministic) direct search methods. Moreover, the proposed stepsizes in [18] depend on the Lipschitz constant of the gradient of the objective function. However, in our approach we proposed some stepsizes which can be easily computed in practice. The extention of the work [18] for an uncontrained problem of minimization of a smooth convex function which is only available through noisy observations of its values were studied in the recent work [8], where the authors proposed accelerated and non-accelerated zeroth-order method, which works in different proximal-setups. They obtained almost dimension-independent rate for the non-accelerated algorithm for the case of ℓ_1-proximal-setup and sparse vector $x_0 - x_*$.

1.2. Outline. We organize this paper as follows. In Section 3 we present our stochastic three points method and give some of its properties. In Section 3.1 we give the main assumptions on the random direction to ensure the convergence of our method. Then, in Section 3.2 we introduce the key lemma for the analysis of the

complexity. Section 4 gives the analysis for the worst case complexity for non-convex problems. While Section 5 deals with the complexity analysis for the convex problems, and Section 6 gives the analysis of the complexity for strongly convex problems. Section 7 proposes a parallel version of STP and gives the corresponding complexity analysis. Numerical tests are illustrated and discussed in Section 8. Conclusions and future improvements are discussed in Section 9.

1.3. Notation. Throughout this paper $\mathcal{D}$ will denote a probability distribution over $\mathbb{R}^n$. We use $\mathbf{E}[\cdot]$ to denote the expectation and $\langle x, y\rangle = x^\top y$ corresponds to the inner product of x and y. We denote also by $\|\cdot\|_2$ the ℓ_2-norm, and by $\|\cdot\|_{\mathcal{D}}$ a norm dependent on $\mathcal{D}$ which we introduce in Section 3.1.

2. Summary of contributions. Here we highlight some of the contributions of this work.

A simple and flexible algorithm. We study a novel variant of direct search based on random directions, which we call Stochastic Three Points (STP). It depends on at most three parameters: The starting point x_0 for the iterate, the probability distribution $\mathcal{D}$ on $\mathbb{R}^n$ to sample the directions, and in some cases an α_0 to define the stepsize. The probability distribution $\mathcal{D}$ may be iteration dependent as far as it satisfies the required assumption (see Assumption 5.4). In fact, Assumption 5.4 may be weakened by letting the probability distribution to depend on the iteration k in the following way

1. The quantity $\gamma_{\mathcal{D}_k} \stackrel{\text{def}}{=} \mathbf{E}_{s\sim\mathcal{D}_k}\|s\|_2^2$ is positive and uniformly bounded away from infinite.
2. There is a constant $\mu_{\mathcal{D}} > 0$ and norm $\|\cdot\|_{\mathcal{D}}$ (independent from k) on $\mathbb{R}^n$ such that

$$\mathbf{E}_{s\sim\mathcal{D}_k}\,|\,\langle g_k, s\rangle\,| \geq \mu_{\mathcal{D}}\|g_k\|_{\mathcal{D}}, \tag{2.1}$$

where $g_k = \nabla f(x_k)$.

This assumption may be weakened even more by letting $\mu_{\mathcal{D}}$ and norm $\|\cdot\|_{\mathcal{D}}$ to dependent on k and assuming i) the uniform boundness of $\mu_{\mathcal{D}_k}$ away from zero, ii) and that $\|\cdot\|_{\mathcal{D}_k}$ is uniformly equivalent to a norm independent from k. To avoid unnecessary notations and for the sake of clarity and simplicity of the presentation, for the analysis we choose the probability distribution to be iteration independent in this paper.

A general setting. Our approach covers some rather exotic first order methods:

- Normalized Gradient Descent (NGD) method: at iteration k, $s \sim \mathcal{D}$ means that $s = \frac{g_k}{\|g_k\|_2}$ with probability 1.
- Signed Gradient Descent (SignGD) method: at iteration k, $s \sim \mathcal{D}$ means that $s = sign\,(g_k)$ with probability 1, where the $sign$ operation is element wise sign.
- Normalized Randomized Coordinate Descent (NRCD) method (equivalently this method can be called also Randomized Signed Gradient Descent): at iteration k, $s \sim \mathcal{D}$ means that $s = \frac{g_k^i}{|g_k^i|}e_i$ if $g_k^i \neq 0$ and $s = 0$ otherwise, with probability $\frac{1}{n}$, where g_k^i is the $i-th$ component of g_k.
- Normalized Stochastic Gradient Descent (NSGD) method: at iteration k, $s \sim \mathcal{D}$ means that $s = \hat{g}_k$ where $\hat{g}_k$ is the stochastic gradient satisfying $\mathbf{E}[\hat{g}_k] = \frac{g_k}{\|g_k\|_2}$, and $\mathbf{E}\left[\|\hat{g}_k\|_2^2\right] \leq \sigma < \infty$.

The required assumption on $\mathcal{D}$ is satisfied in these cases (see Appendix B).

The probability distribution is also allowed to be either continuous or discrete, so that we cover many known strategies of choosing the directions in the DFO setting in the literature. For instance, if $\mathcal{D}$ is the uniform law on the unit sphere we recover the directions proposed in [11, 12, 19, 18]. If it is the discrete law on $\{\pm e_i, i = 1, \ldots, n\}$ we recover the directions proposed in [21]. If it is the discrete law on $\{\pm d_i, i = 1, \ldots, n\}$ where $d_i,\ i = 1, \ldots, n$ form a basis of $\mathbb{R}^n$, `STP` can be seen as a random variant of the Simplified Direct Search (SDS) method studied in [14].

One of the main goals of flexibility in choosing the probability distribution $\mathcal{D}$ is the efficiency for solving some optimization problems which may have some specific properties like:

- The size of the problem to optimize is very large such that even the addition of two vectors may be unfeasible. For instance if the dimension of the problem (i.e., the size of x) is larger than the available memory, then updating all the components of x at each iteration is impossible. One is allowed to update only some components of x at each iteration.
- The objective function is not entirely defined at the beginning of the optimization process, like in the streaming optimization. In other words the data describing the objective function arrives in real time during the optimization process. At a given iteration (time) we can not evaluate the objective function in all points of $\mathbb{R}^n$. We can only evaluate the objective function in a set of directions (only some components of x can be updated).
- Even if we have the entire objective function at the beginning of the optimization process, for some problems the computation of the function value increases with the number of the perturbed variables. In other words, when perturbing all the components of x the evaluation of f takes a lot of time. However by perturbing only one parameter (or a set of parameters) the objective is evaluated in reasonable time.
- Some prior knowledge about Lipschitz constants in some directions is available.

For these kind of situations the choices of $\mathcal{D}$ to be a continuous law is prohibited. However the discrete choices of $\mathcal{D}$ are the most convenient in these cases.

Practicality. `STP` method is extremely simple to use in practice and its analysis is also simple compared to the state-of-the-art direct search methods based on random directions/stepsizes. In fact, the most related work to `STP` is the work in [18]. In the latter work, the proposed stepsizes depend on the Lipschitz constant of the gradient of the objective function, which may not be known in practice. However, for `STP` we proposed several stepsize selection schemes. Some of them can be easily computed in practice. Moreover, our preliminary numerical experiments show that our approach is competitive in practice.

Better bounds. We obtained compact worst case complexity bounds. These bounds are similar to those obtained in [18]. They depend linearly on the dimension of the considered problem, while this dependence is quadratic for deterministic direct search methods [22, 5, 14]. In Table 1 we summarize selected complexity results (bounds on the number of function evaluations) obtained in this paper for `STP` method. In all cases we assume that f is differentiable, bounded below (by f_*), with L-Lipschitz gradient. The assumptions listed in the first column of the table are additional to this. The quantity R_0 measures the size of a specific level set of f. The symbol $\propto$ means proportional. In fact, this symbol appears in the definition of the stepsizes, for instance $\alpha_k \propto \frac{1}{\sqrt{k+1}}$ means that α_k is equal to some constant α_0 (independent

from k) multiplied by $\frac{1}{\sqrt{k+1}}$. This constant α_0 usually depends in the constants of the problem, like the Lipschitz constant and x_0. More details about the definitions of all these quantities are given in the main text.

Assumptions on f (additional to L-smoothness)	Stepsizes	Complexity	Theorems
none	$\alpha_k \propto \frac{1}{\sqrt{k+1}}$ $\alpha_k \propto \varepsilon$	$O\left(\frac{n}{\varepsilon^2}\right)$	4.1, 4.2
convex, R_0 finite	$\alpha_k \propto f(x_k) - f(x_*)$ $\alpha_k \propto \frac{\lvert f(x_k+ts_k)-f(x_k)\rvert}{t}$	$O\left(\frac{n}{\varepsilon}\right)$	5.3, 5.5
λ-strongly convex	$\alpha_k \propto (f(x_k) - f(x_*))^{\frac{1}{2}}$ $\alpha_k \propto \frac{\lvert f(x_k+ts_k)-f(x_k)\rvert}{t}$	$O\left(n \log\left(\frac{1}{\varepsilon}\right)\right)$	6.2 , 6.3

TABLE 1

Summary of the complexity results obtained in this paper for `STP` *method. Column "Complexity" defines the number of iterations needed to guarantee* $\min_k \mathbf{E}\left[\|\nabla f(x_k)\|_{\mathcal{D}}\right] \leq \varepsilon$ *(second row) or* $\mathbf{E}\left[f(x_k) - f(x_*)\right] \leq \varepsilon$ *(third and fourth rows).*

Parallel method. In Table 2 we summarize selected complexity results (bounds on the number of function evaluations) obtained in this paper for the parallel version of the `STP` method. More details about the definitions of all quantities appearing in the table are given in the main text. `PSTP` method gives the same rate as `STP` method with spherical setup but for wider range of distributions.

Assumptions on f (additional to L-smoothness)	Stepsizes	Complexity	Theorems
none	$\alpha_k \propto \frac{1}{\sqrt{k+1}}$	$O\left(\frac{n}{\varepsilon^2}\right)$	7.2
convex, R_0 finite	$\alpha_k \propto f(x_k) - f(x_*)$	$O\left(\frac{n}{\varepsilon}\right)$	7.3
λ-strongly convex	$\alpha_k \propto (f(x_k) - f(x_*))^{\frac{1}{2}}$	$O\left(n \log\left(\frac{1}{\varepsilon}\right)\right)$	7.4

TABLE 2

Summary of the complexity results obtained in this paper for the parallel version of `STP` *method. As before, column "Complexity" defines the number of iterations needed to guarantee* $\min_k \mathbf{E}\left[\|\nabla f(x_k)\|_{\mathcal{D}}\right] \leq \varepsilon$ *(second row) or* $\mathbf{E}\left[f(x_k) - f(x_*)\right] \leq \varepsilon$ *(third and fourth rows).*

Experiments. We provide a number of experimental results, showing that our approach is a competitive algorithm in practice. In fact, we compared on a large set of problems our approach with the method (1.2) as well as with the coordinate search method (the `DDS` method which uses the $2n$ coordinate directions). The experiments show that the use of the random directions leads to a significant improvement in

terms of the number of function evaluation. Indeed, our approach and method (1.2) outperform the DDS method. Moreover, our approach exhibits better performances than the other two methods. See Section 8 for a complete view on the experimental results.

3. Stochastic Three Points method. Our *stochastic three points* (STP) algorithm is formalized below as Algorithm 3.1.

Algorithm 3.1 Stochastic Three Points (STP)

Initialization

Choose $x_0 \in \mathbb{R}^n$, stepsizes $\alpha_k > 0$, probability distribution $\mathcal{D}$ on $\mathbb{R}^n$.

For $k = 0, 1, 2, \ldots$

1. Generate a random vector $s_k \sim \mathcal{D}$
2. Let $x_+ = x_k + \alpha_k s_k$ and $x_- = x_k - \alpha_k s_k$
3. $x_{k+1} = \arg\min\{f(x_-), f(x_+), f(x_k)\}$

Due to the randomness of the search directions s_k for $k \geq 0$, the iterates are also random vectors for all $k \geq 1$. The starting point x_0 is not random (the initial objective function value $f(x_0)$ is deterministic). Note that STP never moves to a point with a larger objective value. This monotonicity property does not depend on $\mathcal{D}$ or the properties of f. Let us formulate this simple observation as a lemma.

LEMMA 3.1 (Monotonicity). *STP produces a monotonic sequence of iterates, i.e., $f(x_{k+1}) \leq f(x_k)$ for all $k \geq 0$. As a consequence,*

$$\mathbf{E}[f(x_{k+1}) \mid x_k] \leq f(x_k). \tag{3.1}$$

Throughout the paper, we assume that f is differentiable, bounded below and has L-Lipschitz gradient.

ASSUMPTION 3.2. *The objective function f is L-smooth with $L > 0$ and bounded from below by $f_* \in \mathbb{R}$. That is, f has a Lipschitz continuous gradient with a Lipschitz constant L:*

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2, \qquad \forall x, y \in \mathbb{R}^n$$

and $f(x) \geq f_*$ for all $x \in \mathbb{R}^n$.

3.1. Random Search Directions. Our analysis in the sequel of the paper will be based on the following key assumption.

ASSUMPTION 3.3. *The probability distribution $\mathcal{D}$ on $\mathbb{R}^n$ has the following properties:*

1. *The quantity $\gamma_{\mathcal{D}} \overset{def}{=} \mathbf{E}_{s\sim\mathcal{D}} \|s\|_2^2$ is positive and finite.*
2. *There is a constant $\mu_{\mathcal{D}} > 0$ and norm $\|\cdot\|_{\mathcal{D}}$ on $\mathbb{R}^n$ such for all $g \in \mathbb{R}^n$,*

$$\mathbf{E}_{s\sim\mathcal{D}} \mid \langle g, s\rangle \mid \geq \mu_{\mathcal{D}}\|g\|_{\mathcal{D}}. \tag{3.2}$$

Note that since all norms in $\mathbb{R}^n$ are equivalent, the second part of the above assumption is satisfied if and only if

$$\inf_{\|g\|_2=1} \mathbf{E}_{s\sim\mathcal{D}} \mid \langle g, s\rangle \mid > 0.$$

However, as the next lemma illustrates, it will be convenient to work with norms that are allowed to depend on $\mathcal{D}$. We now give some examples of distributions for which the above assumption is satisfied.

LEMMA 3.4. *Let $g \in \mathbb{R}^n$.*

1. *If $\mathcal{D}$ is the uniform distribution on the unit sphere in $\mathbb{R}^n$, then*

$$\gamma_{\mathcal{D}} = 1 \quad \text{and} \quad \mathbf{E}_{s\sim\mathcal{D}} \left| \langle g, s \rangle \right| \sim \frac{1}{\sqrt{2\pi n}} \|g\|_2. \tag{3.3}$$

   Hence, $\mathcal{D}$ satisfies Assumption 3.3 with $\gamma_{\mathcal{D}} = 1$, $\|\cdot\|_{\mathcal{D}} = \|\cdot\|_2$ and $\mu_{\mathcal{D}} \sim \frac{1}{\sqrt{2\pi n}}$.

2. *If $\mathcal{D}$ is the normal distribution with zero mean and identity over n as covariance matrix. i.e., $s \sim N(0, \frac{I}{n})$, then*

$$\gamma_{\mathcal{D}} = 1 \quad \text{and} \quad \mathbf{E}_{s\sim\mathcal{D}} \left| \langle g, s \rangle \right| = \frac{\sqrt{2}}{\sqrt{n\pi}} \|g\|_2. \tag{3.4}$$

   Hence, $\mathcal{D}$ satisfies Assumption 3.3 with $\gamma_{\mathcal{D}} = 1$, $\|\cdot\|_{\mathcal{D}} = \|\cdot\|_2$ and $\mu_{\mathcal{D}} = \frac{\sqrt{2}}{\sqrt{n\pi}}$.

3. *If $\mathcal{D}$ is the uniform distribution on $\{e_1, \ldots, e_n\}$, then*

$$\gamma_{\mathcal{D}} = 1 \quad \text{and} \quad \mathbf{E}_{s\sim\mathcal{D}} \left| \langle g, s \rangle \right| = \frac{1}{n} \|g\|_1. \tag{3.5}$$

   Hence, $\mathcal{D}$ satisfies Assumption 3.3 with $\gamma_{\mathcal{D}} = 1$, $\|\cdot\|_{\mathcal{D}} = \|\cdot\|_1$ and $\mu_{\mathcal{D}} = \frac{1}{n}$.

4. *If $\mathcal{D}$ is an arbitrary distribution on $\{e_1, \ldots, e_n\}$ given by $P(s = e_i) = p_i > 0$, then*

$$\gamma_{\mathcal{D}} = 1 \quad \text{and} \quad \mathbf{E}_{s\sim\mathcal{D}} \left| \langle g, s \rangle \right| = \|g\|_{\mathcal{D}} \overset{def}{=} \sum_{i=1}^{n} p_i |g_i|. \tag{3.6}$$

   Hence, $\mathcal{D}$ satisfies Assumption 3.3 with $\gamma_{\mathcal{D}} = 1$ and $\mu_{\mathcal{D}} = 1$.

5. *If $\mathcal{D}$ is a distribution on $D = \{d_1, \ldots, d_n\}$ where $d_1, \ldots, d_n$ form an orthonormal basis of $\mathbb{R}^n$ and $P(s = d_i) = p_i$, then*

$$\gamma_{\mathcal{D}} = 1 \quad \text{and} \quad \mathbf{E}_{s\sim\mathcal{D}} \left| \langle g, s \rangle \right| = \|g\|_{\mathcal{D}} \overset{def}{=} \sum_{i=1}^{n} p_i |g_i|. \tag{3.7}$$

   Hence, $\mathcal{D}$ satisfies Assumption 3.3 with $\gamma_{\mathcal{D}} = 1$ and $\mu_{\mathcal{D}} = 1$.

Proof. See Appendix A. ∎

Without loss of generality, in the rest of this paper we assume that $\gamma_{\mathcal{D}} = 1$. This can be achieved by considering distribution $\mathcal{D}'$ instead, where $s' \sim \mathcal{D}'$ is obtained by first sampling s' from $\mathcal{D}$ and then either normalizing via i) $s = s'/\|s'\|_2$, or ii) $s = s'/\sqrt{\mathbf{E}_{s'\sim\mathcal{D}}\|s'\|_2^2}$.

3.2. Key Lemma. Now, we establish the key result which will be used to prove the main properties of our Algorithm. Its similar result in the case of deterministic direct search (DDS) methods states that the gradient of the objective function for unsuccessful iterations is bounded by a constant multiplied by the stepsize. See for instance [14, Lemma 10].

LEMMA 3.5. *If Assumptions 3.2 and 3.3 hold, then for all $k \geq 0$,*

$$\mathbf{E}\left[f(x_{k+1}) \mid x_k\right] \leq f(x_k) - \mu_{\mathcal{D}}\alpha_k \|\nabla f(x_k)\|_{\mathcal{D}} + \frac{L}{2}\alpha_k^2, \tag{3.8}$$

and

$$\theta_{k+1} \leq \theta_k - \mu_{\mathcal{D}}\alpha_k g_k + \frac{L}{2}\alpha_k^2, \tag{3.9}$$

where $\theta_k = \mathbf{E}[f(x_k)]$ and $g_k = \mathbf{E}[\|\nabla f(x_k)\|_{\mathcal{D}}]$.

Proof. First we notice that from L-smoothness of f we have

$$\begin{aligned} f(x_k + \alpha_k s_k) &\leq f(x_k) + \langle \nabla f(x_k), \alpha_k s_k \rangle + \tfrac{L}{2}\|\alpha_k s_k\|_2^2 \\ &= f(x_k) + \alpha_k \langle \nabla f(x_k), s_k \rangle + \tfrac{L}{2}\alpha_k^2 \|s_k\|_2^2, \end{aligned}$$

and, similarly, $f(x_k - \alpha_k s_k) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), s_k \rangle + \frac{L}{2}\alpha_k^2\|s_k\|_2^2$. Hence,

$$f(x_{k+1}) \leq \min\{f(x_k + \alpha_k s_k), f(x_k - \alpha_k s_k)\} \leq f(x_k) - \alpha_k |\langle \nabla f(x_k), s_k \rangle| + \tfrac{L}{2}\alpha_k^2\|s_k\|_2^2.$$

To conclude (3.8), we only need to take expectation in the above inequality with respect to $s_k \sim \mathcal{D}$, conditional on x_k, and use inequality (3.2). By taking the expectation in (3.8) we get (3.9). ∎

Note that (3.8) can equivalently be written in the following form:

$$\|\nabla f(x_k)\|_{\mathcal{D}} \leq \frac{1}{\mu_{\mathcal{D}}} \left(\frac{f(x_k) - \mathbf{E}\left[f(x_{k+1}) \mid x_k\right]}{\alpha_k} + \frac{L}{2}\alpha_k \right).$$

This form makes it possible to compare this result with a key result used in the analysis of DDS. Indeed, if we assume that the opposite of the following sufficient *expected* decrease condition holds

$$f(x_k) - \mathbf{E}\left[f(x_{k+1}) \mid x_k\right] \geq c\alpha_k^2, \tag{3.10}$$

for some $c > 0$, then we obtain

$$\|\nabla f(x_k)\|_{\mathcal{D}} \leq \frac{1}{\mu_{\mathcal{D}}} \left(c + \frac{L}{2} \right) \alpha_k. \tag{3.11}$$

In DDS, condition (3.10) is equivalent to the sufficient decrease condition $f(x_k) - f(x_{k+1}) \geq c\alpha_k^2$. If such condition does not hold than the step is declared unsuccessful. The inequality in (3.11) is similar with the result in [14, Lemma 10]. In DDS methods, one can check the sufficient decrease condition, so this drives the analysis and allows for simple stepsize update rules to be implemented. In STP, we typically cannot evaluate $\mathbf{E}[f(x_{k+1}) \mid x_k]$ (we can if $\mathcal{D}$ has all its mass on a discrete set – but in that case we would need to do more work per iteration).

4. Non-convex Problems. In this section, we state our most general complexity result where we do not make any additional assumptions on f, besides smoothness and boundedness (see Assumption 3.2).

THEOREM 4.1 (Decreasing stepsize). *Let Assumptions 3.2 and 3.3 hold. Choose* $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$, *where* $\alpha_0 > 0$. *If*

$$K \geq \frac{2\left(\frac{\sqrt{2}(f(x_0) - f_*)}{\alpha_0} + \frac{L\alpha_0}{2}\right)^2}{\mu_{\mathcal{D}}^2 \varepsilon^2}, \tag{4.1}$$

then $\min_{k=0,1,\ldots,K} \mathbf{E}\left[\|\nabla f(x_k)\|_{\mathcal{D}}\right] \leq \varepsilon$.

Proof. We base the proof on the analysis of the recursion (3.9). In particular, it is useful to write it in the following form:

$$g_k \leq \tfrac{1}{\mu_{\mathcal{D}}} \left(\tfrac{\theta_k - \theta_{k+1}}{\alpha_k} + \tfrac{L}{2}\alpha_k \right) = \tfrac{1}{\mu_{\mathcal{D}}} \left(\tfrac{(\theta_k - \theta_{k+1})\sqrt{k+1}}{\alpha_0} + \tfrac{L\alpha_0}{2\sqrt{k+1}} \right). \tag{4.2}$$

We know from (3.1) and the assumption that f is bounded below that $f_* \leq \theta_{k+1} \leq \theta_k \leq f(x_0)$ for all k. Letting $l = \lfloor K/2 \rfloor$, this implies that

$$\sum_{j=l}^{2l} (\theta_j - \theta_{j+1}) = \theta_l - \theta_{2l+1} \leq f(x_0) - f_* \stackrel{\text{def}}{=} C,$$

∎

from which we conclude that there must exist $j \in \{l, \ldots, 2l\}$ such that $\theta_j - \theta_{j+1} \leq C/(l+1)$. This implies that

$$\begin{aligned} g_j &\stackrel{(4.2)}{\leq} \frac{1}{\mu_{\mathcal{D}}} \left(\frac{(\theta_j - \theta_{j+1})\sqrt{j+1}}{\alpha_0} + \frac{L\alpha_0}{2\sqrt{j+1}} \right) \leq \frac{1}{\mu_{\mathcal{D}}} \left(\frac{C\sqrt{j+1}}{\alpha_0(l+1)} + \frac{L\alpha_0}{2\sqrt{j+1}} \right) \\ &\leq \frac{1}{\mu_{\mathcal{D}}} \left(\frac{C\sqrt{2l+1}}{\alpha_0(l+1)} + \frac{L\alpha_0}{2\sqrt{l+1}} \right) \leq \frac{1}{\mu_{\mathcal{D}}\sqrt{l+1}} \left(\frac{\sqrt{2}C}{\alpha_0} + \frac{L\alpha_0}{2} \right) \\ &\leq \frac{1}{\mu_{\mathcal{D}}\sqrt{K/2}} \left(\frac{\sqrt{2}C}{\alpha_0} + \frac{L\alpha_0}{2} \right) \stackrel{(4.1)}{\leq} \varepsilon. \end{aligned}$$

Let us now give some insights into the above theorem.

- **Sphere setup.** If $\mathcal{D}$ is the uniform distribution on the Euclidean sphere, then $\mu_{\mathcal{D}} \sim \frac{1}{\sqrt{2\pi n}}$, and hence the above theorem gives a complexity guarantee of the form
$$O\left(\frac{n}{\varepsilon^2}\right).$$
This is an improvement on `DDS` where the best known complexity bound is $O(n^2/\varepsilon^2)$ [22, 14]. The same conclusion holds for the normal distribution setup.
- **Coordinate setup.** If $\mathcal{D}$ is the uniform distribution on $\{e_1, \ldots, e_n\}$, then $\mu_{\mathcal{D}} = 1/n$ and hence the bound is of the form
$$O\left(\frac{n^2}{\varepsilon^2}\right).$$
However, this is for the ℓ_1 norm of the gradient of f, which is *larger* than the ℓ_2 norm. Indeed, for all x we have $\sqrt{n}\|\nabla f(x)\|_2 \geq \|\nabla f(x)\|_1 \geq \|\nabla f(x)\|_2$, and the first inequality can be tight (for the vector of all ones, for instance). Hence, if we are interested to achieve $\|\nabla f(x)\|_2 \leq \varepsilon'$, in certain situations it may be sufficient to push the ℓ_1 norm of the gradient below $\varepsilon = \sqrt{n}\varepsilon'$ instead. So, the iteration bound can be as good as
$$O\left(\frac{n^2}{(\sqrt{n}\varepsilon')^2}\right) = O\left(\frac{n}{(\varepsilon')^2}\right).$$
- **Quality of the final iterate.** Theorem 4.1 does not guarantee the gradient of f at the *final* point x_K to be small (in expectation). Instead, it guarantees that the gradient of f at *some* point produced by the method will be small. Notice however, that the method is monotonic. Hence, all subsequent points produced by the method will have better functions values than the one which has gradient of minimum norm (in expectation). So, we can say that $f(x_K) \leq f(x_j)$ where $\mathbf{E}\left[\|\nabla f(x_j)\|_{\mathcal{D}}\right] \leq \varepsilon$.
- **Optimal stepsize.** Note that the complexity depends on α_0. The optimal choice (minimizing the complexity bound) is
$$\alpha^* = 8^{1/4}\sqrt{\frac{f(x_0) - f_*}{L}},$$

in which case the complexity bound (4.1) takes the form

$$\frac{4\sqrt{2}(f(x_0)-f_*)L}{\mu_{\mathcal{D}}^2\varepsilon^2}. \tag{4.3}$$

Assume that the lower bound f_* is achieved by some point $x_* \in \mathbb{R}^n$. Necessarily, $\nabla f(x_*) = 0$. Moreover, since f is L-smooth, we can write

$$f(x_0) \le f(x_*) + \langle \nabla f(x_*), x_0 - x_* \rangle + \frac{L}{2}\|x_0 - x_*\|_2^2.$$

Hence, the optimal stepsize is no larger than

$$\alpha^* \le 2^{1/4}\|x_0 - x_*\|_2.$$

Of course, we cannot use this optimal stepsize as we usually do not know L and/or f_*. So, we are paying for the lack of knowledge by an increased complexity bound. This makes intuitive sense: the stepsize should not be much larger than the distance of the initial point to an optimal point.

On the other hand, there are examples of non-convex functions for which the ratio $(f(x_0) - f_*)/L$ is arbitrarily small, and the distance between x_0 and x_* arbitrarily high. This cannot happen for convex functions with bounded level sets or for strongly convex functions, as then $f(x) - f(x_*)$ can be lower bounded by quantity proportional to $\|x - x_*\|_2$ with some positive power.

We now state a complexity theorem for `STP` used with a fixed stepsize.

THEOREM 4.2 (Fixed stepsize). *Let f satisfy Assumption 3.2 and also assume that f is bounded below by $f_* \in \mathbb{R}$. Choose a fixed stepsize $\alpha_k = \alpha$ with $0 < \alpha < 2\mu_{\mathcal{D}}\varepsilon/L$. If*

$$K \ge k(\varepsilon) \overset{def}{=} \left\lceil \frac{f(x_0) - f_*}{(\mu_{\mathcal{D}}\varepsilon - \frac{L}{2}\alpha)\alpha} \right\rceil - 1, \tag{4.4}$$

then $\min_{k=0,1,\dots,K} \mathbf{E}\left[\|\nabla f(x_k)\|_{\mathcal{D}}\right] \le \varepsilon$. In particular, if $\alpha = \mu_{\mathcal{D}}\varepsilon/L$, then

$$k(\varepsilon) = \left\lceil \frac{2L(f(x_0) - f_*)}{\mu_{\mathcal{D}}^2\varepsilon^2} \right\rceil - 1.$$

Proof. If $g_k \le \varepsilon$ for some $k \le k(\varepsilon)$, then we are done. Assume hence by contradiction that $g_k > \varepsilon$ for all $k \le k(\varepsilon)$. By taking expectation in Lemma 3.5, we get

$$\theta_{k+1} \le \theta_k - \mu_{\mathcal{D}}\alpha g_k + \tfrac{L}{2}\alpha^2,$$

where $\theta_k = \mathbf{E}[f(x_k)]$ and $g_k = \mathbf{E}[\|\nabla f(x_k)\|_{\mathcal{D}}]$. Hence,

$$f_* \le \theta_{K+1} < \theta_0 - (K+1)\left(\mu_{\mathcal{D}}\alpha\varepsilon - \tfrac{L}{2}\alpha^2\right) \overset{(4.4)}{\le} \theta_0 - (f(x_0) - f_*) = f_*, \qquad \square$$

which is a contradiction.

Here we give some comments about the `STP` for non-convex functions.

- In some situations, when L is not available, it is impossible to compute optimal $\alpha = \frac{\mu_{\mathcal{D}}\varepsilon}{L}$.
- If we can guess α is close to the optimal, then the method depends linearly on n if $1/\mu_{\mathcal{D}}^2 = O(n)$.

- Also, if we guess α right, we get complexity that depends on $L(f(x_0)-f_*)$, which is similar to the setup with variable stepsizes and optimal α_0.
- As before, we only get guarantee on the best of the points in term of the gradient norm, not on the final point.

5. Convex Problems. In this section we estimate the complexity of the STP in the case of convex f. In this case we need an additional technical assumption.

ASSUMPTION 5.1. *We assume that f is convex, has a minimizer x_*, and has bounded level set at x_0:*

$$R_0 \stackrel{def}{=} \max\{\|x-x_*\|_{\mathcal{D}}^* \;:\; f(x)\le f(x_0)\} < +\infty,$$

where $\|\xi\|_{\mathcal{D}}^* \stackrel{def}{=} \max\{\langle \xi, x\rangle \mid \|x\|_{\mathcal{D}}\le 1\}$ defines the dual norm to $\|\cdot\|_{\mathcal{D}}$.

Note that if the above assumption holds, then whenever $f(x)\le f(x_0)$, we get $f(x)-f(x_*)\le \langle \nabla f(x), x-x_*\rangle \le \|\nabla f(x)\|_{\mathcal{D}}\|x-x_*\|_{\mathcal{D}}^* \le R_0\|\nabla f(x)\|_{\mathcal{D}}$. That is,

$$\|\nabla f(x)\|_{\mathcal{D}} \ge \frac{f(x)-f(x_*)}{R_0}. \tag{5.1}$$

Now, we state our main complexity result of this section. We start with the analysis of STP with constant stepsizes.

THEOREM 5.2 (Constant stepsize). *Let Assumptions 3.2, 3.3 and 5.1 be satisfied. Let $0<\varepsilon<\frac{LR_0^2}{\mu_{\mathcal{D}}^2}$ and choose constant stepsize $\alpha_k=\alpha=\frac{\varepsilon\mu_{\mathcal{D}}}{LR_0}$. If*

$$K \ge \frac{LR_0^2}{\mu_{\mathcal{D}}^2\varepsilon}\log\left(\frac{2(f(x_0)-f(x_*))}{\varepsilon}\right), \tag{5.2}$$

then $\mathbf{E}\left[f(x_K)-f(x_*)\right]\le\varepsilon$.

Proof. Let us substitute (5.1) into Lemma 3.5 and take expectations. We get

$$\theta_{k+1}\le \theta_k - \tfrac{\mu_{\mathcal{D}}\alpha}{R_0}(\theta_k - f(x_*)) + \tfrac{L}{2}\alpha^2. \tag{5.3}$$ ∎

Let $r_k=\theta_k-f(x_*)$ and $c=1-\frac{\mu_{\mathcal{D}}\alpha}{R_0}\in(0,1)$. Subtracting $f(x_*)$ from both sides of (5.3), we obtain

$$\begin{aligned} r_K \le cr_{K-1}+\tfrac{L}{2}\alpha^2 \quad &\le \quad c^K r_0 + \tfrac{L}{2}\alpha^2\sum_{i=0}^{K-1}c^i \\ \le \exp(-\mu_{\mathcal{D}}\alpha K/R_0)r_0 + \tfrac{L\alpha^2}{2(1-c)} \quad &= \quad \exp(-\mu_{\mathcal{D}}\alpha K/R_0)r_0+\tfrac{\varepsilon}{2} \stackrel{(5.2)}{\le} \varepsilon. \end{aligned}$$

If $\mu_{\mathcal{D}}\sim\frac{1}{\sqrt{n}}$, then the above theorem gives a complexity guarantee of the form

$$O\left(\frac{n}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)\right).$$

Comparing this to the best known complexity bound for DDS which is $O(\frac{n^2}{\varepsilon})$ [5, 14], we improve the dependence on n but we deteriorate the dependence on ε because of the presence of the term $\log\left(\frac{1}{\varepsilon}\right)$. In the next theorem we show how we get rid of the $\log\frac{1}{\varepsilon}$ term using variable stepsize.

THEOREM 5.3 (Variable stepsize). *Let Assumptions 3.2, 3.3 and 5.1 be satisfied. Let* $\alpha_k = \alpha_0 \left(f(x_k) - f(x_*)\right)$, *where* $0 < \alpha_0 < \frac{2\mu_{\mathcal{D}}}{R_0 L}$. *Define* $a = \frac{\mu_{\mathcal{D}}\alpha_0}{R_0} - \frac{L\alpha_0^2}{2} > 0$. *If* $k \geq k(\varepsilon) \overset{def}{=} \frac{1}{a}\left(\frac{1}{\varepsilon} - \frac{1}{r_0}\right)$, *then* $\mathbf{E}\left[f(x_k) - f(x_*)\right] \leq \varepsilon$.

Proof. Let us substitute (5.1) into equation (3.8) of Lemma 3.5, and then substrate $f(x_*)$ from both sides we get

$$\mathbf{E}\left[f(x_{k+1}) \mid x_k\right] - f(x_*) \leq f(x_k) - f(x_*) - \mu_{\mathcal{D}}\alpha_k \frac{f(x_k) - f(x_*)}{R_0} + \frac{L}{2}\alpha_k^2.$$

Let $r_k = \mathbf{E}\left[f(x_k)\right] - f(x_*)$. By using our choice of α_k in the previous equation and then taking the expectation we get $r_{k+1} \leq r_k - \left(\frac{\mu_{\mathcal{D}}\alpha_0}{R_0} - \frac{L\alpha_0^2}{2}\right) r_k^2 = r_k - a r_k^2$. Therefore,

$$\frac{1}{r_{k+1}} - \frac{1}{r_k} = \frac{r_k - r_{k+1}}{r_k r_{k+1}} \geq \frac{r_k - r_{k+1}}{r_k^2} \geq a.$$

From this we have $\frac{1}{r_k} \geq \frac{1}{r_0} + ka$ and hence $r_k \leq \frac{1}{\frac{1}{r_0} + ka}$. It remains to notice that for $k \geq \frac{1}{a}\left(\frac{1}{\varepsilon} - \frac{1}{r_0}\right)$ we have $r_k \leq \frac{1}{\frac{1}{r_0} + ka} \leq \varepsilon$. ∎

If $\alpha_0 = \frac{\mu_{\mathcal{D}}}{R_0 L}$, then a is maximal as a function of α_0, for which we get the optimal bound

$$k(\varepsilon) = \frac{2R_0^2 L}{\mu_{\mathcal{D}}^2}\left(\frac{1}{\varepsilon} - \frac{1}{r_0}\right).$$

If $\mu_{\mathcal{D}} \sim \frac{1}{\sqrt{n}}$, then the above theorem gives a complexity guarantee of the form $O\left(\frac{n}{\varepsilon}\right)$.

The stepsizes in the previous theorem depend on $f(x_*)$. Of course, in practice we cannot always use these stepsizes as we usually do not know $f(x_*)$. Next theorem gives a more practical stepsizes for which we get the same complexity as in the previous theorem. We start by stating an extra assumption on the probability distribution $\mathcal{D}$ and show that this assumption is satisfied for all the probability distributions given in Lemma 3.4.

ASSUMPTION 5.4. *The probability distribution* $\mathcal{D}$ *on* $\mathbb{R}^n$ *is such that for all* $s \sim \mathcal{D}$ *are of unit Euclidean norm* ($\|s\|_2 = 1$) *with probability 1.*

Let $C_{\mathcal{D}}$ be the positive constant such that for all $x \in \mathbb{R}^n$ the following inequality holds: $\|x\|_2 \leq C_{\mathcal{D}}\|x\|_{\mathcal{D}}$. Such constant exists due to the equivalence of the norms in $\mathbb{R}^n$.

THEOREM 5.5 (Solution-free stepsize). *Let Assumptions 3.2, 3.3, 5.1 and 5.4 be satisfied. Let* $\alpha_k = \frac{|f(x_k + ts_k) - f(x_k)|}{Lt}$, *where*

$$0 < t \leq \frac{\sqrt{2}\mu_{\mathcal{D}}\mathbf{E}\left[f(x_{K-1}) - f_*\right]}{LR_0}.$$

Define $a = \frac{\mu_{\mathcal{D}}^2}{4LR_0^2}$. *If* $K \geq k(\varepsilon) \overset{def}{=} \frac{1}{a}\left(\frac{1}{\varepsilon} - \frac{1}{r_0}\right)$, *then* $\mathbf{E}\left[f(x_K) - f(x_*)\right] \leq \varepsilon$.

Proof. From Lemma 3.5 we have

$$f(x_{k+1}) \leq f(x_k) - \alpha_k |\langle \nabla f(x_k), s_k\rangle| + \frac{L\alpha_k^2}{2}. \tag{5.4}$$

We know that $\alpha_k^{\text{opt}} = \frac{|\langle \nabla f(x_k), s_k\rangle|}{L}$ minimizes the right hand side of (5.4). But it depends on $\nabla f(x_k)$ which we can not compute *exactly*, because we have zeroth-order

oracle. Actually, we do not need to know the whole gradient, it is enough to know directional derivative of f, which we can approximate by finite difference of function values of f. It is the main idea behind our choice of $\alpha_k^{\text{opt}} = \frac{|f(x_k+ts_k)-f(x_k)|}{Lt}$, which does not depends any more on $f(x_*)$ and can be easily computed in practice. We can rewrite $\alpha_k = \frac{|f(x_k+ts_k)-f(x_k)|}{Lt} = \frac{|\langle \nabla f(x_k), s_k\rangle|}{L} + \frac{|f(x_k+ts_k)-f(x_k)|}{Lt} - \frac{|\langle \nabla f(x_k), s_k\rangle|}{L} \overset{\text{def}}{=} \alpha_k^{\text{opt}} + \delta_k$. Therefore, we have

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \tfrac{|\langle \nabla f(x_k), s_k\rangle|^2}{L} - \delta_k |\langle \nabla f(x_k), x_k\rangle| + \tfrac{|\langle \nabla f(x_k), x_k\rangle|^2}{2L} \\ &\quad + \delta_k |\langle \nabla f(x_k), x_k\rangle| + \tfrac{L}{2}(\delta_k)^2 \\ &= f(x_k) - \tfrac{|\langle \nabla f(x_k), s_k\rangle|^2}{2L} + \tfrac{L}{2}(\delta_k)^2 \end{aligned}$$

Next we estimate $|\delta_k|$ using L-smoothness of f:

$$\begin{aligned} |\delta_k| &= \tfrac{1}{Lt} \left| |f(x_k + ts_k) - f(x_k)| - |\langle \nabla f(x_k), ts_k\rangle| \right| \\ &\leq \tfrac{1}{Lt} |f(x_k + ts_k) - f(x_k) - \langle \nabla f(x_k), ts_k\rangle| \leq \tfrac{1}{Lt} \cdot \tfrac{L}{2} \|ts_k\|_2^2 = \tfrac{t}{2}. \end{aligned}$$

From this we obtain

$$f(x_{k+1}) \quad \leq f(x_k) - \tfrac{|\langle \nabla f(x_k), s_k\rangle|^2}{2L} + \tfrac{Lt^2}{8}. \tag{5.5}$$

Taking mathematical expectation w.r.t. all randomness from the previous inequality we get

$$\begin{aligned} \underbrace{\mathbf{E}[f(x_{k+1})] - f_*}_{r_{k+1}} &\overset{①}{\leq} \underbrace{\mathbf{E}[f(x_k)] - f_*}_{r_k} - \tfrac{\mu_{\mathcal{D}}^2}{2L} \mathbf{E}[\|\nabla f(x_k)\|_{\mathcal{D}}^2] + \tfrac{Lt^2}{8} \\ &\overset{②}{\leq} r_k - \tfrac{\mu_{\mathcal{D}}^2}{2LR_0^2} r_k^2 + \tfrac{Lt^2}{8}, \end{aligned} \tag{5.6}$$

where ① is due to tower property of mathematical expectation and (3.2):

$$\begin{aligned} \mathbf{E}[|\langle \nabla f(x_k), s_k\rangle|^2] &= \mathbf{E}\left[\mathbf{E}[|\langle \nabla f(x_k), s_k\rangle|^2 \mid x_k]\right] \geq \mathbf{E}\left[\left(\mathbf{E}[|\langle \nabla f(x_k), s_k\rangle| \mid x_k]\right)^2\right] \\ &\overset{(3.2)}{\geq} \mu_{\mathcal{D}}^2 \mathbf{E}[\|\nabla f(x_k)\|_{\mathcal{D}}^2]; \end{aligned}$$

② follows from Assumption 5.1: $\mathbf{E}[\|\nabla f(x_k)\|_{\mathcal{D}}^2] \geq \frac{\mathbf{E}\left[(f(x_k)-f_*)^2\right]}{R_0^2} \geq \frac{(\mathbf{E}[f(x_k)-f_*])^2}{R_0^2} = \frac{r_k^2}{R_0^2}$. From this and monotonicity of $\{f(x_k)\}_{k\geq 0}$ we have

$$\tfrac{1}{r_{k+1}} - \tfrac{1}{r_k} \geq \tfrac{r_{k+1}-r_k}{r_k r_{k+1}} \geq \tfrac{\frac{\mu_{\mathcal{D}}^2}{2LR_0^2} r_k^2 - \frac{Lt^2}{8}}{r_k^2} \geq \tfrac{\mu_{\mathcal{D}}^2}{2LR_0^2} - \tfrac{L}{8}\left(\tfrac{t}{r_k}\right)^2. \tag{5.7}$$

If $k \leq K-1$ and $0 < t \leq \frac{\sqrt{2}\mu_{\mathcal{D}} r_{K-1}}{LR_0}$, then we can write

$$\tfrac{1}{r_{k+1}} - \tfrac{1}{r_k} \geq \tfrac{\mu_{\mathcal{D}}^2}{4LR_0^2} = a,$$

since $r_k \leq r_{K-1}$. Finally, we have $\frac{1}{r_k} \geq \frac{1}{r_0} + ka$ and hence $r_k \leq \frac{1}{\frac{1}{r_0}+ka}$. for all $k \leq K$. Thus, if $K \geq \frac{1}{a}\left(\frac{1}{\varepsilon} - \frac{1}{r_0}\right)$, then $r_K \leq \frac{1}{\frac{1}{r_0}+Ka} \leq \varepsilon$. ∎

Actually, requirement $t \leq \frac{\sqrt{2}\mu_{\mathcal{D}}\mathbf{E}[f(x_{K-1})-f_*]}{LR_0}$ could be replaced by $t \leq \frac{\sqrt{2}\mu_{\mathcal{D}}\varepsilon}{LR_0}$ if we additionally require that for all $k \leq K$ we have $r_k \geq \varepsilon$.

6. Strongly Convex Problems. In this section we derive the complexity of the STP method in the case of strongly convex f.

ASSUMPTION 6.1. *f is λ-strongly convex with respect to the norm $\|\cdot\|_{\mathcal{D}}$.*

In this section, we denote by x_* the unique minimizer of f.

THEOREM 6.2. *Let Assumptions 3.2, 3.3 and 6.1 be satisfied. Let stepsize $\alpha_k = \frac{\theta_k \mu_{\mathcal{D}}}{L}\sqrt{2\lambda(f(x_k)-f(x_*))}$, for some $\theta_k \in (0,2)$ such that $\theta \overset{def}{=} \inf_k 2\theta_k - \theta_k^2 > 0$. If*

$$K \geq \frac{L}{\lambda \mu_{\mathcal{D}}^2 \theta} \log\left(\frac{f(x_0)-f(x_*)}{\varepsilon}\right), \tag{6.1}$$

then $\mathbf{E}\left[f(x_K)-f(x_*)\right] \leq \varepsilon$.

Proof. By injecting α_k into equation (3.8) of Lemma 3.5, and then substrate $f(x_*)$ from both sides we get

$$\begin{aligned}\mathbf{E}[f(x_{k+1}) \mid x_k] - f(x_*) \leq f(x_k) - f(x_*) - \frac{\mu_{\mathcal{D}}^2 \theta_k \sqrt{2\lambda(f(x_k)-f(x_*))}\|\nabla f(x_k)\|_{\mathcal{D}}}{L} \\ + \frac{\mu_{\mathcal{D}}^2 \theta_k^2 \lambda (f(x_k)-f(x_*))}{L}.\end{aligned}$$

From the strong convexity property of f we have $\|\nabla f(x_k)\|_{\mathcal{D}}^2 \geq 2\lambda(f(x_k)-f(x_*))$, therefore

$$\begin{aligned}\mathbf{E}[f(x_{k+1})|x_k] - f(x_*) &\leq f(x_k) - f(x_*) - \frac{2\mu_{\mathcal{D}}^2 \theta_k \lambda (f(x_k)-f(x_*))}{L} + \frac{\mu_{\mathcal{D}}^2 \theta_k^2 \lambda (f(x_k)-f(x_*))}{L} \\ &\leq f(x_k) - f(x_*) - \frac{\mu_{\mathcal{D}}^2 \lambda (f(x_k)-f(x_*))}{L}(2\theta_k - \theta_k^2) \\ &\leq f(x_k) - f(x_*) - \frac{\mu_{\mathcal{D}}^2 \theta \lambda (f(x_k)-f(x_*))}{L},\end{aligned}$$

where we used the definition of θ. Let $r_k = \mathbf{E}\left[f(x_k)\right] - f(x_*)$. By taking the expectation of the last inequality we get $r_{k+1} \leq \left(1 - \frac{\mu_{\mathcal{D}}^2 \theta \lambda}{L}\right) r_k$, and therefore

$$r_k \leq \left(1 - \frac{\mu_{\mathcal{D}}^2 \theta \lambda}{L}\right)^k r_0.$$

Hence if K satisfies (6.1), we get $r_K \leq \varepsilon$. ∎

From this theorem we conclude that if there exist $0 < \theta_1 \leq \theta_2 < 2$ such that

$$\frac{\theta_1 \mu_{\mathcal{D}}}{L}\sqrt{2\lambda(f(x_k)-f(x_*))} \leq \alpha_k \leq \frac{\theta_2 \mu_{\mathcal{D}}}{L}\sqrt{2\lambda(f(x_k)-f(x_*))},$$

then the sequence $(r_k)_k$ converges linearly to zero.

The stepsizes from the previous theorem depend on $f(x_*)$. In practice, we cannot always use these stepsizes as we usually do not know $f(x_*)$. Next theorem gives the similar result for STP with stepsizes independent from $f(x_*)$ under additional assumptions that for all $s \sim \mathcal{D}$ we have $\|s\|_2 = 1$ with probability 1.

THEOREM 6.3. *Let Assumptions 3.2, 3.3, 5.4 and 6.1 be satisfied. Let $\alpha_k = \frac{|f(x_k+ts_k)-f(x_k)|}{Lt}$, for $1 < t \leq \frac{2\mu_{\mathcal{D}}\sqrt{\lambda\varepsilon}}{L}$. If*

$$K \geq \frac{L}{\lambda \mu_{\mathcal{D}}^2} \log\left(\frac{2(f(x_0)-f(x_*))}{\varepsilon}\right), \tag{6.2}$$

then $\mathbf{E}\left[f(x_K)-f(x_*)\right] \leq \varepsilon$.

Proof. From (5.5) we have $f(x_{k+1}) \leq f(x_k) - \frac{|\langle \nabla f(x_k), s_k\rangle|^2}{2L} + \frac{Lt^2}{8}$. Taking mathematical expectation w.r.t. all randomness from the previous inequality we get

$$\begin{aligned} \underbrace{\mathbf{E}[f(x_{k+1})] - f_*}_{r_{k+1}} &\overset{①}{\leq} \underbrace{\mathbf{E}[f(x_k)] - f_*}_{r_k} - \tfrac{\mu_{\mathcal{D}}^2}{2L}\mathbf{E}[\|\nabla f(x_k)\|_{\mathcal{D}}^2] + \tfrac{Lt^2}{8} \\ &\overset{②}{\leq} \left(1 - \tfrac{\mu_{\mathcal{D}}^2 \lambda}{L}\right) r_k + \tfrac{Lt^2}{8}, \end{aligned} \tag{6.3}$$

where ① is due to tower property of mathematical expectation and (3.2):

$$\begin{aligned} \mathbf{E}[|\langle \nabla f(x_k), s_k\rangle|^2] &= \mathbf{E}\left[\mathbf{E}[|\langle \nabla f(x_k), s_k\rangle|^2 \mid x_k]\right] \geq \mathbf{E}\left[\left(\mathbf{E}[|\langle \nabla f(x_k), s_k\rangle| \mid x_k]\right)^2\right] \\ &\overset{(3.2)}{\geq} \mu_{\mathcal{D}}^2 \mathbf{E}[\|\nabla f(x_k)\|_{\mathcal{D}}^2]; \end{aligned}$$

② follows from λ-strong convexity of f: $\|\nabla f(x_k)\|_{\mathcal{D}}^2 \geq 2\lambda\left(f(x_k) - f_*\right)$. From (6.3) we have

$$\begin{aligned} r_{k+1} &\leq \left(1 - \tfrac{\mu_{\mathcal{D}}^2 \lambda}{L}\right)^{k+1} r_0 + \tfrac{Lt^2}{8} \sum_{i=0}^{k} \left(1 - \tfrac{\mu_{\mathcal{D}}^2 \lambda}{L}\right)^i \\ &\leq \left(1 - \tfrac{\mu_{\mathcal{D}}^2 \lambda}{L}\right)^{k+1} r_0 + \tfrac{L^2 t^2}{8\mu_{\mathcal{D}}^2 \lambda}. \end{aligned} \tag{6.4}$$

Hence if $t \leq \frac{2\mu_{\mathcal{D}}\sqrt{\lambda\varepsilon}}{L}$ and K satisfies (6.2) we get $r_K \leq \varepsilon$. ∎

7. Parallel Stochastic Three Points Method. Consider the parallel version of `STP` proposed in Algorithm 7.1.

Algorithm 7.1 *Parallel Stochastic Three Points* (`PSTP`)

Initialization

Choose $x_0 \in \mathbb{R}^n$, stepsizes $\alpha_k > 0$, parallelism parameter τ, differentiation stepsize t_0.

For $k = 0, 1, 2, \ldots$

1. For $i = 1, 2, \ldots, \tau$. Generate a random vector $s_{ki} \sim \mathcal{D}$.
2. Let $s_k = \frac{1}{\tau}\sum_{i=1}^{\tau} s_{ki}$.
3. Let $x_+ = x_k + \alpha_k s_k$ and $x_- = x_k - \alpha_k s_k$.
4. $x_{k+1} = \arg\min\{f(x_-), f(x_+), f(x_k)\}$.

We start our analysis of the complexity in this section by stating an extra assumption on the probability distribution $\mathcal{D}$ which is satisfied for all the probability distributions given in Lemma 3.4.

ASSUMPTION 7.1. *The probability distribution $\mathcal{D}$ on $\mathbb{R}^n$ satisfies the following properties.*

1. *If $s_1, s_2 \sim \mathcal{D}$ are independent, then $\mathbf{E}[\langle s_1, s_2\rangle] = 0$.*
2. *There exist a constant $\tilde{\mu}_{\mathcal{D}} > 0$ and $\tau > 0$ such that if $s_1, s_2, \ldots, s_\tau \sim \mathcal{D}$ are independent and for all $g \in \mathbb{R}^n$*

$$\mathbf{E}\left[\left|\left\langle g, \frac{1}{\tau}\sum_{i=1}^{\tau} s_i\right\rangle\right|\right] \geq \frac{\tilde{\mu}_{\mathcal{D}}}{\sqrt{\tau}}\|g\|_2.$$

If the first part of the assumption does not hold for distribution $\mathcal{D}$ we can consider distribution $\bar{\mathcal{D}}$ such that $\mathbf{E}_{s\sim\mathcal{D}}[s] = 0$ by adding opposite vector for each vector from $\mathcal{D}$ and share the probability measure between opposite vectors in equal ratio. The second part of the assumption holds due to Central Limit Theorem for wide range of distributions (this range covers the examples in Lemma 3.4) and due to the second part of Lemma 3.4 we can say that for big enough τ we have $\tilde{\mu}_{\mathcal{D}} \sim \sqrt{\frac{2}{\pi n}}$ in the case when $\gamma_{\mathcal{D}} = 1$.

In the next three subsections we will give the adaptation of the main complexity results obtained for `STP` for `PSTP`.

7.1. Non-convex Case. The following theorem is the adaptation of Theorem 4.1.

THEOREM 7.2. *Let Assumptions 3.2, 3.3, 5.4 and 7.1 hold. Choose* $\alpha_k = \frac{\alpha_0}{\sqrt{k+1}}$, *where* $\alpha_0 > 0$. *If*

$$K \geq \frac{2\left(\frac{\sqrt{2\tau}(f(x_0)-f_*)}{\alpha_0} + \frac{L\alpha_0}{2\sqrt{\tau}}\right)^2}{\tilde{\mu}_{\mathcal{D}}^2\varepsilon^2}, \tag{7.1}$$

then

$$\min_{k=0,1,\ldots,K} \mathbf{E}\left[\|\nabla f(x_k)\|_2\right] \leq \varepsilon.$$

Proof. By definition of x_+ and x_- we have

$$x_+ = x_k + \tfrac{\alpha_k}{\tau}\sum_{i=1}^{\tau} s_{ki}, \qquad x_- = x_k - \tfrac{\alpha_k}{\tau}\sum_{i=1}^{\tau} s_{ki},$$

whence

$$\begin{aligned} f(x_+) &\leq f(x_k) - \alpha_k\langle\nabla f(x_k), \tfrac{1}{\tau}\sum_{i=1}^{\tau} s_{ki}\rangle + \tfrac{L\alpha_k^2}{2\tau^2}\left\|\sum_{i=1}^{\tau} s_{ki}\right\|_2^2, \\ f(x_-) &\leq f(x_k) + \alpha_k\langle\nabla f(x_k), \tfrac{1}{\tau}\sum_{i=1}^{\tau} s_{ki}\rangle + \tfrac{L\alpha_k^2}{2\tau^2}\left\|\sum_{i=1}^{\tau} s_{ki}\right\|_2^2. \end{aligned}$$

Therefore

$$f(x_{k+1}) \leq \min\{f(x_+), f(x_-)\} \leq f(x_k) - \alpha_k\left|\left\langle\nabla f(x_k), \tfrac{1}{\tau}\sum_{i=1}^{\tau} s_{ki}\right\rangle\right| + \tfrac{L\alpha_k^2}{2\tau^2}\left\|\sum_{i=1}^{\tau} s_{ki}\right\|_2^2.$$

Taking conditional mathematical expectation $\mathbf{E}[\,\cdot \mid x_k]$ from the both sides of previous inequality we have

$$\mathbf{E}[f(x_{k+1}) \mid x_k] - f_* \overset{①}{\leq} f(x_k) - f_* - \tfrac{\alpha_k\tilde{\mu}_{\mathcal{D}}}{\sqrt{\tau}}\|\nabla f(x_k)\|_2 + \tfrac{L\alpha_k^2}{2\tau} \tag{7.2}$$

where ① is due to the first part of Assumption 7.1 and Assumption 3.3:

$$\mathbf{E}\left[\left\|\sum_{i=1}^{\tau} s_{ki}\right\|_2^2\right] = \sum_{i=1}^{\tau}\mathbf{E}[\|s_{ki}\|_2^2] + \sum_{i\neq j=1}^{\tau}\mathbf{E}[\langle s_{ki}, s_{kj}\rangle] = \tau.$$

Taking full expectation from the both sides of the inequality (7.2) and rearranging the terms we obtain

$$g_k \leq \frac{\sqrt{\tau}}{\tilde{\mu}_{\mathcal{D}}}\left(\frac{\theta_k-\theta_{k+1}}{\alpha_k}+\frac{L}{2\tau}\alpha_k\right)=\frac{\sqrt{\tau}}{\tilde{\mu}_{\mathcal{D}}}\left(\frac{(\theta_k-\theta_{k+1})\sqrt{k+1}}{\alpha_0}+\frac{L\alpha_0}{2\tau\sqrt{k+1}}\right), \tag{7.3}$$

where $g_k = \|\nabla f(x_k)\|_2$. We know from (3.1) and the assumption that f is bounded below that $f_* \leq \theta_{k+1} \leq \theta_k \leq f(x_0)$ for all k. Letting $l = \lfloor K/2 \rfloor$, this implies that

$$\sum_{j=l}^{2l} \theta_j - \theta_{j+1} = \theta_l - \theta_{2l+1} \leq f(x_0) - f_* \overset{\text{def}}{=} C,$$

∎

from which we conclude that there must exist $j \in \{l, \ldots, 2l\}$ such that $\theta_j - \theta_{j+1} \leq C/(l+1)$. This implies that

$$\begin{aligned} g_j &\overset{(7.3)}{\leq} \frac{\sqrt{\tau}}{\tilde{\mu}_{\mathcal{D}}}\left(\frac{(\theta_j-\theta_{j+1})\sqrt{j+1}}{\alpha_0}+\frac{L\alpha_0}{2\tau\sqrt{j+1}}\right) \leq \frac{\sqrt{\tau}}{\tilde{\mu}_{\mathcal{D}}}\left(\frac{C\sqrt{j+1}}{\alpha_0(l+1)}+\frac{L\alpha_0}{2\tau\sqrt{j+1}}\right) \\ &\leq \frac{\sqrt{\tau}}{\tilde{\mu}_{\mathcal{D}}}\left(\frac{C\sqrt{2l+1}}{\alpha_0(l+1)}+\frac{L\alpha_0}{2\tau\sqrt{l+1}}\right) \leq \frac{\sqrt{\tau}}{\tilde{\mu}_{\mathcal{D}}\sqrt{l+1}}\left(\frac{\sqrt{2}C}{\alpha_0}+\frac{L\alpha_0}{2\tau}\right) \\ &\leq \frac{1}{\tilde{\mu}_{\mathcal{D}}\sqrt{K/2}}\left(\frac{\sqrt{2\tau}C}{\alpha_0}+\frac{L\alpha_0}{2\sqrt{\tau}}\right) \overset{(7.1)}{\leq} \varepsilon. \end{aligned}$$

Note that $\alpha_0 = \frac{\sqrt{2\sqrt{2}\tau}\sqrt{f(x_0)-f_*}}{\sqrt{L}}$ gives the optimal rate which does not depend on τ and coincides with the rate for spherical setup in the `STP` method. It means that for big enough τ the previous theorem gives a complexity guarantee of the form $O(\frac{n}{\varepsilon^2})$.

7.2. Convex Case. In this subsection we state the main complexity result when f is convex.

THEOREM 7.3. *Let Assumptions 3.2, 5.1 (with $\|\cdot\|_{\mathcal{D}} = \|\cdot\|_2$), 5.4 and 7.1 be satisfied. Let $\alpha_k = \alpha_0 (f(x_k) - f_*)$, where $0 < \alpha_0 \leq \frac{2\tau\tilde{\mu}_{\mathcal{D}}}{R_0 L}$. Define $a = \frac{\tilde{\mu}_{\mathcal{D}}\alpha_0}{\sqrt{\tau}R_0} - \frac{L\alpha_0^2}{2\tau}$. If $k \geq k(\varepsilon) \overset{def}{=} \frac{1}{a}\left(\frac{1}{\varepsilon} - \frac{1}{r_0}\right)$, then $\mathbf{E}\left[f(x_k) - f(x_*)\right] \leq \varepsilon$.*

Proof. We have

$$\begin{aligned} \mathbf{E}[f(x_{k+1}) \mid x_k] - f_* &\overset{①}{\leq} f(x_k) - f_* - \frac{\alpha_k\tilde{\mu}_{\mathcal{D}}}{\sqrt{\tau}}\|\nabla f(x_k)\|_2 + \frac{L\alpha_k^2}{2\tau} \\ &\overset{②}{\leq} f(x_k) - f_* - \frac{\tilde{\mu}_{\mathcal{D}}\alpha_k}{R_0\sqrt{\tau}}(f(x_k) - f_*) + \frac{L\alpha_k^2}{2\tau} \end{aligned} \tag{7.4}$$

where ① follows from (7.2), and ② follows from Assumption 5.1. Using our choice of $\alpha_k = \alpha_0(f(x_k) - f_*)$ and taking full mathematical expectation from the both sides of (7.2) we obtain

$$r_{k+1} \leq r_k - \left(\frac{\tilde{\mu}_{\mathcal{D}}\alpha_0}{\sqrt{\tau}R_0} - \frac{L\alpha_0^2}{2\tau}\right) r_k^2 = r_k - a r_k^2.$$

Therefore, $\frac{1}{r_{k+1}} - \frac{1}{r_k} = \frac{r_k - r_{k+1}}{r_k r_{k+1}} \geq \frac{r_k - r_{k+1}}{r_k^2} \geq a$. From this we have $\frac{1}{r_k} \geq \frac{1}{r_0} + ka$ and hence $r_k \leq \frac{1}{\frac{1}{r_0}+ka}$. Finally, if $k \geq \frac{1}{a}\left(\frac{1}{\varepsilon} - \frac{1}{r_0}\right)$, then $r_k \leq \frac{1}{\frac{1}{r_0}+ka} \leq \varepsilon$. ∎

Note that $\alpha_0 = \frac{\sqrt{\tau}\tilde{\mu}_{\mathcal{D}}}{R_0 L}$ maximizes the value a. The optimal value of a is $\frac{\tilde{\mu}_{\mathcal{D}}^2}{2R_0 L^2}$, which is proportional to $\frac{1}{\pi n R_0 L^2}$ due to the second part of Lemma 3.4. It means that for big enough τ the above theorem gives an iteration complexity guarantee of the form $O(\frac{n}{\varepsilon})$.

7.3. Strongly Convex Case. In this subsection we state the main complexity result when f is strongly convex. The following theorem is an adaptation of Theorem 6.3.

THEOREM 7.4. *Let Assumptions 3.3, 3.2, 5.4, 6.1 and 7.1 be satisfied. Let* $\alpha_k = \frac{\theta_k \tilde{\mu}_{\mathcal{D}} \sqrt{\tau}}{L} \sqrt{2\lambda(f(x_k) - f(x_*))}$, *for some* $\theta_k \in (0,2)$ *such that* $\theta \overset{def}{=} \inf_k 2\theta_k - \theta_k^2 > 0$. *If*

$$K \geq \frac{L}{\lambda \tilde{\mu}_{\mathcal{D}}^2 \theta} \log\left(\frac{f(x_0) - f(x_*)}{\varepsilon}\right), \tag{7.5}$$

then $\mathbf{E}\left[f(x_K) - f(x_*)\right] \leq \varepsilon$.

Proof. By injecting α_k into the first inequality of (7.2) we get

$$\begin{aligned}\mathbf{E}[f(x_{k+1})|x_k] - f(x_*) \leq f(x_k) - f(x_*) - \frac{\tilde{\mu}_{\mathcal{D}}^2 \theta_k \sqrt{2\lambda(f(x_k)-f(x_*))} \|\nabla f(x_k)\|_2}{L} \\ + \frac{\tilde{\mu}_{\mathcal{D}}^2 \theta_k^2 \lambda (f(x_k)-f(x_*))}{L}.\end{aligned}$$

From the strong convexity property of f we have $\|\nabla f(x_k)\|_2^2 \geq 2\lambda(f(x_k) - f(x_*))$. Therefore

$$\begin{aligned}\mathbf{E}[f(x_{k+1})|x_k] - f(x_*) &\leq f(x_k) - f(x_*) - \frac{2\tilde{\mu}_{\mathcal{D}}^2 \theta_k \lambda (f(x_k)-f(x_*))}{L} + \frac{\tilde{\mu}_{\mathcal{D}}^2 \theta_k^2 \lambda (f(x_k)-f(x_*))}{L} \\ &\leq f(x_k) - f(x_*) - \frac{\tilde{\mu}_{\mathcal{D}}^2 \lambda (f(x_k)-f(x_*))}{L}(2\theta_k - \theta_k^2) \\ &\leq f(x_k) - f(x_*) - \frac{\tilde{\mu}_{\mathcal{D}}^2 \theta \lambda (f(x_k)-f(x_*))}{L},\end{aligned}$$

by using the definition of θ. Let $r_k = \mathbf{E}\left[f(x_k)\right] - f(x_*)$. By taking the expectation of the last inequality we get $r_{k+1} \leq \left(1 - \frac{\tilde{\mu}_{\mathcal{D}}^2 \theta \lambda}{L}\right) r_k$, hence

$$r_k \leq \left(1 - \frac{\tilde{\mu}_{\mathcal{D}}^2 \theta \lambda}{L}\right)^k r_0.$$

Therefore, if K satisfies (7.5), we get $r_K \leq \varepsilon$. ∎

For big enough τ the above theorem gives an iteration complexity guarantee of the form $O(n \log\left(\frac{1}{\varepsilon}\right))$.

8. Numerical Results. In this section, we report the results of some preliminary experiments performed in order to assess the efficiency and the robustness of the proposed algorithms compared to the coordinate search method (this method will be called `DDS`) and the algorithm proposed in [18]. In the latter approach, at each iteration k, a random vector s_k following the uniform distribution on the unit sphere is generated, then the next iterate is computed as follows

$$x_{k+1} = x_k - \alpha_k \frac{f(x_k + \mu_k s_k) - f(x_k)}{\mu_k} s_k, \tag{8.1}$$

where $\mu_k \in (0,1)$ is the finite differences parameter, and α_k is the stepsize. This method generates a trial step similar to one of the trial steps in our method ($x_- = x_k - \alpha_k s_k$) when the probability distribution $\mathcal{D}$ is chosen to be the uniform distribution on the unit sphere up to the multiplication of the step by $\frac{f(x_k+\mu_k s_k)-f(x_k)}{\mu_k}$. This method will be called `RGF` (Random Gradient free method). All the results presented here are averaged over 10 runs of the algorithms. We did all our experiments using Matlab.

To compare the performance of the algorithms we use performance profiles proposed by Dolan and Moré [6] over a variety of problems. Given a set of problems $\mathcal{P}$ (of cardinality $|\mathcal{P}|$) and a set of algorithms (solvers) $\mathcal{S}$, the performance profile $\rho_s(\tau)$ of an algorithm s is defined as the fraction of problems where the performance ratio $r_{p,s}$ is at most τ

$$\rho_s(\tau) \;=\; \frac{1}{|\mathcal{P}|}\text{size}\{p \in \mathcal{P} : r_{p,s} \leq \tau\}.$$

The performance ratio $r_{p,s}$ is in turn defined by

$$r_{p,s} \;=\; \frac{t_{p,s}}{\min\{t_{p,s} : s \in \mathcal{S}\}},$$

where $t_{p,s} > 0$ measures the performance of the algorithm s when solving problem p, seen here as the number of function evaluation. Better performance of the algorithm s, relatively to the other algorithms on the set of problems, is indicated by higher values of $\rho_s(\tau)$. In particular, efficiency is measured by $\rho_s(1)$ (the fraction of problems for which algorithm s performs the best) and robustness is measured by $\rho_s(\tau)$ for τ sufficiently large (the fraction of problems solved by s). Following what is suggested in [6] for a better visualization, we will plot the performance profiles in a $\log_2$-scale (for which $\tau = 1$ will correspond to $\tau = 0$).

The distribution $\mathcal{D}$ used here for our random direction generation is the uniform distribution on the unit sphere. We performed other experiments (not reported here) with different choices for distributions $\mathcal{D}$. For instance, the distributions listed in Lemma 3.4. We found similar performance as those reported here. The parameters defining the implemented algorithms are set as follows: For `RGF` we choose $\mu_k = 10^{-4}$, and $\alpha_k = \frac{1}{4(n+4)}$ where n is the problem dimension. For this method the authors proposed to use the stepsize $\alpha_k = \frac{1}{4L(n+4)}$, where L is the Lipschitz constant of the gradient of the objective function. Since for our test problems we do not know this constant, we ran `RGF` method with different values for L, for instance 0.1, 1, 10, and 100. The best performance was found for $L = 1$. The stepsize in `DDS` is initialized by $\alpha_0 = 1$, then it is updated dynamically with the iterations by multiplying it by 2 when the step is successful and dividing it by 2 otherwise.

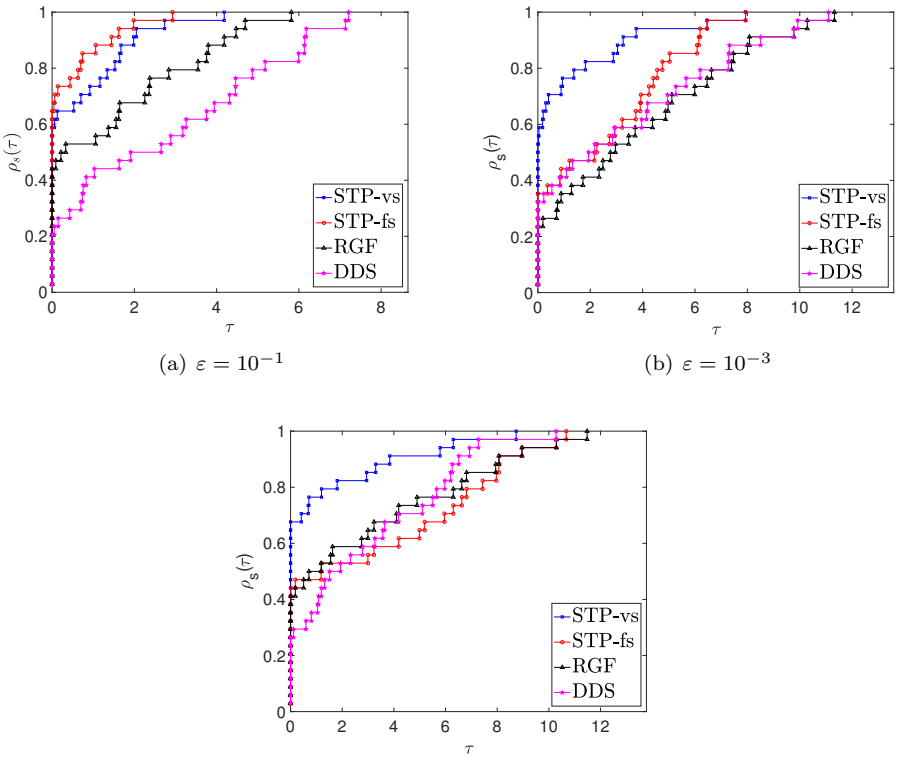
For all algorithms, we counted the number of function evaluations taken to (i) drive the function value below $f^* + \varepsilon\,(f(x_0) - f^*)$, where f^* is a local minimal value of the objective function f, and ε is a tolerance. In our experiments $\varepsilon = 10^{-1},\ 10^{-3}$ and 10^{-5}, (ii) or the maximum number of iterations attains 100000.

8.1. Non-Convex Case. In this section, we report the results of comparison of our approach `STP` for non-convex problems with `DDS` and `RGF`. We will call our `STP` method when using the variable stepsize `STP-vs`, and `STP-fs` when we use a fix stepsize. For `STP-vs` we choose $\alpha_k = \frac{1}{\sqrt{k+1}}$. For `STP-fs` we choose $\alpha_k = \alpha = 0.1\varepsilon$.

We use the Moré/Garbow/Hillstrom 34 test problems [17] which are implemented in Matlab. All the test problems are smooth. The dimension n of the problems changes between $n = 2$ to $n = 100$, typically $n = 2, 10, 50$ and 100. We use the starting points and the values f^* suggested in [17] for all the problems.

Figure 1 depicts the performance profiles of the algorithms. It shows that our approach (the methods `STP-vs` and `STP-fs`) improves the efficiency of the `DDS` and `RGF` algorithms on the tested problems. In fact, the number of the function evaluations performance profiles show that the use of the random directions leads to a significant

improvement on terms of the efficiency (for $\tau = 0$, on about 40% of the tested problems our approach performs the best, and less than 5% for `RGF` and `DDS`). From Figures 1(a) and 1(b), we see that the use of the random directions leads to a better robustness when a small precision is targeted (i.e. $\varepsilon = 10^{-1}$ and $\varepsilon = 10^{-3}$). However, when a big precision ($\varepsilon = 10^{-5}$) is targeted `DDS` becomes competitive. In fact, as shown in Figure 1(c), `DDS` is more robust than `RGF` approach and our method using fix stepsize. Our method `STP-vs` still more robust than `DDS`.



(a) $\varepsilon = 10^{-1}$ (b) $\varepsilon = 10^{-3}$

(c) $\varepsilon = 10^{-5}$

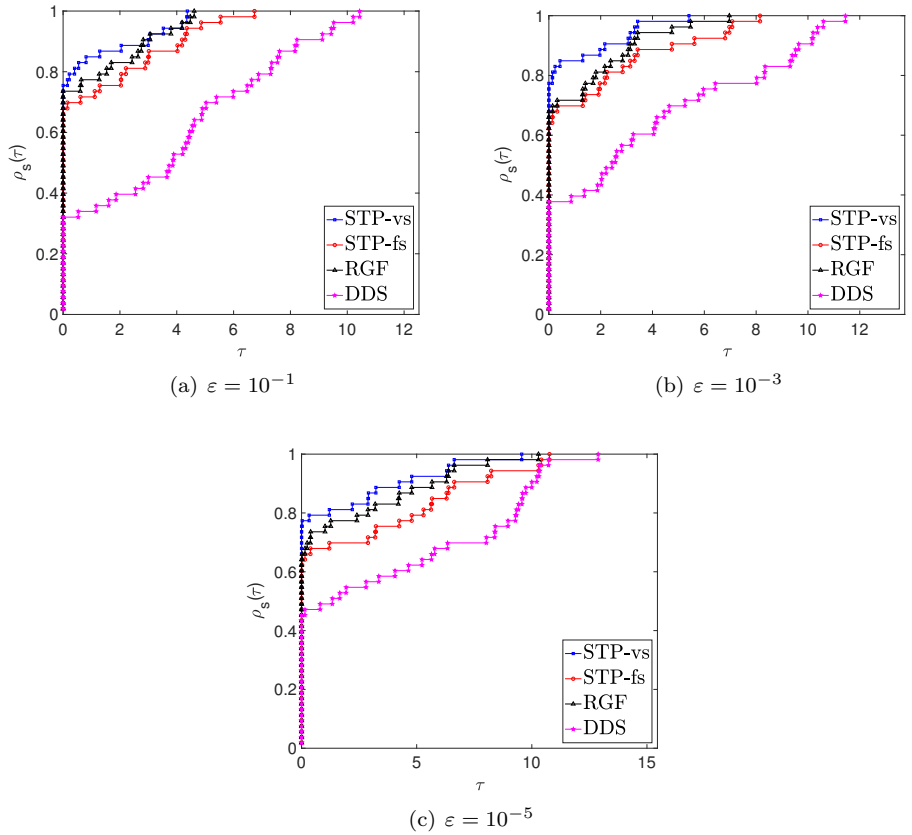
FIG. 1. *Performance profiles on* 34 *optimization problems.*

8.2. Convex Case. In this section, we report the results of comparison of two `STP` methods for convex problems with `DDS` and `RGF`. The first `STP` method is the one using the variable stepsize $\alpha_k = \left|\frac{1}{Lt}(f(x_k + ts_k) - f(x_k))\right|$, where $t = 10^{-4}$. We will call this method `STP-vs`. The second `STP` method is the one using the fix stepsize $\alpha_k = \alpha = 0.1\varepsilon$. It will be called `STP-fs`.

We selected from the Moré/Garbow/Hillstrom problems those with a unique minimum. To have a large bed test, we create different instances for problems by varying the problem dimension n when it is possible. Our test bed in this section contains 40 problems.

In Figure 2, the performance profiles show that the random based methods (`RGF` method and our two methods `STP-vs` and `STP-vs`) outperform by far the `DDS` method. Our method `STP-vs` gives the best performances for small precision (see Figures 2(a) and 2(b)). For big precision ($\varepsilon = 1e-5$), it gives almost similar performances as `RGF`

method ((see Figure 2(c)). Our method `STP-fs` is outperformed by `RGF`.



(a) $\varepsilon = 10^{-1}$ (b) $\varepsilon = 10^{-3}$

(c) $\varepsilon = 10^{-5}$

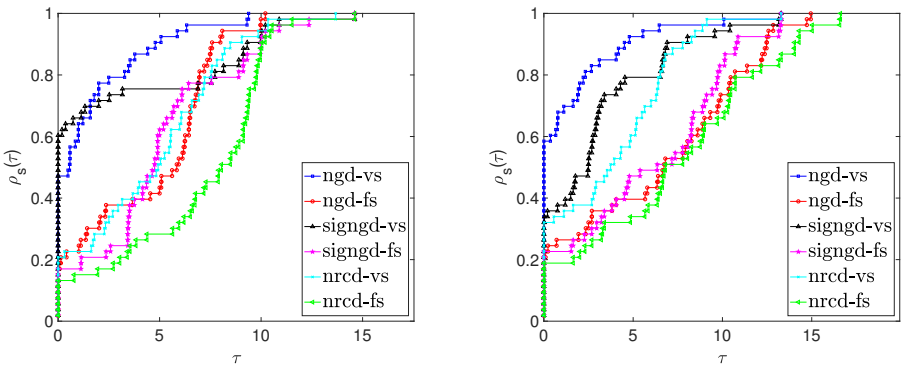
FIG. 2. *Performance profiles on* 40 *optimization problems.*

8.3. First order methods. In this section, we report the results of comparison of gradient based methods that our approach cover, using the variable stepsize $\alpha_k = \frac{1}{\sqrt{k+1}}$ and the fix stepsize $\alpha_k = 0.1\varepsilon$. In fact to select these stepsizes, we run many experiments with different values and found the best results for the chosen stepsizes. We denote with `ngd-vs`, and `ngd-fs` the Normalized Gradient Descent (NGD) methods using the variable stepsize, and the fix stepsize respectively. With similar notation we denote by `signgd-vs`, and `signgd-fs` the Signed Gradient Descent (SignGD) methods and by `nrcd-vs`, and `nrcd-fs` Normalized Randomized Coordinate Descent (NRCD) methods using the variable stepsize, and the fix stepsize respectively.

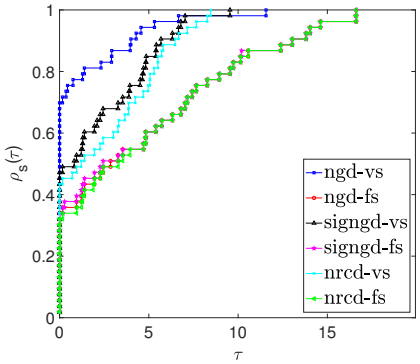
We use the Moré/Garbow/Hillstrom 34 test problems for which we add 20 problems by creating different instances for problems by varying the problem dimension n when it is possible. Our test bed in this section contains 54 problems.

Figure 3 depicts the performance profiles of the algorithms. It shows that the use of the variable stepsize gives better performances than the fix stepsize. As one may expect, the normalized gradient descent method `ngd-vs` exhibits performances better than the other methods, except for small precision ($\varepsilon = 1e-1$), it is less efficient than

signed gradient descent method `signgd-vs`. The latter method is more efficient and less robust than normalized randomized coordinate descent method `nrcd-vs`.



(a) $\varepsilon = 10^{-1}$ (b) $\varepsilon = 10^{-3}$



(c) $\varepsilon = 10^{-5}$

FIG. 3. *Performance profiles on* 54 *optimization problems.*

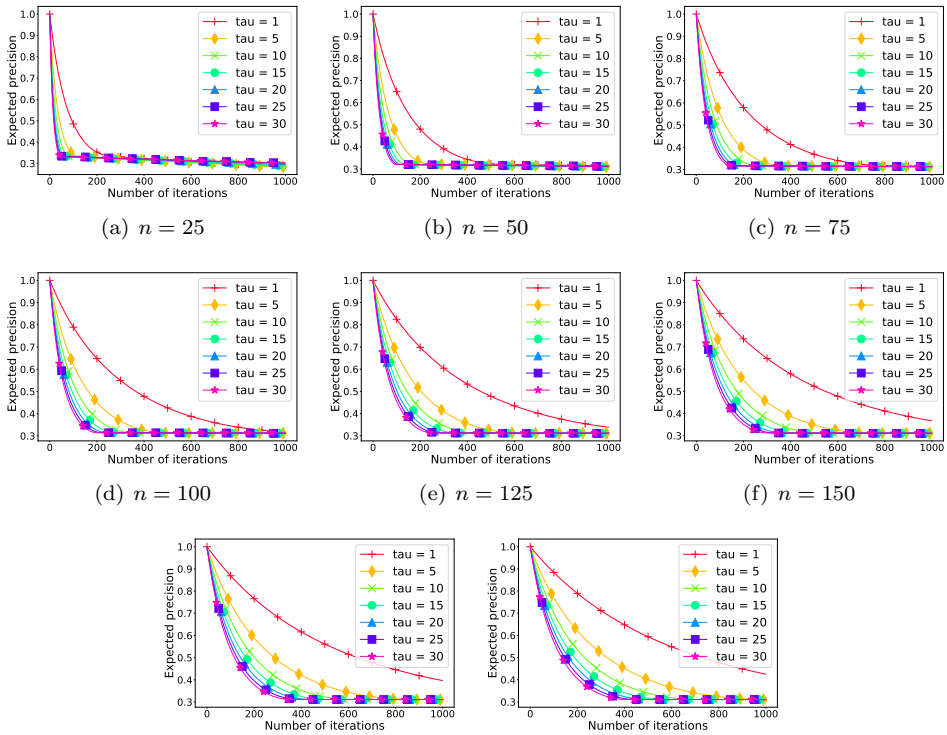
8.4. Experiments for `PSTP`. We considered the following function

$$f(x) = \frac{1}{2}x_1^2 + \frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2 + \frac{1}{2}x_n^2 - x_1$$

and run `PSTP` with different τ (see Figure 4). From the numerical results we see that the rate could be worse for small τ than for τ. It happens because the Assumption 5.4 does not have to be true for small τ with the parameter $\tilde{\mu}_{\mathcal{D}} \sim \sqrt{\frac{2}{\pi n}}$ as we use in the experiments (recall that this parameter corresponds to the statement of Central Limit Theorem and, therefore, we need τ to be big enough). When τ is big enough the Assumption 7.1 will holds and we will obtain the improvement of the rate. We measure $\frac{f(x_k)-f_*}{f(x_0)-f_*}$ on the y-axis and call it "Expected precision".

8.5. `STP` vs `RGF`. We considered the following function

$$f(x) = \frac{1}{2}x_1^2 + \frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2 + \frac{1}{2}x_n^2 - x_1$$



(a) $n = 25$ (b) $n = 50$ (c) $n = 75$

(d) $n = 100$ (e) $n = 125$ (f) $n = 150$

(g) $n = 175$ (h) $n = 200$

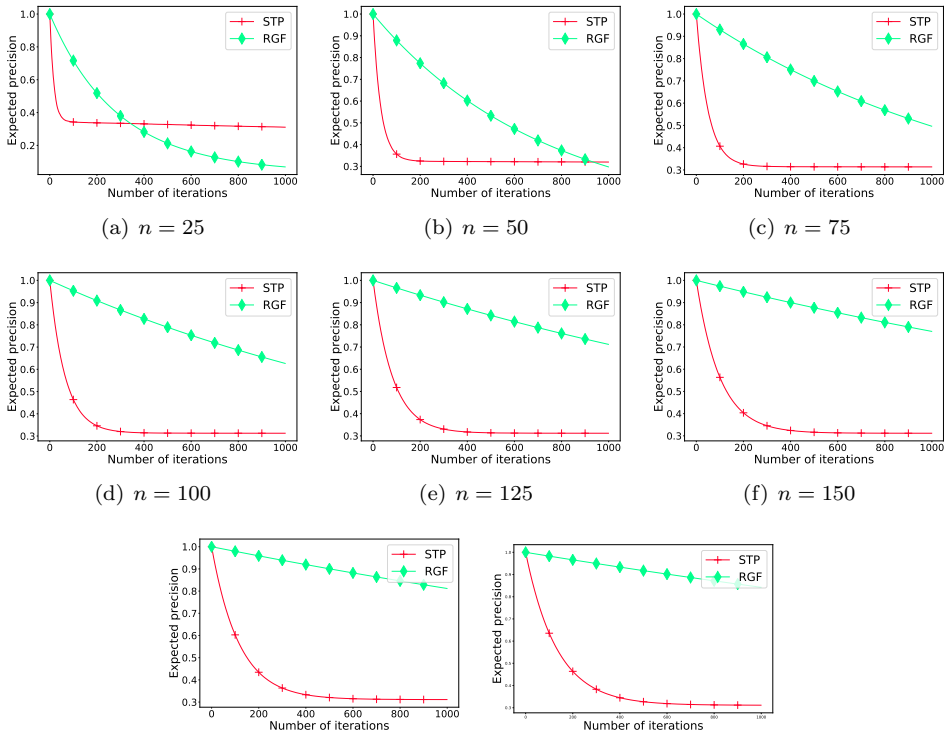
FIG. 4. *Trajectories of PSTP for the different n.*

and run STP and RGF for different n (see Figure 5). We measure $\frac{f(x_k)-f_*}{f(x_0)-f_*}$ on the y-axis and call it "Expected precision". One can notice that STP becomes more beneficial then RGF when n is growing.

9. Conclusions. In this paper, we have proposed a very simple randomized algorithm — Stochastic Three Points (STP) method — for derivative free optimization (DFO). At each iteration, the proposed method try to decrease the objective function along a random direction sampled from a certain fixed probability law. Under mild assumption on this law, we have given the properties of this method for non-convex, convex and strongly convex problems. In fact, we have derived different practical rules for the stepsizes for which this method converges in expectation to a stationary point of the considered problem.

We have derived the worst case complexity of STP. In fact, in the non-convex case, we have shown that STP needs $O(n\varepsilon^{-2})$ function evaluations to find a point at which the gradient of the objective function is below ε, in expectation. In the convex case, the number of iterations to find a point at which the distance between the objective function and its optimal value in expectation is $O(n\varepsilon^{-1})$. STP is shown to converge linearly for the strongly convex problems, i.e. the complexity is $O(n\log(1/\varepsilon))$. The complexity of STP depends linearly on the dimension of the considered problem, while this dependence is quadratic for deterministic direct search (DDS) methods. We have also proposed a parallel version for STP.

Our numerical experiments showed encouraging performance of the proposed STP



(a) $n = 25$ (b) $n = 50$ (c) $n = 75$

(d) $n = 100$ (e) $n = 125$ (f) $n = 150$

(g) $n = 175$ (h) $n = 200$

FIG. 5. *Trajectories of* `STP` *and* `RGF` *for the different* n.

algorithm. A number of issues need further investigation, in particular the best choice of probability law for choosing the random directions. Extending our results to the non smooth problems and/or the constrained problems remains an interesting topic for the future research. It would be also interesting to confirm the potential of the proposed `STP` approach compared to the classical approaches in DFO using extensive numerical tests.

REFERENCES

[1] G. ALLAIRE, *Shape Optimization by the Homogenization Method*, Springer, New York, USA, 2001.

[2] N. BABA, *Convergence of a random optimization method for constrained optimization problems*, Journal of Optimization Theory and Applications, 33 (1981), pp. 1–11.

[3] A. R. CONN, K. SCHEINBERG, AND L. N. VICENTE, *Introduction to Derivative-Free Optimization*, SIAM, Philadelphia, PA, USA, 2009.

[4] M. A. DINIZ-EHRHARDT, J. M. MARTINEZ, AND M. RAYDAN, *A derivative-free nonmonotone line-search technique for unconstrained optimization*, Journal of Optimization Theory and Applications, 219 (2008), pp. 383–397.

[5] M. DODANGEH AND L. N. VICENTE, *Worst case complexity of direct search*, Mathematical Programming, 155 (2016), pp. 307–332.

[6] E. D. DOLAN AND J. J. MORÉ, *Benchmarking optimization software with performance profiles*, Mathematical Programming, 91 (2002), pp. 201–213.

[7] C. DOREA, *Expected number of steps of a random optimization method*, Journal of Optimization Theory and Applications, 39 (1983), pp. 165–171.

[8] E. Gorbunov, P. Dvurechensky, and A. Gasnikov, *An accelerated method for derivative-free smooth stochastic convex optimization*, arXiv preprint arXiv:1802.09022, (2018).
[9] S. Gratton, C. W. Royer, L. N. Vicente, and Z. Zhang, *Direct search based on probabilistic descent*, SIAM Journal on Optimization, 25 (2015), pp. 1515–1541.
[10] J. Haslinger and R. Mckinen, *Introduction to Shape Optimization: Theory, Approximation, and Computation*, SIAM, Philadelphia, PA, USA, 2003.
[11] V. G. Karmanov, *Convergence estimates for iterative minimization methods*, USSR Computational Mathematics and Mathematical Physics, 14 (1974), pp. 1–13.
[12] ———, *On convergence of a random search method in convex minimization problems*, Theory of Probability and its applications, 19 (1974), pp. 788–794.
[13] T. G. Kolda, R. M. Lewis, and V. Torczon, *Optimization by direct search: New perspectives on some classical and modern methods*, SIAM Review, 45 (2003), pp. 385–482.
[14] J. Konečný and P. Richtárik, *Simple complexity analysis of simplified direct search*, arXiv preprint arXiv:1410.0390, (2014).
[15] J. Matyas, *Random optimization*, Automation and Remote Control, 26 (1965), pp. 246–253.
[16] B. Mohammadi and O. Pironneau, *Applied Shape Optimization for Fluids*, Clarendon Press, Oxford, 2001.
[17] J. J. Moré, B. S. Garbow, and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Transactions on Mathematical Software, 7 (1981), pp. 17–41.
[18] Y. Nesterov and V. Spokoiny, *Random gradient-free minimization of convex functions*, Foundations of Computational Mathematics, 17 (2017), pp. 527–566.
[19] B. T. Polyak, *Introduction to Optimization*, Optimization Software, Inc, New York, USA, 1987.
[20] M. Sarma, *On the convergence of the Baba and Dorea random optimization methods*, Journal of Optimization Theory and Applications, 66 (1990), pp. 337–343.
[21] S. U. Stich, C. L. Muller, and B. Gartner, *Optimization of convex functions with random pursuit*, arXiv preprint arXiv:1111.0194, (2011).
[22] L. N. Vicente, *Worst case complexity of direct search*, EURO Journal on Computational Optimization, 1 (2013), pp. 143–153.

Appendix A. Proof of Lemma 3.4.

1. $\gamma_{\mathcal{D}} = \mathbf{E}\|s\|_2^2 = \frac{1}{A_n(1)}\int_{\|s\|_2^2=1}\|s\|_2^2 ds = \frac{1}{A_n(1)}\int_{\|s\|_2^2=1} ds = 1$ where $A_n(1) = \frac{2\pi^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}$ is the air of the $n-1$-unit sphere and Γ is the gamma function.
   Let $\varepsilon_1 = g/\|g\|_2$ and $\varepsilon_2, \ldots, \varepsilon_n$ complete ε_1 to an orthonormal basis of $\mathbb{R}^n$ then

$$\begin{aligned}\mathbf{E}|\langle g, s\rangle| &= \frac{1}{A_n(1)}\int_{\|s\|_2^2=1}|\langle g, s\rangle| ds = \|g\|_2\frac{1}{A_n(1)}\int_{\sum_{i=2}^n s_i^2 = 1-s_1^2}|s_1| ds\\ &= \|g\|_2\frac{1}{A_n(1)}\int_{-1}^{1}|s_1|\int_{\sum_{i=2}^n s_i^2=1-s_1^2} ds_{2:n} ds_1\\ &= \|g\|_2\frac{1}{A_n(1)}\int_{-1}^{1}|s_1| A_{n-1}\left(1-s_1^2\right) ds_1,\end{aligned}$$

   where $A_{n-1}\left(1-s_1^2\right) = \frac{2\pi^{(n-1)/2}\left(1-s_1^2\right)^{n-2}}{\Gamma((n-1)/2)}$ is the volume of the $n-2$ sphere of radius $1-s_1^2$, hence

$$\begin{aligned}\mathbf{E}|\langle g, s\rangle| &= \|g\|_2\frac{1}{A_n(1)}\frac{2\pi^{(n-1)/2}}{\Gamma((n-1)/2)}\int_{-1}^{1}|s_1|\left(1-s_1^2\right)^{n-2} ds_1\\ &= \|g\|_2\frac{1}{A_n(1)}\frac{2\pi^{(n-1)/2}}{\Gamma((n-1)/2)(n-1)}.\end{aligned}$$

   If $n-1 = 2p$ then

$$\mathbf{E}|\langle g, s\rangle| = \|g\|_2\frac{2\pi^p\Gamma(p+1/2)}{2p\Gamma(p)2\pi^p\sqrt{\pi}} = \|g\|_2\frac{(2p)!}{2^{2p+1}(p!)^2} \quad \sim \quad \frac{\|g\|_2}{2\sqrt{\pi p}},$$

since according to Stirling formula, $p! \sim p^p e^{-p}\sqrt{2\pi p}$. If $n-1=2p+1$ then

$$\mathbf{E}|\langle g,s\rangle| = \|g\|_2 \frac{2\pi^p\sqrt{\pi}\Gamma(p+1)}{2\pi^{p+1}(2p+1)\Gamma(p+1/2)} = \|g\|_2 \frac{(p!)^2 2^{2p}}{(2p+1)!\pi} \sim \frac{\sqrt{p}}{\sqrt{\pi}(2p+1)} \sim \frac{\|g\|_2}{2\sqrt{\pi p}}$$

In the both cases, $\mathbf{E}|\langle g,s\rangle| \sim \frac{\|g\|_2}{2\sqrt{\pi p}} \sim \frac{\|g\|_2}{\sqrt{2\pi n}}$.

2. $\gamma_{\mathcal{D}} = \mathbf{E}\|s\|_2^2 = \frac{1}{n}\mathbf{E}\|x\|_2^2 = 1$, where $x \sim N(0,I)$.
Note that $s \sim \frac{1}{\sqrt{n}}N(0,I)$ implies $\langle g,s\rangle \sim \frac{1}{\sqrt{n}}N(0,\|g\|_2^2)$, hence

$$\mathbf{E}|\langle g,s\rangle| = \frac{1}{\|g\|_2\sqrt{2n\pi}}\int_{-\infty}^{+\infty}|x|e^{-\frac{x^2}{2\|g\|_2^2}}dx = \frac{\sqrt{2}}{\sqrt{n\pi}}\|g\|_2.$$

3. $\gamma_{\mathcal{D}} = \sum_{i=1}^n \|e_i\|_2^2 P(s=e_i) = 1$ and $\mathbf{E}|\langle g,s\rangle| = \frac{1}{n}\sum_{i=1}^n |g_i| = \frac{1}{n}\|g\|_1$.
4. $\gamma_{\mathcal{D}} = \sum_{i=1}^n \|e_i\|_2^2 P(s=e_i) = 1$ and $\mathbf{E}|\langle g,s\rangle| = \sum_{i=1}^n |g_i| P(s=e_i) = \sum_{i=1}^n p_i|g_i|$.
5. $\gamma_{\mathcal{D}} = \sum_{i=1}^n \|d_i\|_2^2 P(s=d_i) = \sum_{i=1}^n p_i = 1$ and $\mathbf{E}|\langle g,s\rangle| = \sum_{i=1}^n p_i|g_i d_i| = \|g\|_{\mathcal{D}}$.

Appendix B. Proof that our approach covers some first order methods.

- Normalized Gradient Descent (NGD) method:
At iteration k, $s \sim \mathcal{D}_k$ means that $s = \frac{g_k}{\|g_k\|_2}$ with probability 1.

$$\gamma_{\mathcal{D}_k} = \mathbf{E}_{s\sim\mathcal{D}_k}\|s\|_2^2 = 1,$$
$$\mathbf{E}_{s\sim\mathcal{D}_k}|\langle g_k,s\rangle| = \|g_k\|_2.$$

- Signed Gradient Descent (SignGD) method:
At iteration k, $s \sim \mathcal{D}_k$ means that $s = sign\,(g_k)$ with probability 1, where the *sign* operation is element wise sign.

$$\gamma_{\mathcal{D}_k} = \mathbf{E}_{s\sim\mathcal{D}_k}\|s\|_2^2 = \mathbf{E}_{s\sim\mathcal{D}_k}\|sign\,(g_k)\|_2^2 \le \sum_{i=1}^n 1 = n,$$
$$\mathbf{E}_{s\sim\mathcal{D}_k}|\langle g_k,s\rangle| = \mathbf{E}_{s\sim\mathcal{D}_k}|\langle g_k, sign\,(g_k)\rangle| = \|g_k\|_1.$$

- Normalized Randomized Coordinate Descent (NRCD) method (equivalently this method can be called Randomized Signed Gradient Descent):
At iteration k, $s \sim \mathcal{D}_k$ means that $s = \frac{g_k^i}{|g_k^i|}e_i$ with probability $\frac{1}{n}$, where g_k^i is the $i-th$ component of g_k.

$$\gamma_{\mathcal{D}_k} = \mathbf{E}_{s\sim\mathcal{D}_k}\|s\|_2^2 = \frac{1}{n}\sum_{i=1}^n 1 = 1$$
$$\mathbf{E}_{s\sim\mathcal{D}_k}|\langle g_k,s\rangle| = \mathbf{E}_{i\sim U[1,\dots,n]}\left|\left\langle g_k, \frac{g_k^i}{|g_k^i|}e_i\right\rangle\right| = \frac{1}{n}\sum_{i=1}^n |g_k^i| = \frac{1}{n}\|g_k\|_1.$$

- Normalized Stochastic Gradient Descent (NSGD) method:
At iteration k, $s \sim \mathcal{D}_k$ means that $s = \hat{g}_k$ where $\hat{g}_k$ is the stochastic gradient satisfying $\mathbf{E}\left[\hat{g}_k\right] = \frac{g_k}{\|g_k\|_2}$, and $\mathbf{E}\left[\|\hat{g}_k\|_2^2\right] \le \sigma < \infty$.

$$\mathbf{E}_{s\sim\mathcal{D}_k}|\langle g_k,s\rangle| = \mathbf{E}_{s\sim\mathcal{D}_k}|\langle g_k,\hat{g}_k\rangle| \ge \mathbf{E}_{s\sim\mathcal{D}_k}\langle g_k,\hat{g}_k\rangle = \|g_k\|_2.$$