Zeroth order optimization with orthogonal random directions

David Kozak[*]   Cesare Molinari[†]   Lorenzo Rosasco[‡]   Luis Tenorio [§]   Silvia Villa [¶]

Abstract. We propose and analyze a randomized zeroth-order approach based on approximating the exact gradient by finite differences computed in a set of orthogonal random directions that changes with each iteration. A number of previously proposed methods are recovered as special cases including spherical smoothing, coordinate descent, as well as discretized gradient descent. Our main contribution is proving convergence guarantees as well as convergence rates under different parameter choices and assumptions. In particular, we consider convex objectives, but also possibly non-convex objectives satisfying the Polyak-Łojasiewicz (PL) condition. Theoretical results are complemented and illustrated by numerical experiments.

1 Introduction

It is common in engineering, economics, statistics, and machine learning to try to minimize a function for which no analytical form is readily accessible, and only a zeroth-order oracle giving the function value at a given point is available. Sometimes an analytical form for the function exists, but the gradient does not have an explicit expression or it is infeasible to obtain. Specific examples of both scenarios are highlighted in the first chapter of [18], to which we add reinforcement learning [48, 35, 15, 22]. When only function evaluations are available the options for optimization are somewhat limited. Some of the earliest theoretical analysis of zeroth order optimization was done on *random search* [37, 44]. In random search, a step is proposed in a randomly chosen direction and the objective function is computed; if it provides an improvement over the function value at the current position then the step is taken, otherwise a new random sample is drawn and the process is repeated. This approach has the downside of not using derivative information, therefore relying on inexpensive function evaluations and converging slowly due to the uninformed search directions. Older methods, such as discrete gradient descent and discrete coordinate descent [56, 30], have similar aims but actually use approximate derivative information to determine how far to step.Extensions to functions that are differentiable almost-everywhere have been considered in [25, 26]. There are alternative algorithms that behave similarly to (randomized) finite difference coordinate descent but use a random basis for choosing descent directions [34, 40, 20, 22] and ample experimental results showcase the practical utility of these approaches [48, 35, 15]. The Itoh–Abe discrete gradient method studied in [24, 45, 21] is another approach to zeroth-order optimization, but at each iteration and for each coordinate it requires to solve a scalar equation involving the objective function. Similar considerations apply for its stochastic version, where the stepsize is still defined in an implicit way.

---

[*]Solea Energy. E-mail: dkozak@soleaenergy.com

[†]Istituto Italiano di Tecnologia. E-mail: cecio.molinari@gmail.com

[‡]MaLGa, DIBRIS, Università di Genova CBMM, MIT Istituto Italiano di Tecnologia. E-mail: lorenzo.rosasco@unige.it

[§]Department of Applied Mathematics and Statistics, Colorado School of Mines. E-mail: ltenorio@mines.edu

[¶]MaLGa, DIMA, Università degli Studi di Genova. E-mail: villa@dima.unige.it

Our approach is based on approximating the exact gradient by finite differences computed in a set of orthogonal random directions. Different randomized projection can be considered, recovering different approaches as special cases. Indeed, our general approach recovers finite difference versions of coordinate descent, descent in subspaces defined by random orthogonal matrices, and spherical smoothing as special cases. All these methods can be treated in a unified way with our approach. Our main contribution is proving convergence results as well as convergence rates. For convex objectives, we study convergence in function value and give realizable conditions for convergence of the iterates. Beyond convexity, we consider the case when the objective satisfies the Polyak-Łojasiewicz (PL) condition for which we can give stronger guarantees with faster rates. The PL condition is necessarily satisfied by functions that are strongly convex in which case our results are trivially extended to convergence of the iterates to the unique minimizer. For many instances of our general algorithm the derived results are new. Simple experiments are also provided to show that the rates described by the theorems are achievable in practice.

The paper is organized as follows. In Section 2, we describe the general setting, proposed approach and its special cases. In Section 3, we summarize and discuss our main results. In Section 4, we present useful preliminary results. In Section 5 and 6, we detail and prove the results for the convex case and for objective functions satisfying a PL condition - respectively. In Section 7 we present numerical experiments and conclude in Section 8 with some remarks and open questions.

2 Problem statement

We consider the problem of finding a point $x_* \in \mathbb{R}^d$ such that $f(x_*) = f_*$, where

$$f_* := \inf_{x \in \mathbb{R}^d} f(x). \tag{P}$$

The function $f \colon \mathbb{R}^d \to \mathbb{R}$ satisfies the following hypotheses:

(**H.1**) $f_* > -\infty$, f is differentiable and ∇f is λ-Lipschitz; that is, for every x_1 and $x_2 \in \mathbb{R}^d$,

$$\|\nabla f(x_2) - \nabla f(x_1)\| \le \lambda \|x_2 - x_1\|.$$

(**H.2**) f is convex and has a minimizer in $\mathbb{R}^d$.

The goal of this paper is to study the convergence properties of the stochastic iterative procedure described in Algorithm 1, designed to solve numerically problem (P). This Algorithm is based on the finite difference approximations of a subset of directional derivatives of f randomly chosen at each iteration. More precisely, the steps of the algorithm are summarized as follows:

Algorithm 1: Zeroth order stochastic subspace algorithm

Let $x_0 \in \mathbb{R}^d$, let $(\alpha_k)_{k \in \mathbb{N}}$ and $(h_k)_{k \in \mathbb{N}}$ be sequences of positive real values and let $(P_k)_{k \in \mathbb{N}}$ be a
  sequence of independent random matrices in $\mathbb{R}^{d \times \ell}$ defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$.
for $k = 0, 1, \ldots$ **do**
 | $x_{k+1} = x_k - \alpha_k P_k \nabla_{(P_k, h_k)} f(x_k)$.
end

We briefly introduce the notation and explain the main ideas behind the algorithm. Given $\ell \le d$, a matrix $P \in \mathbb{R}^{d \times \ell}$ and an index $j \in [\ell] := \{1, \ldots, \ell\}$, we let $p^{(j)} \in \mathbb{R}^d$ denote the j-th column of P. For $h > 0$,

define the vector $\nabla_{(P,h)}f(x) \in \mathbb{R}^\ell$ with entries

$$\left[\nabla_{(P,h)}f(x)\right]_j := \frac{f(x + hp^{(j)}) - f(x)}{h}, \qquad \forall j \in \{1, \ldots, \ell\}. \tag{2.1}$$

Note that $\nabla_{(P,h)}f(x)$ is the finite difference approximation of the directional derivatives of f in the directions identified by the columns of P. Since f is differentiable, $\lim_{h \to 0^+} \left[\nabla_{(P,h)}f(x)\right]_j = \langle \nabla f(x), p^{(j)} \rangle$ for every $j \in [\ell]$ and so

$$\lim_{h \to 0^+} \nabla_{(P,h)}f(x) = P^\top \nabla f(x).$$

Using the above notation, Algorithm 1 can be re-written more explicitly as

$$x_{k+1} = x_k - \alpha_k \sum_{j=1}^\ell \frac{f(x_k + h_k p_k^{(j)}) - f(x_k)}{h_k} \, p_k^{(j)}.$$

The recursion has the same structure as classical gradient descent, but the gradient at x_k is computed with two different forms of approximations. On the one hand, the directional derivatives are replaced by finite differences, no derivative of the function f is required. On the other hand, only the ℓ directions defined by the columns of P_k are used. Note that ℓ may be smaller than the dimension of the ambient space d. Moreover, the directions are chosen randomly at each iteration and are not necessarily drawn from the canonical basis. Indeed, throughout the paper the following properties are the only ones required on the random matrices P_k: for every $k \in \mathbb{N}$,

    (**P.1**) $P_k^\top P_k \overset{a.s.}{=} (d/\ell)\,\mathbb{I}$;
    (**P.2**) $\mathbb{E}P_k P_k^\top = \mathbb{I}$.

Next, we discuss several examples of algorithms that can be derived as special cases, see also Section 9.1.

Example 2.1 (Coordinate descent). If P_k contains ℓ columns of $\mathbb{I}_d$ chosen uniformly at random without replacement and with random sign, say $\pm e_{k_1}, \ldots \pm e_{k_\ell}$, then $p_k^{(j)} = \pm e_{k_j}$ for $j = 1, \ldots, \ell$, and Algorithm 1 corresponds to a discretized version of parallel block-coordinate descent.

Example 2.2 (Spherical smoothing). Consider $P_k = (\sqrt{d/\ell})\,Q_k \mathbb{I}_{d \times \ell}$, where Q_k is as in the QR-decomposition of a matrix $Z_k = Q_k R_k \in \mathbb{R}^{d \times d}$ with $R_{k_{ii}} > 0$, and each element of Z_k is drawn independently from $\mathcal{N}(0,1)$. The matrix $\mathbb{I}_{d \times \ell}$ truncates Q_k to its first ℓ columns so $Q_k \mathbb{I}_{d \times \ell}$ corresponds to ℓ columns of the random orthogonal matrix distributed according to the Haar measure on orthogonal matrices [38]. In other words, the columns $p_k^{(j)}$ are orthogonal and distributed uniformly on the sphere for all j. Then, when $\ell = 1$, $P_k \nabla_{(P_k,h)}f(x_k)$ is a spherical smoothing estimate of the gradient, as described in, e.g., [22, 7]. For the case $\ell > 1$ it is more common to sample $p^{(j)}$ independently and uniformly on the sphere [7], but in our case, to satisfy Assumptions (**P.1**) and (**P.2**), the columns of P must be orthonormal, similar to [33]. The advantage of a matrix P with orthonormal columns is discussed at length in [33]. We remark here that this property is a valid extension of traditional spherical smoothing to subspaces of dimension greater than one, and is required to obtain our results and connect Algorithm (1) with discrete gradient descent with $\ell = d$.

Example 2.3 (Gradient descent on random subspaces). If $h_k = h$ for every $k \in \mathbb{N}$ and $h \to 0$, then $\nabla_{(P_k,h_k)}f(x_k)$ becomes $P_k^\top \nabla f(x_k)$ and the algorithm reduces to

$$x_{k+1} = x_k - \alpha_k P_k P_k^\top \nabla f(x_k). \tag{2.2}$$

For shorthand, when we reference Algorithm 1 with $h_k = 0$, we are referring to the use of exact directional derivatives as in recursion 2.2, which has been introduced and studied in [33]. Our analysis allows to recover and improve on previous results for the iteration in (2.2).

Example 2.4 (Gradient descent). When $\ell = d$ we have $PP^\top = P^\top P = \mathbb{I}_d$. Hence, P is an orthonormal basis for $\mathbb{R}^d$, so that discrete gradient descent and gradient descent are recovered as special cases of Algorithm 1 and recursion 2.2 respectively.

Before stating and discussing our main results, we add one remark.

Remark 2.5 (Derivative-free optimization and automatic differentiation). From a practical point of view, the implementation of methods based on exact gradient computations, such as (2.2), is restricted to cases where exact directional derivatives are available. Under certain conditions, one can use automatic differentiation to obtain directional derivatives, however this requirement restricts the user to particular software, and precludes experiments wherein the function is accessed only via blackbox function evaluations, as is more common in derivative-free optimization. Importantly, the recursion based on exact gradient computations cannot be used when the simulations are physical (such as robotics, and many engineering examples), or when the objective can not be described by the elementary functions available to automatic differentiation software (such as in reinforcement learning).

3   Main results

In this section, we summarize the main results of the paper. We provide convergence results, explain the dependence on the discretization parameters h_k and the choice of the stepsizes α_k, and provide context for the results within the larger body of literature. The section contains two theorems. Theorem 3.1 establishes convergence properties of Algorithm 1 in the convex case, while Theorem 3.7 deals with objective functions satisfying the Polyak-Łojasiewicz condition **H.3** without requiring convexity. The theorems are simplified versions of the results in Sections 5 and 6, where more detailed statements and proofs can be found.

Theorem 3.1 (Convergence - convex case). *Assume that conditions **H.1** and **H.2** are satisfied and suppose that **P.1** and **P.2** hold. Let (x_k) be a random sequence generated by Algorithm 1. Set $\Lambda := \lambda d/\ell$.*

(i) *Let $0 \leq h_k \leq \overline{h}$ and $\alpha_k = \alpha$ for some $\overline{h} > 0$ and $\alpha \in (0, 1/\Lambda)$. Then, for explicit constants C_1, C_2, C_3 and C_4,*

$$\min_{j \in [k]} \mathbb{E}(f(x_j) - f_*) \quad \leq \quad \max\{C_1/(k+1) + C_2\overline{h}^2 + C_3\overline{h}, \ C_2\overline{h}^2 + C_4\overline{h}\}.$$

(ii) *Set $h_k = 1/k^r$ with $r > 0$ and $\alpha_k = \alpha/k$ with $\alpha \in (0, 1/\Lambda)$. Then,*

$$\lim_k f(x_k) \quad \overset{a.s.}{=} \quad f_*$$

and, for an explicit constant $D_0 > 0$,

$$\min_{j \in [k]} \mathbb{E}\left[f(x_j) - f_*\right] \quad \leq \quad D_0/\ln k.$$

(iii) *Set $h_k = 1/k^r$ with $r > 1$ and $\alpha_k = \alpha$ with $\alpha \in (0, 1/\Lambda)$. Then,*

$$\lim_k f(x_k) \quad \overset{a.s.}{=} \quad f_*$$

and, for an explicit constant $D > 0$,

$$\min_{j \in [k]} \mathbb{E}\left[f(x_j) - f_*\right] \quad \leq \quad D/k.$$

4

Moreover, there exists a random variable x_* with values in $\operatorname{argmin} f$ *such that* $x_k \overset{a.s.}{\to} x_*$.

(iv) Consider recursion 2.2 with $\alpha_k = \alpha$ and $\alpha \in (0, 2/\Lambda)$. Then

$$f(x_k) - f_* \quad \overset{a.s.}{=} \quad o\left(1/k\right)$$

and there exists a random variable x_* with values in $\operatorname{argmin} f$ *such that* $x_k \overset{a.s.}{\to} x_*$.

Proof. The results in $(i), (ii), (iii)$ and (iv) are special cases of Theorems 5.2, 5.4, 5.6 and 5.12 - respectively. $\qquad\square$

More detailed results - and under milder assumptions - are given in the theorems cited above. There, for instance, one can find explicit computations of the constants C_1, C_2, C_3, C_4 and D. Some comments on Theorem 3.1 are in order. In (i) we show the case in which the discretization h_k does not decrease to zero. We get a stability estimate for the expectation of the function values at the best iterate, depending on the upper-bound on the error noise $\overline{h}$. Notice that we recover the rate $C_1/(k+1)$ as $\overline{h} \to 0$. On the other hand, for $\overline{h} > 0$, running the algorithm beyond a certain number of iterations does not lead to any guarantee of improvement to the best iterate. More precisely, if we consider a stopping time proportional to $1/\overline{h}$, then overall accuracy will be $O(\overline{h})$ (assuming $\overline{h} \leq 1$ for simplicity). In the rest of Theorem 3.1 we require $h_k \to 0$. In (ii) the discretization sequence is allowed to converge to zero (polynomially but) arbitrarily slowly, but also a vanishing stepsize α_k is required and the results we can obtain are quite weak: the function values converge to the optimum with a logarithmic rate (in expectation and for the best iterate). Better results can be obtained if h_k converges to zero fast enough, as shown in (iii). In this case we gain on two different sides: first, a constant stepsize α can be used, which is convenient from a numerical point of view; second, we get faster convergence rates for the function values of the form $1/k$ (again in expectation and for the best iterate); third, we have a.s. convergence of the iterates. To the best of our knowledge, this is the first result showing convergence of the iterates for these types of zeroth-order methods in the general convex case; recall that special cases of our method include several well-known methods such as coordinate descent and smoothing on a sphere. Finally, for recursion (2.2) in (iv), we obtain an a.s. convergence rate asymptotically faster than $1/k$ for the last iterate (and not only for the best one). We add four remarks.

Remark 3.2. In the setting of Theorem 3.1, $\Lambda = \lambda d/\ell$ plays the role of the Lipschitz constant. In (iv) the choice of stepsize is bounded above by the classical limit $2/\Lambda$. On the other hand, as will be evident in the proofs, (i)-(iii) require the stepsize to be bounded above by $1/\Lambda$.

Remark 3.3. The ergodic iterate $\bar{x}_k := \left(\sum_{j=1}^{k} \alpha_j\right)^{-1} \sum_{j=1}^{k} \alpha_j x_j$ attains the same rates as the ones above for the best iterate, but it is of little practical interest. Indeed, in order to apply the algorithm, f has to be evaluated at each iterate x_k and so it is possible to just keep the one that achieves the minimal value function.

Remark 3.4. Let N denote the number of function evaluations required to perform k iterations. At each iteration Algorithm 1 uses $\ell + 1$ function evaluations. Then $N = (\ell + 1)k$ and the rates in Theorem 3.1 (i)-(iii) can be easily re-written in terms of function evaluations. The same observation holds for the results in Theorem 3.7.

Remark 3.5 (Comparison with previous work). Here we compare to the results in the literature dealing with a convex objective function.
- The case $\ell = 1$ with Gaussian sampling is considered in [39] where one can find results similar to ours in terms of convergence rates but with worse constants. See Remark 5.9 for a precise comparison.

5

- In [20] a mirror-descent variant of Algorithm 1 is proposed to deal with a stochastic objective, similar to the setting in [30]. A rate $O(1/\sqrt{k})$ in expectation is derived, however the results should not be compared directly to our related results in (iii), as the setting is more challenging. See Remark 5.10.
- In [33] only the setting (iv) is studied. Stricter assumptions are required (e.g., existence of a finite R such that

$$\max_{x_*} \max_{x} \{\|x - x_*\| : f(x) \le f(x_0)\} \le R,$$

and only convergence in expectation is provided. The special case of parallel coordinate descent has been considered in [52] and [49].

The following Corollary considers the case where the algorithm is run for a finite number of iterations known a priori.

Corollary 3.6. *Under the assumptions of Theorem 3.1 (i), let $K \in \mathbb{N}$ and $\overline{h} \le 1/K$. Then, for some constant C,*

$$\min_{j \in [K]} \mathbb{E}(f(x_j) - f_*) \le C/K. \tag{3.1}$$

In particular, given a tolerance $\varepsilon > 0$, it is possible to choose a number of iteration $K \in \mathbb{N}$ such that $C/K \le \varepsilon$ and a discretization $\overline{h} \le 1/K$, in order to get

$$\min_{j \in [K]} \mathbb{E}(f(x_j) - f_*) \le \varepsilon. \tag{3.2}$$

Next we state and discuss a second set of results derived under different assumptions on the objective function. It is well-known that first order methods exhibit favorable convergence results for strongly convex functions. More recently, it has been proved that improved convergence rates can also be obtained when the objective function satisfies weaker geometrical assumptions that do not require convexity [2, 29]. In this paper we consider the (global) Polyak-Łojasiewicz condition:

(**H.3**) the function f is γ-PL; namely, for every $x \in \mathbb{R}^d$,

$$\|\nabla f(x)\|^2 \ge \gamma \left(f(x) - f_*\right).$$

The main example of functions satisfying the global PL condition is the class of strongly convex functions. For more examples in the non-convex setting, see [2] and the numerical experiments in Section 7. We stress again that in the next result we assume **H.3** but not **H.2**; that is, we do not require convexity of the objective function.

Theorem 3.7 (Convergence - PL case). *Assume that conditions **H.1** and **H.3** are satisfied and suppose that **P.1** and **P.2** hold. Let (x_k) be a random sequence generated by Algorithm 1. Set $\Lambda := \lambda d/\ell$ and let $\alpha \in (0, 2/\Lambda)$. Fix a constant $w \le 1$ such that $0 < w < 2 - \Lambda\overline{\alpha}$, and define $\eta := 1 - w\alpha\gamma/2$.*

(i') *Set $h_k \le \overline{h}$ for some $\overline{h} > 0$ and $\alpha_k = \alpha$. Then, for an explicit constant $C_1 > 0$,*

$$\mathbb{E}\left[f(x_k) - f_*\right] \quad \le \quad \eta^k \left(f(x_0) - f_*\right) + C_1 \, \overline{h}^2 \left[1 - \eta^k\right].$$

(ii') *Set $h_k = 1/k^r$ with $r > 0$ and $\alpha_k = \alpha$. Then, there exists a constant $C_2 > 0$ such that*

$$\mathbb{E}\left[f(x_k) - f_*\right] \quad \le \quad C_2/k^{2r}.$$

6

(iii') Set $h_k = \sqrt{\eta^k/k^r}$ with $r > 1$ and $\alpha_k = \alpha$. Then, for an explicit constant $C_2 > 0$,

$$\mathbb{E}\left[f(x_k) - f_*\right] \quad \leq \quad C_2 \eta^k.$$

(iv') Consider recursion (2.2) with $\alpha_k = \alpha$. Then,

$$\mathbb{E}\left[f(x_k) - f_*\right] \quad \leq \quad \eta^k \left(f(x_0) - f_*\right).$$

Proof. The results in (i'), (ii') and (iii') are special cases of Theorems 6.2, 6.4 and 6.8 - respectively; while the result in (iv') is presented here for completeness but is already considered in [33] - see Corollary 2.3. $\square$

Sharper and more detailed results are given in the theorems cited above, where the reader can also find the explicit computations of the constants involved in the rates. We remark only that, for a constant stepsize $\alpha_k = 1/\Lambda$, the decreasing rate in (i'), (ii') and (iii') is given by $\eta = 1 - \gamma/(2\Lambda)$. As a general comparison with Theorem 3.1, note that the results in Theorem 3.7 do not involve the best iterate but only the last one. As in the first result of Theorem 3.1, in Theorem 3.7 (i') the parameter h_k does not necessarily vanish and so the error produced by the finite difference discretization of the gradient does not converge to zero. This explains the substantial difference between this result and the ones in (ii')-(iv'): in the upper-bound of (i'), for $k \to +\infty$ the right-hand side does not vanish. We can only guarantee that the expectation of the function evaluations converges with a linear rate to a sublevel set of f with value $f_* + C_1\overline{h}^2$, also called *error dominated region*. On the other hand note that, in comparison with Theorem 3.1 (i), the upper-bound does not diverge with the iterations but it remains bounded. In (iii'), we study the case of polynomial decay of the discretization parameter; namely, a decay of the form $h_k = 1/k^r$ with $r > 0$. In this case, the upper-bound is proportional to $1/k^{2r}$ and the rate gets better for a faster decay of the discretization parameter. In (iii'), under the assumption of a sufficiently fast (exponential) decay of h_k and with a constant stepsize α_k, we get a linear rate of convergence to the optimal value. Finally in (iv'), for the recursion in (2.2) with α_k constant, we recover the linear rate already shown in [33]. We add three remarks.

Remark 3.8 (Stepsize bounds). Following the same discussion as in Remark 3.2, in all the results of Theorem 3.7 the stepsize must be bounded above by the classical limit $2/\Lambda$.

Remark 3.9 (Adaptivity). An important consequence of the analysis in the previous theorems is that Algorithm 1 is *adaptive*; that is, knowledge of the specific properties of f is not needed to ensure the corresponding convergence results. We give an explicit example. Consider an objective for which assumptions **H.1**, **H.2** are satisfied and we run the algorithm with $h_k = 1/k^r$ for $r > 1$ and $\alpha_k = \alpha$ for $\alpha \in (0, \ell/(\lambda d))$. Then the results in Theorem 3.1 (iii) hold; namely, we have convergence of the iterates, almost sure convergence of the function values and a $1/k$ rate for the best iterate in expectation. In the same setting, if the function satisfies the PL inequality (**H.3**), from Theorem 3.7 (ii') we get automatically a $1/k^{2r}$ rate for the last iterate in expectation.

Remark 3.10 (Almost sure convergence). Under the assumptions of Theorem 3.7 (ii'), we have that $\mathbb{E}\left[f(x_k) - f_*\right] \lesssim C_2/k^{2r}$. Then, If $r > 1/2$ the right hand side is summable and, by Markov's inequality and Borel-Cantelli Lemma, $f(x_k) \overset{\text{a.s.}}{\to} f_*$, so that $d(x_k) \overset{\text{a.s.}}{\to} 0$. In particular, if the function f has a unique minimizer x_*, $x_k \overset{\text{a.s.}}{\to} x_*$. Analogous reasoning holds for Theorem 3.7 (iii') and (iv').

Finally, the following Corollary considers the case where the algorithm is run for a finite number of iterations known a priori.

Corollary 3.11. *Under the assumptions of Theorem 3.7 (i'), let $K \in \mathbb{N}$ and $\overline{h} \le \eta^{K/2}$. Then,*

$$\mathbb{E}\left[f(x_K) - f_*\right] \le \left(f(x_0) - f_* + C_1\right)\eta^K. \tag{3.3}$$

Moreover, if $d(\cdot)$ is the Euclidean distance of the argument to the set $\operatorname{argmin}_x f$, then

$$\mathbb{E}\, d(x_K) \le \frac{2}{\gamma}\left(f(x_0) - f_* + C_1\right)\eta^K.$$

Proof. The first result follows directly from the assumption $\overline{h} \le \eta^{K/2}$ and Theorem 3.7 (i'). The second claim follows from the previous one and the following inequality, that holds under **H.3** - see e.g. [11]: for every $x \in \mathbb{R}^d$, $d(x) \le 2/\gamma\left(f(x) - f_*\right)$. □

Equation (3.3) allows for a solution minimizing f up to an arbitrary desired accuracy $\varepsilon > 0$ in expectation. To this aim it is sufficient to choose

$$K = \frac{\ln\left(\varepsilon/(f(x_0) - f_* + C_1)\right)}{\ln\eta} \quad \text{and} \quad \overline{h} \le \eta^{K/2}.$$

Note that the dependence on the number of iterations is logarithmic in ε. A similar observation holds for $\mathbb{E}d(x_K)$. If f has a unique minimizer x_* (e.g. f is strongly convex) Corollary 3.11 provides a convergence rate for the iterates to the minimizer.

In the next sections we provide more detailed statements and proofs of the results. We start with some useful preliminary results that are the basis for the development in later sections.

4 Preliminaries

4.1 Notation and stepsize assumptions

For $k \in \mathbb{N}$, we define

$$f_k := f(x_k), \quad \nabla f_k := \nabla f(x_k), \quad \nabla_k f_k := \nabla_{(P_k, h_k)} f(x_k).$$

Now we provide the main conditions on the parameters of the proposed algorithm. Setting $\Lambda := \lambda d/\ell$, the assumptions we consider on the sequences of stepsizes (α_k) and discretizations (h_k) are as follows:

(A.1) $0 < \alpha_k \le \overline{\alpha} < 2/\Lambda$;
(A.2) $0 < \alpha_k \le \overline{\alpha} < 1/\Lambda$;
(A.3) $(\alpha_k) \notin \ell^1$ and (h_k) is bounded above by some $\overline{h} \ge 0$;
(A.4) (α_k) is bounded below by some $\underline{\alpha} > 0$ and (h_k) is bounded above by some $\overline{h} \ge 0$.

Note that **A.2** implies **A.1** while **A.4** implies **A.3**. Also note that Λ plays a role analogous to the Lipschitz constant in classical gradient descent, and $\Lambda = \lambda$ when $\ell = d$.

4.2 Auxiliary bound

We start the analysis with an auxiliary lemma that estimates the distance between the surrogate of the gradient used in the algorithm and the projected exact gradient. For this first result we assume only Lipschitz continuity of the gradient and condition **P.1** on the matrix P_k; in particular, convexity of f is not needed. The upper-bound is a simple consequence of the Descent Lemma 9.2 and the proof is in the appendix.

Lemma 4.1. *Let f be a function satisfying* **H.1** *and $P_k \in \mathbb{R}^{d \times \ell}$ a random matrix satisfying* **P.1**. *Then, for every $x \in \mathbb{R}^d$ and every $k \in \mathbb{N}$,*

$$\left\| \nabla_{(P_k, h_k)} f(x) - P_k^\top \nabla f(x) \right\| \leq \frac{\lambda d h_k}{2\sqrt{\ell}} \qquad a.s.$$

Remark 4.2 (Discussion on the bias). Consider a sufficiently regular function $f : \mathbb{R}^d \to \mathbb{R}$. Fix $h > 0$, $\ell = 1$ and a $\mathbb{R}^d$-valued random vector p satisfying Assumptions **P.1**-**P.2**; namely,

$$\|p\| \stackrel{\text{a.s.}}{=} \sqrt{d} \quad \text{and} \quad \mathbb{E}\left[pp^\top\right] = \mathbb{I}.$$

For every $x \in \mathbb{R}^d$, by Taylor expansion we have that

$$\mathbb{E}\left[\nabla_{(p,h)} f(x)\right] = \mathbb{E}\left[\frac{f(x + hp) - f(x)}{h} p\right] \tag{4.1}$$

$$= \nabla f(x) + \frac{h}{2} \mathbb{E}\left[\langle \nabla^2 f(x) p, p \rangle p\right] + \frac{h^2}{6} \mathbb{E}\left[\langle [\nabla^3 f(x) p] p, p \rangle p\right] + o(h^2).$$

We focus on two fundamental choices for the distribution of p.

- In the coordinate-wise framework, p is distributed uniformly on the (discrete) set $(\sqrt{d}\, e_i)_{i=1}^d \cup (-\sqrt{d}\, e_i)_{i=1}^d$, where e_i represents the i-th vector of the canonical basis. Then the first-order term in (4.1) is zero:

$$\mathbb{E}\left[\langle \nabla^2 f(x) p, p \rangle p\right] = \frac{1}{2d} \sum_{i=1}^d \langle \nabla^2 f(x) \sqrt{d}\, e_i, \sqrt{d}\, e_i \rangle \sqrt{d}\, e_i$$

$$- \frac{1}{2d} \sum_{i=1}^d \langle \nabla^2 f(x) \sqrt{d}\, e_i, \sqrt{d}\, e_i \rangle \sqrt{d}\, e_i = 0.$$

- The same holds for spherical smoothing, where p is distributed uniformly on the (continuous) set given by the sphere of radius $\sqrt{d}$ and centered at the origin, that we denote by ∂B. Also in this case, indeed,

$$\mathbb{E}\left[\langle \nabla^2 f(x) p, p \rangle p\right] = \int_{\partial B} \langle \nabla^2 f(x) p, p \rangle p \, dp = 0,$$

as we are integrating an odd function on a domain that is symmetric with respect to the origin. So, in the sufficiently regular case analysed above, the surrogate gradient proposed by coordinate-wise and spherical smoothing is unbiased up to the first-order term in h. But notice also that, in both cases, the term in h^2 is non-zero in general and the surrogate gradient is biased. This excludes the direct applicability of many theoretical studies for general stochastic algorithms such as [9] and motivates the introduction of Lemma 9.2, in which we bound a.s. the distance between the surrogate gradient and the projection of the exact one. On the other hand, the case of quadratic functions - in which both coordinate-wise and spherical smoothing are unbiased - deserves a tailored convergence analysis that we leave for future work. For refined properties of spherical smoothing and coordinate-wise descent, see [7] and Lemma 1, Remark 1 in [33].

4.3 A quasi-descent lemma

In the next proposition and corollary we obtain an a.s. estimate for the decrease of the objective function values. This result is the fundamental tool for the analysis in Section 6, and it is also of standalone interest.

Proposition 4.3. *Assume* **H.1**, **P.1** *and* **A.1** *and let the random sequence* (x_k) *be generated by Algorithm 1. Fix a constant w such that $0 < w < 2 - \Lambda\overline{\alpha}$ and $w \leq 1$. Define $C := (\ell\Lambda^2)/(8\min(1, 2 - \Lambda\overline{\alpha} - w))$. Then, for every $k \in \mathbb{N}$,*

$$f_{k+1} - f_k \leq -\frac{w\alpha_k}{2}\|P_k^\top \nabla f_k\|^2 + C\alpha_k h_k^2 \qquad a.s.$$

In particular, $f_{k+1} - f_k \leq C\alpha_k h_k^2$ a.s.

Before showing the proof of Proposition 4.3, we make some comments on the assumptions and conclusion and provide an important consequence under additional assumptions. First note that Proposition4.3 requires neither **P.2** nor **H.2**-**H.3**. The difference $f_{k+1} - f_k$ is bounded a.s. by the sum of two terms: a negative term depending on the squared norm of the projected gradient $P_k^\top \nabla f_k$ and an error term proportional to $\alpha_k h_k^2$.

Remark 4.4. Notice that the term mutiplying $\|P_k^\top \nabla f_k\|^2$ does not depend on h_k. A result analogous to Proposition 4.3 can be obtained for recursion (2.2), corresponding to the limit $h \to 0$. In this case, for every $k \in \mathbb{N}$,

$$f_{k+1} - f_k \leq -\frac{w\alpha_k}{2}\|P_k^\top \nabla f_k\|^2 \qquad a.s.$$

In particular, the objective function values are a.s. decreasing; namely, $f_{k+1} \leq f_k$ a.s. This fact is used later, in the proof of Theorem 5.12.

Finally, adding condition **A.4** to the hypothesis of Proposition 4.3, we get the following corollary.

Corollary 4.5. *Assume* **H.1**, **P.1**, **A.1** *and* **A.4**. *Let* (x_k) *be a random sequence generated by Algorithm 1. Let $0 < w < 2 - \Lambda\overline{\alpha}$ and $w \leq 1$. Then, for every $k \in \mathbb{N}$,*

$$f_{k+1} - f_k \leq -\frac{w\overline{\alpha}}{2}\|P_k^\top \nabla f_k\|^2 + C\overline{\alpha}\overline{h}^2 \qquad a.s. \tag{4.2}$$

Proof of Proposition 4.3. For notational simplicity, we denote P_k by P and set $e_k := \nabla_k f_k - P_k^\top \nabla f_k$. By Assumption **P.1**, for every $v \in \mathbb{R}^\ell$,

$$\|Pv\|^2 \overset{a.s.}{=} \frac{d}{\ell}\|v\|^2. \tag{4.3}$$

By Hypothesis **H.1** combined with Descent Lemma 9.2 and Algorithm 1, we obtain

$$f_{k+1} - f_k \leq \langle \nabla f_k, x_{k+1} - x_k \rangle + \frac{\lambda}{2}\|x_{k+1} - x_k\|^2$$

$$= -\alpha_k \langle P^\top \nabla f_k, \nabla_k f_k \rangle + \frac{\alpha_k^2 \lambda}{2}\|P\nabla_k f_k\|^2. \tag{4.4}$$

For the last term, we add and subtract $PP^\top \nabla f_k$ and recall the definition of e_k, to get

$$\frac{\alpha_k^2 \lambda}{2}\|P\nabla_k f_k\|^2 = \frac{\alpha_k^2 \lambda}{2}\|P\nabla_k f_k - PP^\top \nabla f_k + PP^\top \nabla f_k\|^2$$

$$= \frac{\alpha_k^2 \lambda}{2}\|Pe_k\|^2 + \frac{\alpha_k^2 \lambda}{2}\|PP^\top \nabla f_k\|^2 + \alpha_k^2 \lambda \langle Pe_k, PP^\top \nabla f_k \rangle.$$

Now, using again the algorithm and relation (4.3) in inequality (4.4), we have

$$
\begin{aligned}
f_{k+1} - f_k \le{}& -\alpha_k \langle P^\top \nabla f_k, \nabla_k f_k \rangle \\
& + \frac{\alpha_k^2 \lambda d}{2\ell} \|e_k\|^2 + \frac{\alpha_k^2 \lambda d}{2\ell} \|P^\top \nabla f_k\|^2 + \frac{\alpha_k^2 \lambda d}{\ell} \langle \nabla_k f_k - P^\top \nabla f_k, P^\top \nabla f_k \rangle \\
={}& \alpha_k (\Lambda \alpha_k - 1) \langle P^\top \nabla f_k, \nabla_k f_k \rangle + \frac{\Lambda \alpha_k^2}{2} \|e_k\|^2 - \frac{\Lambda \alpha_k^2}{2} \|P^\top \nabla f_k\|^2,
\end{aligned}
$$

where we recall that Λ is defined as $\Lambda := \lambda d/\ell$. Write $\nabla_k f_k = e_k + P^\top \nabla f_k$, to get

$$
\begin{aligned}
f_{k+1} - f_k \le{}& \alpha_k (\Lambda \alpha_k - 1) \langle P^\top \nabla f_k, e_k + P^\top \nabla f_k \rangle \\
& + \frac{\Lambda \alpha_k^2}{2} \|e_k\|^2 - \frac{\Lambda \alpha_k^2}{2} \|P^\top \nabla f_k\|^2 \\
={}& \alpha_k (\Lambda \alpha_k - 1) \langle P^\top \nabla f_k, e_k \rangle + \frac{\Lambda \alpha_k^2}{2} \|e_k\|^2 + \alpha_k \left(\frac{\Lambda \alpha_k}{2} - 1 \right) \|P^\top \nabla f_k\|^2.
\end{aligned}
$$

By Young's inequality with parameter τ_k and the estimate of $\|e_k\|$ from Lemma 4.1,

$$
\begin{aligned}
f_{k+1} - f_k \le{}& \frac{\alpha_k}{2\tau_k} |\Lambda \alpha_k - 1| \|P^\top \nabla f_k\|^2 + \frac{\alpha_k \tau_k}{2} |\Lambda \alpha_k - 1| \|e_k\|^2 + \frac{\Lambda \alpha_k^2}{2} \|e_k\|^2 \\
& + \alpha_k \left(\frac{\Lambda \alpha_k}{2} - 1 \right) \|P^\top \nabla f_k\|^2 \\
\le{}& -\alpha_k \left(1 - \frac{\Lambda \alpha_k}{2} - \frac{1}{2\tau_k} |\Lambda \alpha_k - 1| \right) \|P^\top \nabla f_k\|^2 \\
& + \frac{\alpha_k}{2} (\tau_k |\Lambda \alpha_k - 1| + \Lambda \alpha_k) \frac{\lambda^2 h_k^2 d^2}{4\ell}.
\end{aligned}
$$

If $\alpha_k \le 1/\Lambda$, choosing $\tau_k = 1$, we obtain

$$
\alpha_k \left(1 - \frac{\Lambda \alpha_k}{2} - \frac{1}{2\tau_k} |\Lambda \alpha_k - 1| \right) = \frac{\alpha_k}{2} \ge \frac{w \alpha_k}{2} \tag{4.5}
$$

and

$$
(\omega_k |\Lambda \alpha_k - 1| + \Lambda \alpha_k) \frac{\alpha_k \lambda^2 h_k^2 d^2}{8\ell} = \frac{\ell \Lambda^2}{8} \alpha_k h_k^2 \le C \alpha_k h_k^2. \tag{4.6}
$$

Similarly, if $1/\Lambda < \alpha_k \le \overline{\alpha} < 2/\Lambda$, choose $\tau_k = \frac{\Lambda \alpha_k - 1}{2 - \Lambda \alpha_k - w}$. Notice that $\tau_k > 0$ because $\Lambda \alpha_k - 1 > 0$ and

$$
0 < w < 2 - \Lambda \overline{\alpha} \le 2 - \Lambda \alpha_k.
$$

Then,

$$
\alpha_k \left(1 - \frac{\Lambda \alpha_k}{2} - \frac{1}{2\tau_k} |\Lambda \alpha_k - 1| \right) = \frac{w \alpha_k}{2},
$$

and, since $1 - w \Lambda \alpha_k \le 1$,

$$
(\tau_k |\Lambda \alpha_k - 1| + \Lambda \alpha_k) \frac{\alpha_k \lambda^2 h_k^2 d^2}{8\ell} = \left(\frac{1 - w \Lambda \alpha_k}{2 - \Lambda \alpha_k - w} \right) \frac{\ell \Lambda^2}{8} \alpha_k h_k^2 \le C \alpha_k h_k^2.
$$

$\square$

11

Remark 4.6. If the stepsize satisfies $\alpha_k \leq 1/\Lambda$ for every $k \in \mathbb{N}$ (instead of the weaker condition $\alpha_k < 2/\Lambda$ from Assumption **A.1**), it is clear from the proof that the result in Proposition 4.3 reduces to

$$f_{k+1} - f_k \leq -\frac{\alpha_k}{2} \|P_k^\top \nabla f_k\|^2 + \frac{\ell \Lambda^2}{8} \alpha_k h_k^2 \qquad a.s.$$

An analogous result holds for Corollary 4.5.

5 Convex case

In this section we study the case of convex objective functions. First, in Lemma 5.1 we derive a key preliminary result establishing a stochastic Fejér monotonicity property for our method. Using this Lemma we derive convergence results for different parameter choices. In Section 5.2 we consider the case of a non-vanishing discretization. Then, in Section 5.3, and 5.4 we consider two different regimes where the discretization is made finer as the iteration proceeds, considering both a more and a less aggressive strategy. Finally, in Section 5.5 we consider the limit case where no discretization is considered, namely recursion 2.2. Throughout this section we use the filtration $\mathcal{F}_k = \sigma(P_1, \ldots, P_{k-1})$.

5.1 Fejér monotonicity

We derive an a.s. energy estimate for the Lyapunov sequence $\left(\|x_k - x_*\|^2 \right)$, where (x_k) is the random sequence generated by Algorithm 1 and x_* is any minimizer of f.

Lemma 5.1. *Assume* **H.1**, **H.2** *and* **P.1**, **P.2**. *Let* (x_k) *be a random sequence generated by Algorithm 1. Then, for every* $k \in \mathbb{N}$ *and* $x_* \in \operatorname{argmin} f$,

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2 \,\middle|\, \mathcal{F}_k \right] - \left(1 + \frac{\alpha_k h_k d^2}{2\ell\sqrt{\ell}} \right) \|x_k - x_*\|^2$$

$$\overset{a.s.}{\leq} \quad D_k + \frac{2d}{\ell} \alpha_k^2 \|\nabla f_k\|^2 + 2\alpha_k \langle \nabla f_k, x_* - x_k \rangle,$$

where $\Lambda := \lambda d/\ell$ *and*

$$D_k := \frac{\Lambda^2}{2} \left[d \left(\alpha_k h_k \right)^2 + \sqrt{\ell} \left(\alpha_k h_k \right) \right]. \tag{5.1}$$

Lemma 5.1 is the fundamental inequality used in the remainder of this section. We combine it with another important ingredient, that is the cocoercivity of ∇f (see the Baillon-Haddad Theorem, Lemma 9.1) which allows us to show that the sequence $\left(\|x_k - x_*\|^2 \right)$ is stochastically Fejér monotone up to the error generated at each iteration by D_k. Under different assumptions on the stepsize and discretization sequences the sequence (x_k) exhibits different behaviors in terms of optimization and convergence rates.

Proof. Recall the notation $\nabla_k f_k := \nabla_{(P_k, h_k)} f(x_k)$. Lemma 4.1 provides an estimate for $\|e_k\|$, where $e_k := \nabla_k f_k - P_k \nabla f(x_k)$. By Assumption **P.1**, for every $v \in \mathbb{R}^\ell$,

$$\|P_k v\|^2 \overset{a.s.}{=} \frac{d}{\ell} \|v\|^2 ; \tag{5.2}$$

12

and, by **P.2**, for every fixed $w \in \mathbb{R}^d$,

$$\mathbb{E}\left[\left\|P_k^\top w\right\|^2 \mid \mathcal{F}_k\right] = \|w\|^2. \tag{5.3}$$

Thus, for every $x_* \in \arg\min f$ and every $k \in \mathbb{N}$, the following equality holds a.s.

$$
\begin{aligned}
&\|x_{k+1} - x_*\|^2 - \|x_k - x_*\|^2 \\
=\quad &\|x_{k+1} - x_k\|^2 + 2\langle x_{k+1} - x_k, x_k - x_*\rangle \\
(\text{Alg. } 1) \quad =\quad &\alpha_k^2 \|P_k \nabla_k f_k\|^2 - 2\alpha_k \langle P_k \nabla_k f_k, x_k - x_*\rangle \\
=\quad &\alpha_k^2 \left\|P_k \nabla_k f_k - P_k P_k^\top \nabla f_k + P_k P_k^\top \nabla f_k\right\|^2 + 2\alpha_k \langle \nabla f_k, x_* - x_k\rangle \\
&+ 2\alpha_k \langle P_k P_k^\top \nabla f_k - \nabla f_k, x_* - x_k\rangle \\
&+ 2\alpha_k \langle P_k \nabla_k f_k - P_k P_k^\top \nabla f_k, x_* - x_k\rangle.
\end{aligned}
$$

Using the fact that $(x + y)^2 \leq 2x^2 + 2y^2$ and the Cauchy-Schwarz inequality, we get

$$
\begin{aligned}
&\|x_{k+1} - x_*\|^2 - \|x_k - x_*\|^2 \\
\leq\quad &2\alpha_k^2 \|P_k e_k\|^2 + 2\alpha_k^2 \left\|P_k P_k^\top \nabla f_k\right\|^2 + 2\alpha_k \langle \nabla f_k, x_* - x_k\rangle \\
&+ 2\alpha_k \langle P_k P_k^\top \nabla f_k - \nabla f_k, x_* - x_k\rangle + 2\alpha_k \|P_k e_k\| \, \|x_* - x_k\|.
\end{aligned}
$$

Then, by **P.1** and Lemma 4.1,

$$
\begin{aligned}
&\|x_{k+1} - x_*\|^2 - \|x_k - x_*\|^2 \\
\leq\quad &\frac{2\alpha_k^2 d \left(\lambda h_k d\right)^2}{4\ell^2} + \frac{2\alpha_k^2 d}{\ell} \left\|P_k^\top \nabla f_k\right\|^2 + 2\alpha_k \langle \nabla f_k, x_* - x_k\rangle \\
&+ 2\alpha_k \langle P_k P_k^\top \nabla f_k - \nabla f_k, x_* - x_k\rangle + \frac{\alpha_k \lambda h_k d^2}{\ell \sqrt{\ell}} \|x_* - x_k\|.
\end{aligned} \tag{5.4}
$$

Finally,

$$
\begin{aligned}
&\|x_{k+1} - x_*\|^2 - \|x_k - x_*\|^2 \\
\leq\quad &\frac{d^3 \lambda^2}{2\ell^2} \left(\alpha_k h_k\right)^2 + \frac{2d}{\ell} \alpha_k^2 \left\|P_k^\top \nabla f_k\right\|^2 + 2\alpha_k \langle \nabla f_k, x_* - x_k\rangle \\
&+ 2\alpha_k \langle P_k P_k^\top \nabla f_k - \nabla f_k, x_* - x_k\rangle + \frac{\lambda^2 d^2 \left(\alpha_k h_k\right)^2}{2\ell \sqrt{\ell} \alpha_k h_k} + \frac{\alpha_k h_k d^2}{2\ell \sqrt{\ell}} \|x_* - x_k\|^2,
\end{aligned}
$$

where we used Young inequality with parameter $(\alpha_k h_k d^2)/(\ell \sqrt{\ell})$. Taking the conditional expectation given $\mathcal{F}_k$ and using Assumption **P.2**, we get the claim. For integrability considerations, see Remark 9.12. $\qquad\square$

Given the above inequality, we next derive a number of different convergence results considering different choices of the discretization parameter and the stepsize.

13

5.2 Non-vanishing dicretization

Using an intermediate estimate from the proof of Lemma 5.1, we get the following result: with only a boundedness assumption on the discretization sequences, we derive an upper bound on the expectation of the function values for the best iterate.

Theorem 5.2. *Assume* **H.1**, **H.2**, **P.1**, **P.2** *and* **A.2**. *Let* (x_k) *be a random sequence generated by Algorithm 1. Then, for every* $k \in \mathbb{N}$,

$$\min_{j \in [k]} \mathbb{E}(f_j - f_*) \leq \frac{\tilde{D}_k}{\sum_{j=0}^k \alpha_j},$$

where

$$\tilde{D}_k := \frac{d + \xi \ell}{2\xi \ell} \left[S_k + \sum_{j=0}^k \rho_j \left(\sqrt{S_{j-1}} + \sum_{i=0}^j \rho_i \right) \right],$$

with $\xi := \frac{1}{\lambda} - \frac{d\bar{\alpha}}{\ell} > 0$, $\rho_i := \frac{\lambda d^2}{\ell \sqrt{\ell}} (\alpha_i h_i)$ *and* $S_j := \|x_0 - x_*\|^2 + \sum_{i=0}^j \frac{\Lambda^2}{2} (\alpha_i h_i)^2$. *In particular, for* $\alpha_k = \alpha$ *and* $h_k \leq \bar{h}$, *we get*

$$\min_{j \in [k]} \mathbb{E}(f_j - f_*) \leq \max\{C_1/(k+1) + C_2\bar{h}^2 + C_3\bar{h},\ C_2\bar{h}^2 + C_4\bar{h}\},$$

where we made explicit only the dependence on the iteration number k *and the bound on the discretization error* $\bar{h}$, *while* C_1, C_2 *and* C_3 *are appropriate constants derived from the proof.*

The above result is an extended version of Theorem 3.1(i). As mentioned after Theorem 3.1, it suggests that the accuracy will stop improving after a given number of iterations depending on the discretization level. The proof follows.

Proof. Start from the inequality in 5.4. Recalling that we defined $\rho_j := \frac{\lambda d^2}{\ell \sqrt{\ell}} (\alpha_j h_j)$, for every $j \in \mathbb{N}$ we have that

$$
\begin{aligned}
\|x_{j+1} - x_*\|^2 - \|x_j - x_*\|^2 &\leq \frac{\Lambda^2}{2} (\alpha_j h_j)^2 + \frac{2d}{\ell} \alpha_j^2 \left\| P_j^\top \nabla f_j \right\|^2 + 2\alpha_j \langle \nabla f_j, x_* - x_j \rangle \\
&\quad + 2\alpha_j \langle P_j P_j^\top \nabla f_j - \nabla f_j, x_* - x_j \rangle + \rho_j \|x_* - x_j\| \\
\text{(Baillon-Haddad Theorem 9.1)} \quad &\leq \frac{\Lambda^2}{2} (\alpha_j h_j)^2 + \frac{2d}{\ell} \alpha_j \left\| P_j^\top \nabla f_j \right\|^2 - \frac{2}{\lambda} \alpha_j \|\nabla f_j\|^2 \\
&\quad + 2\alpha_j \langle P_j P_j^\top \nabla f_j - \nabla f_j, x_* - x_j \rangle + \rho_j \|x_* - x_j\|.
\end{aligned}
$$

Taking expectations in the previous bound and denoting $u_j := \sqrt{\mathbb{E} \|x_j - x_*\|^2}$, we get that for every $j \in \mathbb{N}$

$$
\begin{aligned}
u_{j+1}^2 - u_j^2 &\leq \frac{\Lambda^2}{2} (\alpha_j h_j)^2 + \frac{2d}{\ell} \alpha_j^2 \mathbb{E} \left[\|\nabla f_j\|^2 \right] - \frac{2}{\lambda} \alpha_j \mathbb{E} \left[\|\nabla f_j\|^2 \right] + \rho_j \mathbb{E} \sqrt{\|x_* - x_j\|^2} \\
&\leq \frac{\Lambda^2}{2} (\alpha_j h_j)^2 - 2\xi \alpha_j \mathbb{E} \left[\|\nabla f_j\|^2 \right] + \rho_j u_j,
\end{aligned}
$$

where we recall that $\xi := \frac{1}{\lambda} - \frac{d\alpha}{\ell} > 0$. Then, summing from $j = 0$ to $j = k$,

$$u_{k+1}^2 + 2\xi \sum_{j=0}^{k} \alpha_j \mathbb{E}\left[\|\nabla f_j\|^2\right] \le \underbrace{u_0^2 + \sum_{j=0}^{k} \frac{\Lambda^2}{2}(\alpha_j h_j)^2}_{=S_k} + \sum_{j=0}^{k} \rho_j u_j,$$

and

$$\sum_{j=0}^{k} \alpha_j \mathbb{E}\left[\|\nabla f_j\|^2\right] \le \frac{1}{2\xi}\left[S_k + \sum_{j=0}^{k} \rho_j u_j\right] \tag{5.5}$$

and, by discrete Bihari's Lemma 9.8, for every $j \in \mathbb{N}$

$$u_j \le \frac{1}{2}\sum_{i=0}^{j} \rho_i + \left[S_{j-1} + \left(\frac{1}{2}\sum_{i=0}^{j} \rho_i\right)^2\right]^{1/2} \le \sqrt{S_{j-1}} + \sum_{i=0}^{j} \rho_i. \tag{5.6}$$

Starting again from the intermediate inequality in 5.4, for every $j \in \mathbb{N}$,

$$\|x_{j+1} - x_*\|^2 - \|x_j - x_*\|^2 \le \frac{\Lambda^2}{2}(\alpha_j h_j)^2 + \frac{2d}{\ell}\alpha_j^2\left\|P_j^\top \nabla f_j\right\|^2 + 2\alpha_j\langle\nabla f_j, x_* - x_j\rangle$$
$$+ 2\alpha_j\langle P_j P_j^\top \nabla f_j - \nabla f_j, x_* - x_j\rangle + \rho_j\|x_* - x_j\|$$
$$(\text{convexity}) \le \frac{\Lambda^2}{2}(\alpha_j h_j)^2 + \frac{2d}{\ell}\alpha_j\left\|P_j^\top \nabla f_j\right\|^2 - 2\alpha_j(f_j - f_*)$$
$$+ 2\alpha_j\langle P_j P_j^\top \nabla f_j - \nabla f_j, x_* - x_j\rangle + \rho_j\|x_* - x_j\|.$$

Taking expectations, we get

$$u_{j+1}^2 - u_j^2 + 2\alpha_j \mathbb{E}(f_j - f_*) \le \frac{\Lambda^2}{2}(\alpha_j h_j)^2 + \frac{2d}{\ell}\alpha_j \mathbb{E}\left[\|\nabla f_j\|^2\right] + \rho_j u_j.$$

Summing from $j = 0$ to $j = k$, we conclude the first claim:

$$u_{k+1}^2 + 2\sum_{j=0}^{k} \alpha_j \mathbb{E}(f_j - f_*) \le S_k + \frac{2d}{\ell}\sum_{j=0}^{k} \alpha_j \mathbb{E}\left[\|\nabla f_j\|^2\right] + \sum_{j=0}^{k} \rho_j u_j$$

$$(5.5) \quad \le S_k + \frac{d}{\xi\ell}\left[S_k + \sum_{j=0}^{k} \rho_j u_j\right] + \sum_{j=0}^{k} \rho_j u_j$$

$$= \frac{d + \xi\ell}{\xi\ell}\left[S_k + \sum_{j=0}^{k} \rho_j u_j\right]$$

$$(5.6) \quad \le \frac{d + \xi\ell}{\xi\ell}\left[S_k + \sum_{j=0}^{k} \rho_j\left(\sqrt{S_{j-1}} + \sum_{i=0}^{j} \rho_i\right)\right]$$

$$= 2\tilde{D}_k.$$

15

For the second claim, for $\alpha_k = \alpha$ and $h_k \leq \overline{h}$, define $\overline{\rho} := \frac{\lambda d^2}{\ell \sqrt{\ell}} \left(\alpha \overline{h}\right)$ and note that

$$\tilde{D}_k \leq \frac{d + \xi\ell}{2\xi\ell} \left[\|x_0 - x_*\|^2 + \sum_{j=0}^{k} \frac{\Lambda^2}{2} (\alpha_j h_j)^2 + \|x_0 - x_*\| \sum_{j=0}^{k} \rho_j \right.$$

$$\left. + \sum_{j=0}^{k} \rho_j \left(\sum_{i=0}^{j} \frac{\Lambda}{\sqrt{2}} (\alpha_i h_i) \right) + \sum_{j=0}^{k} \rho_j \left(\sum_{i=0}^{j} \rho_i \right) \right]$$

$$\leq \frac{d + \xi\ell}{2\xi\ell} \left[\|x_0 - x_*\|^2 + \frac{\Lambda^2}{2} \alpha^2 \overline{h}^2 \sum_{j=0}^{k} 1 + \overline{\rho} \|x_0 - x_*\| \sum_{j=0}^{k} 1 \right.$$

$$\left. + \frac{\Lambda}{\sqrt{2}} \alpha \overline{h} \overline{\rho} \sum_{j=0}^{k} \sum_{i=0}^{j} 1 + \overline{\rho}^2 \sum_{j=0}^{k} \sum_{i=0}^{j} 1 \right]$$

$$\leq \tilde{C}_1 + \tilde{C}_2 \overline{h}^2 (k+1) + \tilde{C}_3 \overline{h}(k+1) + \tilde{C}_4 \overline{h}^2 (k+1)(k+2),$$

where $\tilde{C}_1, ..., \tilde{C}_4$ are appropriate constants. Recalling that α_k is assumed to be constant and so that $\sum_{j=0}^{k} \alpha_j = \alpha(k+1)$, by trivial manipulations we get that

$$\min_{j \in [k]} \mathbb{E}(f_j - f_*) \leq \frac{\bar{C}_1}{k+1} + \bar{C}_2(h + \overline{h}^2) + \bar{C}_3 \overline{h}^2 k,$$

where $\bar{C}_1, \bar{C}_2$ and $\bar{C}_3$ are again appropriate constants. We conclude simply by noticing that, for $k \leq 1/\overline{h}$, the right hand side is bounded by $\bar{C}_1/(k+1) + \bar{C}_2(h + \overline{h}^2) + \bar{C}_3 \overline{h}$; while, for $k > 1/\overline{h}$, the sequence $\min_{j \in [k]} \mathbb{E}(f_j - f_*)$ is non-increasing in k and so controlled by the bound at $k = 1/\overline{h}$, that is $\bar{C}_1/(1/\overline{h} + 1) + \bar{C}_2(h + \overline{h}^2) + \bar{C}_3 \overline{h} \leq (\bar{C}_1 + \bar{C}_2 + \bar{C}_3)\overline{h} + \bar{C}_2 \overline{h}^2$. $\qquad\square$

Remark 5.3. The case of a single random direction at each iterations ($\ell = 1$) sampled from a normal distribution is studied in [39, 40]. Under Hypothesis **H.1** and **H.2**, [40, Theorem 8] states the following result: taking constant $\alpha_k = \frac{1}{4\lambda(d+4)}$ and $h_k = h$,

$$\min_{j \in [k]} \mathbb{E}\left[f(x_j) - f_*\right] \leq \frac{4\lambda(d+4)\|x_0 - x_*\|^2}{k} + \frac{9\lambda h^2 (d+4)^2}{25}. \tag{5.7}$$

Comparing this bound with the one obtained in Theorem 5.2, we see that this is tighter. The main difference between the two approaches is due to the sampling of the random direction. The one adopted in [40], is such that the expectation of the finite difference approximation of the directional derivative is the gradient of a (Gaussian) smoothing of f, while such a property does not hold under our assumptions, and a different proof is needed. We will see in Remark 5.9, that the different bounds lead to very similar results in terms of accuracy if the discretization error in our method is allowed to go to zero.

Next, we develop our analysis for the case of a decreasing sequence of discretization parameters (h_k) which allows for a finer discretization and an increasingly accurate approximation of the exact gradient.

5.3 Basic results with coarser discretization

We begin considering very mild assumptions on the speed at which the discretization sequence vanishes. Using Lemma 5.1 we prove a.s. convergence for the function value of the best iterate, as well as a sublinear rate in expectation. The following result is an extended version of item (ii) in Theorem 3.1.

Theorem 5.4. *Assume* **H.1**, **H.2**, **P.1**, **P.2**, **A.2** *and* **A.3**. *Assume also that* $(\alpha_k h_k) \in \ell^1$. *Let* (x_k) *be a random sequence generated by Algorithm 1. Then,*

$$\lim_k f(x_k) \overset{a.s.}{=} f_*.$$

Moreover, we have the following convergence rate for the best iterate in expectation:

$$\min_{j \in [k]} \mathbb{E}\left[f_j - f_*\right] \quad \leq \quad \frac{D}{\sum_{j=0}^{k} \alpha_j},$$

where the constant $D > 0$ is provided in the proof.

Example 5.5. For every $k \in \mathbb{N}$ let $h_k = h/k^r$ and $\alpha_k = \alpha/k^s$. Let $\alpha \in (0, 1/\Lambda)$, $h > 0$ and $s \geq 0$ with $r > 1 - s \geq 0$. Then $(\alpha_k) \notin \ell^1$, $(\alpha_k h_k) \in \ell^1$, so the assumptions of Theorem 5.4 are satisfied. For example, the latter holds for $\alpha_k = \alpha/k$ and $h_k = h/k^r$ with $r > 0$ (α_k vanishing and h_k going to zero arbitrarily slow); or for $\alpha_k = \alpha$ and $h_k = h/k^r$ with $r > 1$ (α_k constant and h_k going to zero sufficiently fast).

Proof. Consider the inequality from Lemma 5.1, namely

$$\mathbb{E}_k \|x_{k+1} - x_*\|^2 - \left(1 + \frac{\alpha_k h_k d^2}{2\ell\sqrt{\ell}}\right) \|x_k - x_*\|^2$$

$$\leq D_k + \frac{2d}{\ell}\alpha_k^2 \|\nabla f_k\|^2 + 2\alpha_k\langle\nabla f_k, x_* - x_k\rangle$$

$$(\text{cocoercivity of } \nabla f, \text{ see Thm. 9.1}) \ \leq D_k + \frac{2d}{\ell}\alpha_k^2 \|\nabla f_k\|^2 - \frac{2\alpha_k}{\lambda} \|\nabla f_k\|^2 \tag{5.8}$$

$$(\textbf{A.2}) \ \leq D_k + 2\alpha_k\left(\frac{d\overline{\alpha}}{\ell} - \frac{1}{\lambda}\right)\|\nabla f_k\|^2$$

$$= D_k - 2\xi\alpha_k\|\nabla f_k\|^2,$$

where we defined $\xi := \frac{1}{\lambda} - \frac{d\overline{\alpha}}{\ell} > 0$. By the assumptions, $(\alpha_k h_k) \in \ell^1$ and so $(D_k) \in \ell^1$. Using Robbins-Siegmund Lemma 9.11, we know that $(\|x_k - x_*\|)$ is a.s. convergent for every $x_* \in \text{argmin}\, f$ and that a.s.

$$\left(\alpha_k \|\nabla f_k\|^2\right) \in \ell^1.$$

By Lemma 5.1,

$$\mathbb{E}_k\left[\|x_{k+1} - x_*\|^2\right] - \left(1 + \frac{\alpha_k h_k d^2}{2\ell\sqrt{\ell}}\right)\|x_k - x_*\|^2$$

$$\leq D_k + \frac{2d}{\ell}\alpha_k^2 \|\nabla f_k\|^2 + 2\alpha_k\langle\nabla f_k, x_* - x_k\rangle \tag{5.9}$$

$$(\text{convexity of } f \text{ and } \textbf{A.2}) \ \leq D_k + \frac{2d\overline{\alpha}}{\ell}\alpha_k \|\nabla f_k\|^2 - 2\alpha_k\left(f_k - f_*\right).$$

17

By the Robbins-Siegmund Lemma 9.11, we get that a.s.

$$(\alpha_k (f_k - f_*)) \in \ell^1. \tag{5.10}$$

We know that $f_k \geq f_*$ and, by Assumption A.3, the sequence (α_k) is positive and does not belong to ℓ^1. So,

$$\liminf_k f(x_k) \overset{\text{a.s.}}{=} f_*. \tag{5.11}$$

Recall that, from Proposition 4.3, we have that

$$f_{k+1} - f_k \leq C \alpha_k h_k^2 \qquad a.s.$$

By the assumptions $(\alpha_k h_k) \in \ell^1$ and h_k bounded, we know that also $(\alpha_k h_k^2)$ belongs to ℓ^1. Then, from Lemma 9.7, f_k is a.s. convergent. This implies, joint with 5.11, that $\lim_k f(x_k) \overset{\text{a.s.}}{=} f_*$.
For the convergence rate, first take the total expectation in (5.8),

$$\mathbb{E} \|x_{j+1} - x_*\|^2 - \left(1 + \frac{\alpha_k h_k d^2}{2\ell\sqrt{\ell}}\right) \mathbb{E} \|x_j - x_*\|^2 + 2\xi\alpha_j \mathbb{E} \|\nabla f_j\|^2 \leq D_j.$$

Applying Lemma 9.7 to the deterministic sequence $\left(\mathbb{E}\left[\|x_k - x_*\|^2\right]\right)$ we get convergence for every $x_* \in \arg\min f$ (and so the sequence is bounded above by some constant $C(x_*)$). Moreover, summing from $j = 0$ to $j = k$,

$$\begin{aligned}
\sum_{j=0}^k \alpha_j \mathbb{E} \|\nabla f_j\|^2 &\leq \frac{1}{2\xi} \sum_{j=0}^k \left(\mathbb{E} \|x_j - x_*\|^2 - \mathbb{E} \|x_{j+1} - x_*\|^2\right) \\
&\quad + \frac{C(x_*)d^2}{4\xi\ell\sqrt{\ell}} \sum_{j=0}^k \alpha_j h_j + \frac{1}{2\xi} \sum_{j=0}^k D_j \\
&= \frac{1}{2\xi} \left(\mathbb{E} \|x_0 - x_*\|^2 - \mathbb{E} \|x_{k+1} - x_*\|^2\right) \\
&\quad + \frac{C(x_*)d^2}{4\xi\ell\sqrt{\ell}} \sum_{j=0}^k \alpha_j h_j + \frac{1}{2\xi} \sum_{j=0}^k D_j \\
&\leq \frac{1}{2\xi} \left(\|x_0 - x_*\|^2 + \frac{C(x_*)d^2}{2\ell\sqrt{\ell}} \sum_{j=0}^{+\infty} \alpha_j h_j + \sum_{j=0}^{+\infty} D_j\right).
\end{aligned} \tag{5.12}$$

Taking the total expectation in inequality (5.9) and recalling that $\mathbb{E}\left[\|x_k - x_*\|^2\right]$ is bounded above by some constant $C(x_*)$, we get that, for every $j \in \mathbb{N}$,

$$\begin{aligned}
\alpha_j \, \mathbb{E} \left(f_j - f_*\right) &\leq \frac{1}{2}\mathbb{E} \|x_j - x_*\|^2 - \frac{1}{2}\mathbb{E} \|x_{j+1} - x_*\|^2 + \frac{C(x_*)d^2}{4\ell\sqrt{\ell}} \, \alpha_j h_j \\
&\quad + \frac{D_j}{2} + \frac{d\overline{\alpha}}{\ell}\alpha_j \mathbb{E} \|\nabla f_j\|^2.
\end{aligned}$$

18

Summing from $j = 0$ to $j = k$,

$$\sum_{j=0}^{k} \alpha_j \, \mathbb{E} \left(f_j - f_* \right)$$

$$\leq \quad \frac{1}{2} \mathbb{E} \left\| x_0 - x_* \right\|^2 - \frac{1}{2} \mathbb{E} \left\| x_{k+1} - x_* \right\|^2 + \frac{C(x_*) d^2}{4\ell\sqrt{\ell}} \sum_{j=0}^{k} \alpha_j h_j$$

$$+ \frac{1}{2} \sum_{j=0}^{k} D_j + \frac{d\overline{\alpha}}{\ell} \sum_{j=0}^{k} \alpha_j \mathbb{E} \left\| \nabla f_j \right\|^2$$

$$(5.12) \quad \leq \quad \frac{1}{2} \left\| x_0 - x_* \right\|^2 + \frac{C(x_*) d^2}{4\ell\sqrt{\ell}} \sum_{j=0}^{+\infty} \alpha_j h_j + \frac{1}{2} \sum_{j=0}^{+\infty} D_j$$

$$+ \frac{d\overline{\alpha}}{2\xi\ell} \left(\left\| x_0 - x_* \right\|^2 + \frac{C(x_*) d^2}{2\ell\sqrt{\ell}} \sum_{j=0}^{+\infty} \alpha_j h_j + \sum_{j=0}^{+\infty} D_j \right)$$

$$= \quad \frac{\xi\ell + d\overline{\alpha}}{2\xi\ell} \left(\left\| x_0 - x_* \right\|^2 + \frac{C(x_*) d^2}{2\ell\sqrt{\ell}} \sum_{j=0}^{+\infty} \alpha_j h_j \right.$$

$$\left. + \frac{\Lambda^2}{2} \sum_{j=0}^{+\infty} \left[d \left(\alpha_j h_j \right)^2 + \sqrt{\ell} \left(\alpha_j h_j \right) \right] \right)$$

$$\leq \quad \frac{1}{2 \left(1 - \Lambda\overline{\alpha} \right)} \left(\left\| x_0 - x_* \right\|^2 + \frac{1}{2} \left(\frac{C(x_*) d^2}{\ell\sqrt{\ell}} + \Lambda d\overline{h} + \Lambda^2 \sqrt{\ell} \right) \sum_{j=0}^{+\infty} \alpha_j h_j \right)$$

$$=: D < +\infty.$$

We obtain the bound by noticing that

$$\min_{j \in [k]} \mathbb{E} \left[f_j - f_* \right] \sum_{j=0}^{k} \alpha_j \leq \sum_{j=0}^{k} \alpha_j \, \mathbb{E} \left(f_j - f_* \right).$$

$\square$

5.4 Improved results with finer discretization

We next make stronger assumptions on the sequences (α_k) and (h_k) allowing us to derive an a.s. convergence result for the function values and a.s. convergence of the iterates to a solution. The following result is an extended version of Theorem 3.1 (iii).

Theorem 5.6. *Under the same conditions as in Theorem 5.4, but with the stronger Assumption **A.4** instead of **A.3**. Namely, assume **H.1**, **H.2**, **P.1**, **P.2** and $0 < \underline{\alpha} \leq \alpha_k \leq \overline{\alpha} < 1/\Lambda$ and $(h_k) \in \ell^1$. Let (x_k) be a sequence generated by Algorithm 1. Then,*

$$\lim_k f_k \overset{a.s.}{=} f_*$$

and there exists a random variable x_* with values in $\operatorname{argmin} f$ such that $x_k \overset{a.s.}{\to} x_*$.

Example 5.7. For every $k \in \mathbb{N}$, let $\alpha_k = \alpha$ constant in $(0, 1/\Lambda)$ and $h_k = h/k^r$ with $h > 0$ and $r > 1$. Then the assumptions of Theorem 5.6 hold. These conditions are a special case of those in Example 5.5. In general, under Assumption **A.4** (required for Theorem 5.6), the stepsize α_k is uniformly bounded below by a strictly positive constant and so it can not converge to zero. Then, to get the condition $(\alpha_k h_k) \in \ell^1$, h_k can not converge to zero arbitrarily slowly as in Example 5.5. Indeed, for Theorem 5.6 to hold with h_k of the form h/k^r, h_k has to converge to zero strictly faster than $1/k$.

Remark 5.8. Under the assumptions in Theorem 5.6, the convergence rate in Theorem 5.4 holds and reads as

$$\min_{j \in [k]} \mathbb{E}\left[f_j - f_*\right] \quad \leq \quad \frac{D}{\underline{\alpha} k}.$$

Proof. From the proof of the previous theorem, we see that for every $x_* \in \operatorname{argmin} f$ the sequence $(\|x_k - x_*\|)$ is a.s. convergent and a.s. $(\alpha_k (f_k - f_*)) \in \ell^1$. From Assumption **A.4**, $0 < \underline{\alpha} \leq \alpha_k$ and so the sequence $(f_k - f_*)$ is non-negative and belongs to ℓ^1 a.s. In particular,

$$\lim_k f_k \overset{\text{a.s.}}{=} f_*.$$

More precisely, there is a $\bar{\Omega} \subseteq \Omega$ with $\mathbb{P}(\bar{\Omega}) = 1$ such that, for every $\omega \in \bar{\Omega}$,

$$\lim_k f(x_k(\omega)) = f_*. \tag{5.13}$$

For $\omega \in \bar{\Omega}$, let $x_{k_j}(\omega)$ be a convergent subsequence of $x_k(\omega)$; say $x_{k_j}(\omega) \to x_\infty$. Then, by continuity of the function f and the limit in (5.13),

$$f(x_\infty) = \lim_k f(x_k(\omega)) = f_*.$$

Then $x_\infty \in \operatorname{argmin} f$, as it is a minimizer of f. Summarizing, there is a full measure set for which every cluster point of the random sequence (x_k) belongs to $\operatorname{argmin} f$. Finally, combining the latter result with the fact that, for every $x_* \in \operatorname{argmin} f$, $(\|x_k - x_*\|)$ is a.s. convergent, the stochastic version of Opial's Lemma 9.10 guarantees the existence of a random variable x_* with values in $\operatorname{argmin} f$ such that $x_k \overset{\text{a.s.}}{\to} x_*$. $\square$

In the next remarks we compare our rates on the objective function with results available in the literature. Recall that none of the considered papers prove the convergence of the iterates.

Remark 5.9. We compare our results for a vanishing discretization with the ones obtained in [40, Theorem 8] for a single direction sampled according to a normal distribution (see also Remark 5.3). Choosing $\alpha_k = \alpha$ constant in $(0, 1/\Lambda)$ and $h_k = h/k^r$ with $h > 0$ and $r > 1$, from Remark 5.8 we get that

$$\min_{j \in [k]} \mathbb{E}\left[f(x_j) - f_*\right] \leq \frac{D}{\alpha k},$$

where

$$D := \frac{1}{2(1 - \Lambda\alpha)} \left(\|x_0 - x_*\|^2 + \frac{\alpha}{2} \left(\frac{C(x_*)d^2}{\ell\sqrt{\ell}} + \Lambda d h + \Lambda^2 \sqrt{\ell} \right) \sum_{j=0}^{+\infty} h_j \right).$$

For the special case $\alpha = \ell/(2\lambda d)$ and $\ell = 1$ we derive (recalling that $\sum_{j=0}^{+\infty} 1/k^r = \zeta(r) < +\infty$ where ζ is the Riemann zeta function),

$$\min_{j \in [k]} \mathbb{E}\left[f(x_j) - f_*\right] \le \frac{2\lambda d \|x_0 - x_*\|^2}{k} + \frac{hd^2\zeta(r)}{2k}\left(C(x_*) + \lambda h + \lambda^2\right). \tag{5.14}$$

Comparing equations (5.7) and (5.14) we see that the dependence on the dimension is the same however our result converges to the optimum because we chose a decreasing discretization parameter. In addition, we are free to choose the stepsize bigger than the one proposed in [39, 40] resulting in slightly better constants. On the other hand, in [40], they also study the case of accelerated inertial algorithms. A similar comparison to the one above can be done also with the results in [23].

Remark 5.10. The minimization of a smooth function via a zeroth-order oracle is also considered in [20]. The assumptions in that paper are different from ours, both in terms of properties of the objective function, as well as of the available zeroth-order oracle. Regarding the objective function, in addition to the Lipschitz continuity of the gradient, the authors of [20] require more restrictive assumptions, such as boundedness of the gradient itself on the entire feasible set, which is assumed to be compact. The zeroth-order oracle instead is more general than ours, and consists of noisy function evaluations. In their setting G denotes the bound on the gradient of f, and R is the diameter of the feasible set. With the choice

$$\alpha_k = \frac{\alpha R}{2G\sqrt{d/\ell}\sqrt{k}} \qquad \text{and} \qquad h_k = \frac{uG}{\lambda d^{3/2} k}, \tag{5.15}$$

for some $\alpha > 0$ and $u > 0$, they derive a bound of the form

$$\min_{j \in [k]} \mathbb{E}\left[f(x_j) - f_*\right] \le \frac{5RG\sqrt{1 + d/\ell}}{\sqrt{k}}\left(\max\{\alpha, \alpha^{-1}\} + \frac{\alpha u^2}{\sqrt{k}} + \frac{u\log(2k)}{k}\right). \tag{5.16}$$

If we choose $\alpha_k = \ell/(2\lambda d)$ and $h_k = \frac{h}{\lambda d^{3/2} k^r}$ with $r > 1$, we get

$$\min_{j \in [k]} \mathbb{E}\left[f(x_j) - f_*\right] \le \frac{2\lambda d \|x_0 - x_*\|^2}{\ell k} + \frac{H}{2k}\left(\frac{\lambda^2 C(x_*)\sqrt{d}}{\ell\sqrt{\ell}} + \frac{h}{\sqrt{d}\ell} + \frac{\lambda\sqrt{d}}{\ell\sqrt{\ell}}\right) \tag{5.17}$$

Comparing (5.16) and (5.17), we observe that we obtain a better convergence rate, due to the fact that we consider a noise-free oracle, but our analysis leads to a worse dependence on the ratio d/ℓ. Since the two settings are very different the significance of the comparison is somewhat limited.

Finally, in the next section, we consider the case where $h = 0$. i.e. Recursion (2.2).

5.5 Convergence results for recursion (2.2)

This section covers the special case of recursion (2.2), corresponding to the limiting case of Algorithm 1 when exact directional derivatives are available. Lemma 5.11 provides a sharper energy estimate than the one in Lemma 5.1, which in turns leads to the improved convergence results of Theorem 5.12. The result is an extended version of Theorem 3.1(iv).

Lemma 5.11. *Assume* **H.1**, **H.2**, **P.1** *and* **P.2**. *Let* (x_k) *be a random sequence generated by recursion* (2.2). *Then, for every* $k \in \mathbb{N}$ *and every* $x_* \in \arg\min f$,

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2 \,\big|\, \mathcal{F}_k\right] - \|x_k - x_*\|^2 \stackrel{a.s.}{=} \frac{\alpha_k^2 d}{\ell}\|\nabla f_k\|^2 + 2\alpha_k\langle\nabla f_k, x_* - x_k\rangle.$$

21

Proof. For every $k \in \mathbb{N}$ and every $x_* \in \operatorname{argmin} f$, we have that a.s.

$$\|x_{k+1} - x_*\|^2 - \|x_k - x_*\|^2 = \|x_{k+1} - x_k\|^2 + 2\langle x_{k+1} - x_k, x_k - x_* \rangle$$

$$(2.2) \quad = \alpha_k^2 \left\| P_k P_k^\top \nabla f_k \right\|^2 - 2\alpha_k \langle P_k P_k^\top \nabla f_k, x_k - x_* \rangle$$

$$(\textbf{P.1}, \text{ see } (5.2)) \quad = \frac{\alpha_k^2 d}{\ell} \langle \nabla f_k, P_k P_k^\top \nabla f_k \rangle + 2\alpha_k \langle P_k P_k^\top \nabla f_k, x_* - x_k \rangle.$$

The claim follows taking the conditional expectation given $\mathcal{F}_k$ and using **P.2**. $\qquad\square$

Using the estimate from the above lemma and with very mild assumptions on α_k, we get the following result ensuring convergence of the iterates, a rate in expectation for the function values and an asymptotic a.s. convergence rate of the form $1/k$.

Theorem 5.12. *Assume* **H.1**, **H.2**, **P.1**, **P.2**, **A.1** *and* **A.3**. *Let* (x_k) *be a random sequence generated by Algorithm* 2.2. *Then there is a random variable* x_* *with values in* $\operatorname{argmin} f$ *such that*

$$x_k \overset{a.s.}{\to} x_*. \tag{5.18}$$

Moreover, the sequence (f_k) *is a.s. non-increasing with* $\lim_k f_k \overset{a.s.}{=} f_*$ *and the following convergence rate in expectation holds:*

$$\mathbb{E}\left[f_k - f_*\right] \leq D_0 / \sum_{j=0}^{k} \alpha_j,$$

where the constant $D_0 > 0$ *is provided in the proof. Finally, if* **A.4** *also holds,*

$$f_k - f_* \overset{a.s.}{=} o(k^{-1}). \tag{5.19}$$

Proof. We recall the equality from Lemma 5.11: for every $x_* \in \operatorname{argmin} f$ and every $k \in \mathbb{N}$,

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2 \, \big| \, \mathcal{F}_k\right] - \|x_k - x_*\|^2 \overset{a.s.}{=} \frac{\alpha_k^2 d}{\ell} \|\nabla f_k\|^2 + 2\alpha_k \langle \nabla f_k, x_* - x_k \rangle. \tag{5.20}$$

By Baillon-Haddad Theorem 9.1, ∇f is $1/\lambda$ co-coercive and so

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2 \, \big| \, \mathcal{F}_k\right] - \|x_k - x_*\|^2 \overset{a.s.}{\leq} \frac{\alpha_k^2 d}{\ell} \|\nabla f_k\|^2 - \frac{2\alpha_k}{\lambda} \|\nabla f_k\|^2$$

$$(\textbf{A.1}) \quad \overset{a.s.}{\leq} -\frac{2 - \overline{\alpha}\Lambda}{\lambda} \, \alpha_k \|\nabla f_k\|^2. \tag{5.21}$$

Define $\xi_0 := 2 - \overline{\alpha}\Lambda/\lambda$, a strictly positive quantity. Then by Robbins-Siegmund Lemma 9.11, for every $x_* \in \operatorname{argmin} f$ the random variable $(\|x_k - x_*\|)$ is a.s. convergent and that a.s. $\left(\alpha_k \|\nabla f_k\|^2\right) \in \ell^1$. Beginning again from the equality in Lemma 5.11, we estimate the term $2\alpha_k \langle \nabla f_k, x_* - x_k \rangle$ using the convexity of f and the gradient inequality:

$$f_k + \langle \nabla f_k, x_* - x_k \rangle \leq f_*.$$

Recalling that $\alpha_k \leq \overline{\alpha}$ by Assumption **A.1**, it leads to

$$\mathbb{E}\left[\|x_{k+1} - x_*\|^2 \, \big| \, \mathcal{F}_k\right] - \|x_k - x_*\|^2 + 2\alpha_k \left(f_k - f_*\right) \overset{a.s.}{\leq} \frac{\overline{\alpha} d}{\ell} \alpha_k \|\nabla f_k\|^2. \tag{5.22}$$

Robbins-Siegmund Lemma 9.11 reveals that $(\alpha_k(f_k - f_*)) \in \ell^1$ a.s. Since by assumption (α_k) is not summable, $\liminf_k (f_k - f_*) \overset{\text{a.s.}}{=} 0$. By Remark 4.4, the sequence (f_k) is a.s. non-increasing and bounded below by f_*. In particular, it is a.s. convergent and $f_k \overset{\text{a.s.}}{\to} f_*$. Following the same reasoning as in the proof of Theorem 5.6, there is a random variable x_* with values in $\operatorname{argmin} f$ such that $x_k \overset{\text{a.s.}}{\to} x_*$. To obtain the convergence rate first take the total expectation in inequality (5.21) and sum from $j = 0$ to $j = k$ to get

$$
\begin{aligned}
\sum_{j=0}^{k} \alpha_j \mathbb{E}\|\nabla f_j\|^2 &\leq \frac{1}{\xi_0} \sum_{j=0}^{k} \left(\mathbb{E}\|x_j - x_*\|^2 - \mathbb{E}\|x_{j+1} - x_*\|^2 \right) \\
&= \frac{1}{\xi_0} \left(\mathbb{E}\|x_0 - x_*\|^2 - \mathbb{E}\|x_{k+1} - x_*\|^2 \right) \\
&\leq \frac{1}{\xi_0} \|x_0 - x_*\|^2.
\end{aligned}
\tag{5.23}
$$

Summing (5.22) over k and combining it with (5.23), an expectation yields

$$
\begin{aligned}
\sum_{j=0}^{k} \alpha_j \, \mathbb{E}\left[f_j - f_*\right] &\leq \frac{1}{2} \mathbb{E}\|x_0 - x_*\|^2 - \frac{1}{2} \mathbb{E}\|x_{k+1} - x_*\|^2 + \frac{\overline{\alpha} d}{2\ell} \sum_{j=0}^{k} \alpha_j \mathbb{E}\|\nabla f_j\|^2 \\
(5.23) \quad &\leq \frac{1}{2} \|x_0 - x_*\|^2 + \frac{\overline{\alpha} d}{2\ell \xi_0} \|x_0 - x_*\|^2 \\
&= \frac{\ell \xi_0 + \overline{\alpha} d}{2\ell \xi_0} \|x_0 - x_*\|^2 \\
&= \frac{1}{2 - \Lambda \overline{\alpha}} \|x_0 - x_*\|^2 =: D_0 < +\infty.
\end{aligned}
$$

Since $(\mathbb{E}f_j)$ is non-increasing, $\mathbb{E}[f_k - f_*] \sum_{j=0}^{k} \alpha_j \leq \sum_{j=0}^{k} \alpha_j \, \mathbb{E}[f_j - f_*]$. Dividing by the sum over α_j yields the rate. Finally, assuming also **A.4** and using the fact that $(\alpha_k(f_k - f_*)) \in \ell^1$ a.s., we get $(f_k - f_*) \in \ell^1$ a.s. Since $(f_k - f_*)$ is also a.s. non-increasing, we conclude by Lemma 9.6 that $f_k - f_* \overset{\text{a.s.}}{=} o(k^{-1})$. $\quad\square$

Remark 5.13. Note that the constant D_0 in Theorem 5.12 indeed corresponds to D in Theorem 5.4 when the discretization is set $h_k = 0$ for every $k \in \mathbb{N}$.

Remark 5.14. Under the same conditions of Remark 5.9 but considering the analogue of recursion (2.2), the following result is obtained in [40, Theorem 8]: taking constant $\alpha_k = 1/(4\lambda(d+4))$ and $h_k = h$,

$$
\mathbb{E}\left[f(\bar{x}_k) - f_*\right] \leq 4\lambda(d+4)\|x_0 - x_*\|^2/(2k).
\tag{5.24}
$$

Under the same assumptions, consider the case of the sampling that we proposed (see Assumptions **P.1** and **P.2**). From Theorem 5.12, we get that

$$
\mathbb{E}\left[f(\bar{x}_k) - f_*\right] \leq \frac{\|x_0 - x_*\|^2}{2(1 - \Lambda\alpha)\alpha k}.
$$

For the case $\alpha = \ell/(2\lambda d)$ and $\ell = 1$ we derive

$$
\mathbb{E}\left[f(\bar{x}_k) - f_*\right] \leq 2\lambda d \|x_0 - x_*\|^2/k
\tag{5.25}
$$

and the same observations of Remark 5.9 hold.

6 Polyak-Łojasiewicz case

In contrast with the case of a general convex f considered thus far, this section assumes the PL inequality (**H.3**), but convexity (**H.2**) is not needed. Since strong convexity implies the PL inequality, all of the results in this section hold when the objective function is strongly convex. We use the PL inequality in the a.s. quasi-decreasing estimates of Section 4.3 to get the main estimate of Lemma 6.1. The application of Lemma 6.1 in different settings leads to the convergence rates in expectation for the function values given in Theorems 6.2, 6.4 (based on Lemma 9.5), 6.5 (based on Lemma 9.4) and 6.8. These results are similar but intrinsically different. In Theorem 6.2, we study the case in which both sequences (h_k) and (α_k) are bounded above, but not converging to zero; specifically, the error generated by the discretization does not vanish. In this context, we obtain a *linear* rate in expectation not to the optimal value, but to a sublevel of the objective function depending on $\overline{C}, \underline{\tau}$ and γ, see (6.1). In Theorem 6.4, for the case of α_k constant and vanishing h_k, we get *sublinear* rates in expectation *to the optimum*. In Theorem 6.5 we obtain similar rates assuming that both (α_k) and (h_k) converge to zero polynomially. Note that the algorithm does not converge to the optimal value if h_k does not vanish, even with rapid decay of α_k, a fact that may be surprising to readers more familiar with first-order stochastic approximation algorithms though, as discussed, it is easy to see why. Finally, in Theorem 6.8 we show linear convergence rates to the optimal value for a fast decay of h_k .

6.1 Main estimate

The following basic estimate will be used repeatedly.

Lemma 6.1. *Let* (x_k) *be generated by Algorithm 1. Assume* **H.1**, **H.3**, **P.1**, **P.2** *and* **A.1**. *Then, for every* $k \in \mathbb{N}$,

$$\mathbb{E}\left[f_{k+1} - f_*\right] \leq \left(1 - \frac{w\alpha_k\gamma}{2}\right)\mathbb{E}\left[f_k - f_*\right] + C\alpha_k h_k^2.$$

Assuming also **A.4** *and defining* $\eta := 1 - w\underline{\alpha}\gamma/2$, *we get*

$$\mathbb{E}\left[f_k - f_*\right] \leq \eta^k \left[(f_0 - f_*) + \frac{C}{\eta}\sum_{j=0}^{k-1}\frac{\alpha_j h_j^2}{\eta^j}\right].$$

Proof. Taking the conditional expectation given $\mathcal{F}_k$ in the a.s. inequality of Proposition 4.3, we have

$$\mathbb{E}\left[f_{k+1}\,\middle|\,\mathcal{F}_k\right] - f_k \leq -\frac{w\alpha_k}{2}\mathbb{E}\left[\|P_k^\top\nabla f_k\|^2\,\middle|\,\mathcal{F}_k\right] + C\alpha_k h_k^2$$

$$(\textbf{P.2}) \;\; = -\frac{w\alpha_k}{2}\|\nabla f_k\|^2 + C\alpha_k h_k^2$$

$$(\textbf{H.3}) \;\; \leq -\frac{w\alpha_k}{2}\gamma\left(f_k - f_*\right) + C\alpha_k h_k^2.$$

An expectation yields the first claim. Under Assumption **A.4**, the last inequality yields

$$\mathbb{E}\left[f_{k+1} - f_*\right] \leq \left(1 - \frac{w\underline{\alpha}\gamma}{2}\right)\mathbb{E}\left[f_k - f_*\right] + C\alpha_k h_k^2.$$

24

Iterating leads to the second claim:

$$
\begin{aligned}
\mathbb{E}\left[f_k - f_*\right] &\leq \eta^k \left(f_0 - f_*\right) \\
&\quad + C\left[\alpha_{k-1}h_{k-1}^2 + \eta\alpha_{k-2}h_{k-2}^2 + \eta^2\alpha_{k-3}h_{k-3}^2 + ... + \eta^{k-1}\alpha_0 h_0^2\right] \\
&= \eta^k\left(f_0 - f_*\right) + C\sum_{j=0}^{k-1}\eta^{k-1-j}\alpha_j h_j^2 \\
&= \eta^k\left[\left(f_0 - f_*\right) + \frac{C}{\eta}\sum_{j=0}^{k-1}\frac{\alpha_j h_j^2}{\eta^j}\right].
\end{aligned}
$$

$\square$

6.2   Linear quasi-rate

We first provide an extended version of Theorem 3.7 (i'). In this case both (α_k) and (h_k) are bounded above but not vanishing, leading to the following result as a direct consequence of Lemma 6.1. In particular, the following bound suggests to stop iterating when $(f_k - f_*) \leq C\overline{\alpha}\overline{h}^2/(1-\eta)$.

Theorem 6.2. *Let (x_k) be generated by Algorithm 1. Assume* **H.1**, **H.3**, **P.1**, **P.2**, **A.1** *and* **A.4**. *Then, for every $k \in \mathbb{N}$,*

$$
\mathbb{E}\left[f_k - f_*\right] \leq \eta^k\left(f_0 - f_*\right) + \frac{C\overline{\alpha}\overline{h}^2}{1-\eta}\left[1 - \eta^k\right],
$$

where the constant C is defined in Proposition 4.3 and $\eta = 1 - w\underline{\alpha}\gamma/2$.

Remark 6.3. For recursion 2.2 we recover the linear rate proved in [32]; namely,

$$
\mathbb{E}\left[f_k - f_*\right] \leq \left(1 - \frac{w\underline{\alpha}\gamma}{2}\right)^k\left(f_0 - f_*\right).
$$

On the other hand, for $0 \leq h_k \leq \overline{h}$, the cumulative error term does not vanish:

$$
\lim_k \frac{C\overline{\alpha}\overline{h}^2}{1-\eta}\left[1 - \eta^k\right] = \frac{2C\overline{\alpha}\overline{h}^2}{w\underline{\alpha}\gamma} = \frac{\ell\Lambda^2\overline{\alpha}\overline{h}^2}{4w\underline{\alpha}\gamma\min\left(1, 2 - \Lambda\overline{\alpha} - w\right)}, \tag{6.1}
$$

where we recall that $0 < w < 2 - \Lambda\overline{\alpha}$ and $w \leq 1$. Finally, note that for $\alpha_k = 1/\Lambda$, by Remark 4.6 we have that the decreasing rate is $1 - \gamma/(2\Lambda)$.

6.3   Sublinear rates

We now state the results obtained with bounded step-size and vanishing discretization. Assuming only that (h_k) converges to zero, the objective function values converge in expectation to the optimum; while, for a polynomial decay of h_k, we get sublinear convergence rates. The following is an extended version of Theorem 3.7 (ii').

Theorem 6.4. *Let (x_k) be generated by Algorithm 1. Assume* **H.1**, **H.3**, **P.1**, **P.2**, **A.1** *and* **A.4**. *If (h_k) converges to zero, then*

$$\lim_k \mathbb{E} f_k = f_*.$$

Moreover, if $h_k = h/k^r$ for some $r > 0$ and $h > 0$, then

$$\limsup_k \ k^{2r} \mathbb{E}\left[f_k - f_*\right] \leq \frac{2C\overline{\alpha}h^2}{w\underline{\alpha}\gamma},$$

In particular, there is a constant $\tilde{C} > 0$ such that

$$\mathbb{E}\left[f_k - f_*\right] \leq \tilde{C}/k^{2r}$$

and, for every $t \in (0, r)$,

$$\mathbb{E}\left[f_k - f_*\right] = o\left(\frac{1}{k^{2t}}\right).$$

Proof. From Lemma 6.1, we get

$$\mathbb{E}\left[f_{k+1} - f_*\right] \leq \eta \mathbb{E}\left[f_k - f_*\right] + C\overline{\alpha}\, h_k^2.$$

The remainder follows from Lemma 9.5 with $c_k = C\overline{\alpha}h_k^2$, $c = C\overline{\alpha}h^2$ and $t = 2r$. $\qquad\square$

In the next result we allow both the step-size and the discretization to converge to zero polynomially. In this case, we get again sublinear rates in expectation similar to the ones obtained in Theorem 6.4.

Theorem 6.5. *Let (x_k) be generated by Algorithm 1. Assume* **H.1**, **H.3**, **P.1** *and* **P.2**. *For $0 < \alpha < 2/\Lambda$ and $h > 0$, set $\alpha_k = \alpha/k^s$, $h_k = h/k^r$ and define*

$$c := \ w\alpha\gamma/2 \quad and \quad d := \ C\alpha h^2. \tag{6.2}$$

Then, for $s = 1$ and $r > 0$, we get

$$\mathbb{E}\left[f_k - f_*\right] \leq \begin{cases} \frac{d}{(c-2r)k^{2r}} + o\left(\frac{1}{k^{2r}}\right) & \text{if } 2r < c; \\ O\left(\frac{\log k}{k^c}\right) & \text{if } 2r = c; \\ O\left(\frac{1}{k^c}\right) & \text{if } 2r > c. \end{cases} \tag{6.3}$$

If $0 < s < 1$, for every $r > 0$, we have

$$\mathbb{E}\left[f_k - f_*\right] \leq \frac{d}{c}\frac{1}{k^{2r}} + o\left(\frac{1}{k^{2r}}\right).$$

Remark 6.6. In order for the previous results to hold, both (α_k) and (h_k) must converge to zero. In the case of $s = 1$ (and so α_k proportional to $1/k$), as the intuition suggests, the rate improves for larger r (and thus for h_k vanishing faster) up to the value $c/2$. But eventually a saturation effect occurs: increasing r beyond $c/2$ does not improve the bound. On the other hand, for $\alpha_k = \alpha/k^s$ with $0 < s < 1$, the convergence rates improve with larger values of r similar to the ones in Theorem 6.4.

Remark 6.7. Note the difference between the results of this section. In Theorem 6.2, for non-vanishing h_k, we obtain a linear rate *with an error*; while in Theorems 6.4 and 6.5, with h_k going to zero, the convergence rates are sublinear but *to the optimum*.

Proof. From Lemma 6.1,

$$\mathbb{E}\left[f_{k+1} - f_*\right] \leq \left(1 - \frac{w\alpha_k\gamma}{2}\right)\mathbb{E}\left[f_k - f_*\right] + C\alpha_k h_k^2.$$

To conclude the first result, apply the first part of Lemma 9.4 with $p = 2r$ and c, d as in (6.2). For the second result, apply the second part of Lemma 9.4 with the same c, d as before and $s = s$ and $t = s + 2r$. $\quad\square$

6.4 Linear rate

Finally, assuming a fast decay of h_k, we derive linear convergence rates in expectation. The following result is an extended version of Theorem 3.7(iii'). The proof is a simple consequence of Lemma 6.1.

Theorem 6.8. *Let* (x_k) *be generated by Algorithm 1. Assume* **H.1**, **H.3**, **P.1**, **P.2**, **A.1** *and* **A.4**. *Assume that* $\left(h_k^2/\eta_k\right) \in \ell^1$. *Then, for every* $k \in \mathbb{N}$,

$$\mathbb{E}\left[f_k - f_*\right] \leq \eta^k\left[(f_0 - f_*) + \frac{C\overline{\alpha}}{\eta}\sum_{j=0}^{+\infty}\frac{h_j^2}{\eta^j}\right],$$

where the constant C *is defined in Proposition 4.3 and* $\eta = 1 - w\underline{\alpha}\gamma/2$.

7 Numerical results

In this section we present synthetic examples illustrating the different results we derived and discussed. Our analysis unifies many algorithms that have been thoroughly empirically examined, e.g., [15, 7, 33]. Hence, we present only toy problems to illustrate our theoretical results. We omit the convex, non-PL case as we were unable to identify a function that resulted in a qualitative difference in performance of the algorithm for any $\ell \in \{1, \ldots, d\}$.

Convex function satisfying PL inequality. Many convex functions satisfy the PL inequality leading to an improved rate of convergence. An example of this case is $f(x) = \|Ax\|^2$ where $A \in \mathbb{R}^{n \times d}$ is fixed but not necessarily full column rank and $\lambda = 100$. Specifically, if any eigenvalue of A is 0 then f is not strongly convex, however because it is PL we are still able to apply Theorem 3.7. We choose $n = d = 100$, and we force at least one eigenvalue of A to be 0. For the cases $\ell < d$ we take the average of 10 runs. While in the long run the discrete gradient method catches up to the subspace approaches, it is important to recognize that for many practical problems of interest the dimension of the objective function may be very high relative to the budget for function evaluations (this budget could be due to time, money, computational power, etc). In very high-dimensional cases, it may not even be possible to perform a single iteration of gradient descent, due to the $d+1$ function evaluations required at each iteration; requiring only ℓ function evaluations per iteration, may allow for substantial progress with identical budget constraints. This effect is even more apparent with the rapid initial progress made with $\ell < d$ in the left panel of Figure 1.

Non-convex function satisfying PL inequality. $f(x) = \|Ax\|^2 + 3\sin^2(c^\top x)$, with A fixed but not necessarily full rank and $Ac = c$. Again, let $n = d = 100$ and $\lambda = 100$. For the cases $\ell < d$ we take the average of 10 runs. The most notable feature in Figures 1 and 2 is that when the step-size is not fixed, choosing
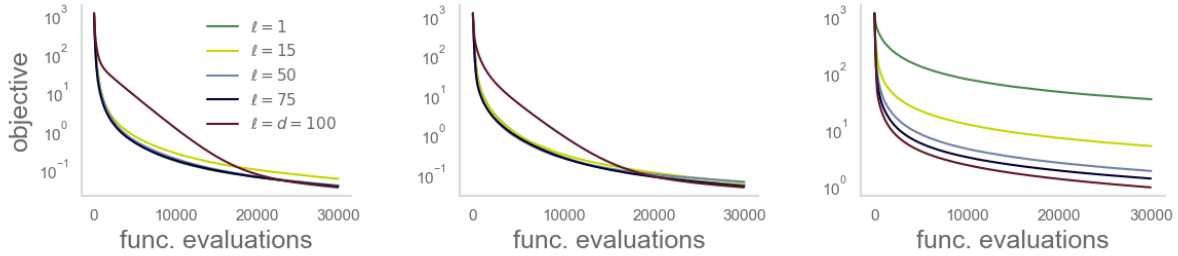
Figure 1: Optimization of a convex objective function satisfying the PL inequality. **left:** $\alpha = \ell/(d\lambda)$, $h = 10^{-7}$. **center:** $\alpha = \ell/(d\lambda)$, $h = 10^{-7}/k^{0.0001}$. **right:** $\alpha = \ell/(d\lambda\sqrt{k})$, $h = 10^{-7}/k^{0.0001}$. Note the different axes between the figures.
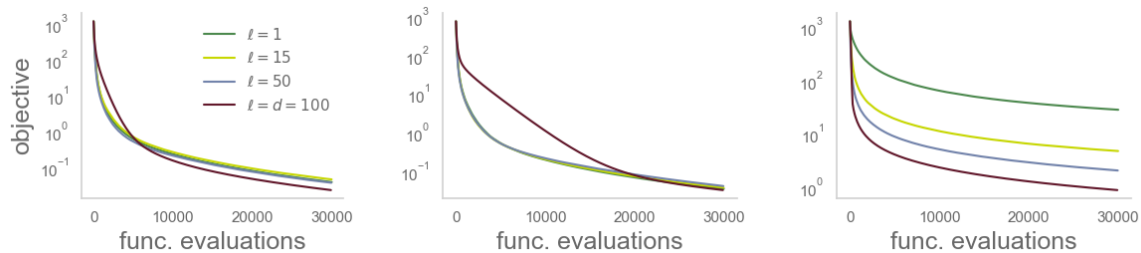

Figure 2: Optimization of a non-convex objective function satisfying the PL inequality. **left:** $\alpha = \ell/(d\lambda)$, $h = 10^{-7}$. **center:** $\alpha = \ell/(d\lambda)$, $h = 10^{-7}/k^{0.0001}$. **right:** $\alpha = \ell/(d\lambda\sqrt{k})$, $h = 10^{-7}/k^{0.0001}$. Note the different axes between the figures.

$\ell < d$ severely under performs the discrete gradient method. When $\ell < d$ only a subset of the available information is being used at each iteration, and with the step-size diminishing every successive iteration has less impact than those that precede it. Thus, the trade-off between cost-per-iteration and progress-per-iteration favors a higher per-iteration cost in return for more progress, particularly in the early iterations. This trade-off flips when the step-size is fixed: much faster progress is made early on when $\ell < d$ and many directions provide improvement of the objective, but of course the discrete gradient method ultimately catches up.

Variability due to stochasticity. One potential benefit of letting $\ell = d$ and performing the full discrete gradient method is that there is no randomness involved so the results are deterministic. The theorems provide guarantees for $\mathbb{E}f(x_k) - f_*$, but here we investigate how much variability can be expected between runs with identical initializations when $\ell < d$. We use the same non-convex function as previously, and perform 100 runs using the same initialization in each case, considering 15000 function evaluations. The substantial overlap between the $\ell < d$ cases makes it difficult to discern between various values of ℓ, thus in this figure we provide only the extremes, $\ell = 1$ and $\ell = d$. The bold line represents the mean of the $\ell = 1$ case, and since $\ell = d$ is deterministic it is run only once. As expected, there is variability between runs when $\ell < d$, however in the early iterations even the worst case performs better than discretized gradient descent. We have only a heuristic explanation for this phenomenon: in the early iterations there are many directions that may lead to improvement so the inexpensive $\ell = 1$ directions are more efficient than a full gradient estimate, but as we approach the optimum a judicious choice of direction is rewarded. The fact that the discrete gradient method
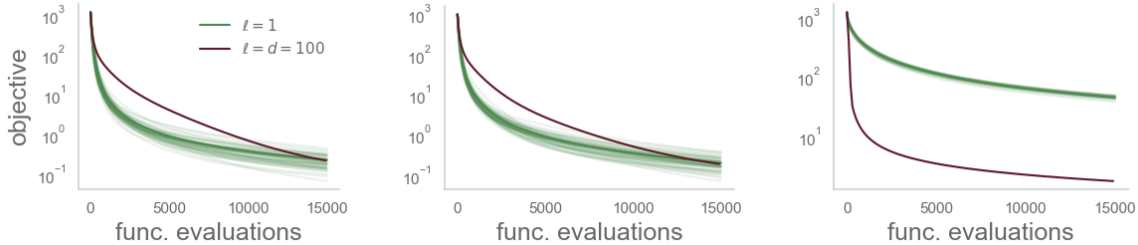
28

Figure 3: Optimization of a convex objective function with 100 restarts at the same initialization. **left:** $\alpha = \ell/(d\lambda)$, $h = 10^{-7}$. **center:** $\alpha = \ell/(d\lambda)$, $h = 10^{-7}/k^{.0001}$. **right:** $\alpha = \ell/(d\lambda\sqrt{k})$, $h = 10^{-7}/k^{1.0001}$. Note the different axes between the figures.

catches up and eventually outperforms the others is consistent with the theory since all else equal, larger ℓ implies better progress per iteration. Of greater practical interest is the out-performance in the scenario $\ell \ll d$ in the early iterations. This suggests that for low-precision optimization, or when relatively few iterations are possible due to time or money constraints, it may be beneficial to choose $\ell < d$. Indeed, this is precisely the scenario where subspace descent methods are used, cf. [33, 15, 35]. Our theory does not cover the use of a backtracking line search such as the one presented in [8], so we do not provide figures detailing its performance; however, in practice a line search is a necessary component to achieve outperformance compared to the gradient method as discussed at length in [33].

Effect of h. Recall that there is an additive error term at each iteration due to the use of finite differences to approximate the gradient. In light of this fact, it may be surprising that the preceding figures appear to have objective function values that decrease monotonically with the increase in function evaluations even when h is fixed. Note, however, that for a fixed h as in Theorem 6.2 the error is asymptotically $\mathcal{O}(h^2)$, whereas in Theorem 6.4 a decaying $h_k = h/k^r$ is used with $r, h > 0$, resulting in an objective that decays to zero. Figure 4 demonstrates that the limiting error, or lack thereof, guaranteed by the theorems is observed in practice. The first figure is a convex function satisfying the PL inequality with various fixed values for h. The second is a non-convex function satisfying the PL inequality with the same values of h, and in the third figure we set $h_k = h/k^r$ with $h = 10^{-5}$ and $r = 1$ and run the algorithm 100 times on the non-convex function. In all cases, $d = 5$, $\ell = 1$ and $\lambda = 4$, resulting in fast convergence.

Several conclusions can be drawn from Figure 4. First, as expected by the theorems, h does not appear to play a role in the rate of convergence, only in the magnitude of the asymptotic error. Further, for all of the algorithms that fit our theory the finite difference error can essentially be ignored, provided the desired accuracy is less than $\mathcal{O}(h^2)$. So, with h the square root of machine precision, our theorems seem to indicate (and our figures support) that the error due to finite differences can be safely ignored. The figure on the right shows that the variance due to different stochastic realizations of the algorithm is small, suggesting that the theoretical analysis done in [33] and [32] may be extended to the discrete setting.

8   Conclusions

We presented a method that generalizes several well-known derivative-free optimization algorithms including, for example, spherical smoothing and discretized versions of coordinate and gradient descent. We provide convergence analysis of this generic method considering objective functions that are either convex or satisfy a Polyak-Łojasiewicz (PL) condition. Multiple possible choices for the stepsizes and the finite dif-
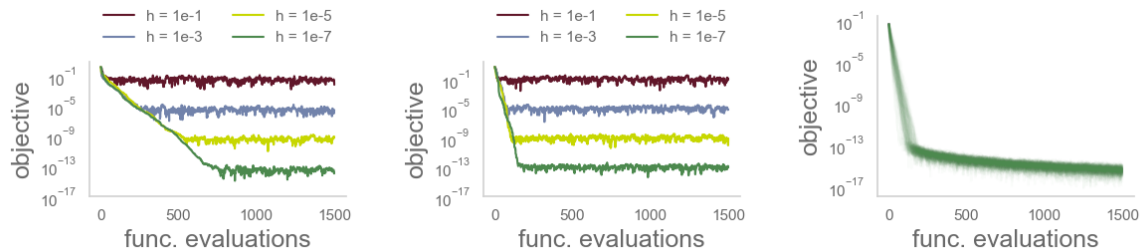
Figure 4: **left:** Optimization of a convex PL function using various values of h. **center:** Optimization of a non-convex PL function using various values of h. **right:** Optimization of a non-convex PL function with $h_k = 10^{-5}/k$

ference parameter are studied. The best choice depends on the error tolerance of the user. By allowing the stepsize and discretization to decay the algorithm achieves a slower rate of convergence but is able to converge to an optimum, but by fixing these values a faster convergence rate is obtained at the expense of converging only to within a region of an optimum. To the best of our knowledge, this work provides the first convergence guarantees for the iterates of spherical smoothing and discretized coordinate descent to a minimizer when the objective function is convex.

There are several possible extensions of this work. Our analysis does not cover the use of an adaptive stepsize that has been shown empirically to be highly advantageous [35, 33]. Theoretical analysis using a stochastic linesearch may be possible, several recent papers offer promising results that may extend to our case [13, 8, 42]. It would be interesting to analyze the algorithm in the presence of noisy function evaluations as in [30, 34]. Such an extension would be of great practical consequence as many applications of interest have noisy objective functions. Finally, faster convergence may be possible using derivative-free quasi-Newton methods or any methods that exploit the curvature of the objective as in [6, 27, 28, 10].

References

[1] A. AGARWAL, O. DEKEL, AND L. XIAO, *Optimal algorithms for online convex optimization with multi-point bandit feedback*, in Proceedings of the Twenty Third Annual Conference on Computational Learning Theory, Citeseer, 2010, pp. 28–40.

[2] H. ATTOUCH, J. BOLTE, AND B. F. SVAITER, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods*, Mathematical Programming, 137 (2013), pp. 91–129.

[3] H. AVRON, P. MAYMOUNKOV, AND S. TOLEDO, *Blendenpik: supercharging Lapack's least-squares solver*, SIAM J. Sci. Comput., 32 (2010), pp. 1217–1236.

[4] J.-B. BAILLON AND G. HADDAD, *Quelques propriétés des opérateurs angle-bornés etn-cycliquement monotones*, Israel Journal of Mathematics, 26 (1977), pp. 137–150.

[5] A. BECK AND L. TETRUASHVILI, *On the convergence of block coordinate descent type methods*, SIAM J. Optim., 23 (2013), pp. 2037–2060.

[6] A. S. Berahas, R. H. Byrd, and J. Nocedal, *Derivative-free optimization of noisy functions via quasi-Newton methods*, SIAM J. Optim., 29 (2019), pp. 965–993.

[7] A. S. Berahas, L. Cao, K. Choromanski, and K. Scheinberg, *A theoretical and empirical comparison of gradient approximations in derivative-free optimization*, arXiv preprint arXiv:1905.01332, (2019).

[8] A. S. Berahas, L. Cao, and K. Scheinberg, *Global convergence rate analysis of a generic line search algorithm with noise*, arXiv preprint arXiv:1910.04055, (2019).

[9] D. P. Bertsekas and J. N. Tsitsiklis, *Gradient convergence in gradient methods with errors*, SIAM Journal on Optimization, 10 (2000), pp. 627–642.

[10] R. Bollapragada and S. M. Wild, *Adaptive sampling quasi-newton methods for derivative-free stochastic optimization*, arXiv preprint arXiv:1910.13516, (2019).

[11] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, Mathematical Programming, 165 (2017), pp. 471–507.

[12] C. Cartis and L. Roberts, *Scalable subspace methods for derivative-free nonlinear least-squares optimization*, arXiv preprint arXiv:2102.12016, (2021).

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

[14] A. Cauchy et al., *Méthode générale pour la résolution des systemes d'équations simultanées*, Comp. Rend. Sci. Paris, 25 (1847), pp. 536–538.

[15] K. Choromanski, M. Rowland, V. Sindhwani, R. Turner, and A. Weller, *Structured evolution with compact architectures for scalable policy optimization*, in Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause, eds., vol. 80 of Proceedings of Machine Learning Research, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018, PMLR, pp. 970–978.

[16] K. L. Chung, *On a stochastic approximation method*, Ann. Math. Statistics, 25 (1954), pp. 463–483.

[17] P. L. Combettes and J.-C. Pesquet, *Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping*, SIAM Journal on Optimization, 25 (2015), pp. 1221–1248.

[18] A. R. Conn, K. Scheinberg, and L. N. Vicente, *Introduction to derivative-free optimization*, vol. 8 of MPS/SIAM Series on Optimization, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, 2009.

[19] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, *Faster least squares approximation*, Numer. Math., 117 (2011), pp. 219–249.

[20] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono, *Optimal rates for zero-order convex optimization: The power of two function evaluations*, IEEE Transactions on Information Theory, 61 (2015), pp. 2788–2806.

[21] M. J. Ehrhardt, E. S. Riis, T. Ringholm, and C.-B. Schönlieb, *A geometric integration approach to smooth optimisation: Foundations of the discrete gradient method*, arXiv preprint arXiv:1805.06444, (2018).

[22] A. D. Flaxman, A. T. Kalai, and H. B. McMahan, *Online convex optimization in the bandit setting: gradient descent without a gradient*, in Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, ACM, New York, 2005, pp. 385–394.

[23] S. Ghadimi and G. Lan, *Stochastic first-and zeroth-order methods for nonconvex stochastic programming*, SIAM Journal on Optimization, 23 (2013), pp. 2341–2368.

[24] V. Grimm, R. I. McLachlan, D. I. McLaren, G. Quispel, and C. Schönlieb, *Discrete gradient methods for solving variational image regularisation models*, Journal of Physics A: Mathematical and Theoretical, 50 (2017), p. 295201.

[25] A. Gupal, *A method for the minimization of almost-differentiable functions*, Cybernetics, 13 (1977), pp. 115–117.

[26] A. Gupal and V. Norkin, *Algorithm for the minimization of discontinuous functions*, Cybernetics, 13 (1977), pp. 220–223.

[27] F. Hanzely, N. Doikov, Y. Nesterov, and P. Richtarik, *Stochastic subspace cubic Newton method*, in Proceedings of the 37th International Conference on Machine Learning, H. D. III and A. Singh, eds., vol. 119 of Proceedings of Machine Learning Research, PMLR, 13–18 Jul 2020, pp. 4027–4038.

[28] F. Hanzely, D. Kovalev, and P. Richtarik, *Variance reduced coordinate descent with acceleration: New method with a surprising application to finite-sum problems*, in Proceedings of the 37th International Conference on Machine Learning, H. D. III and A. Singh, eds., vol. 119 of Proceedings of Machine Learning Research, PMLR, 13–18 Jul 2020, pp. 4039–4048.

[29] H. Karimi, J. Nutini, and M. 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, Springer, 2016, pp. 795–811.

[30] J. Kiefer and J. Wolfowitz, *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statistics, 23 (1952), pp. 462–466.

[31] K. Knopp, *Theory and application of infinite series*, Courier Corporation, 1990.

[32] D. Kozak, S. Becker, A. Doostan, and L. Tenorio, *Stochastic subspace descent*, arXiv preprint arXiv:1904.01145, (2019).

[33] D. Kozak, S. Becker, A. Doostan, and L. Tenorio, *A stochastic subspace approach to gradient-free optimization in high dimensions*, Comput. Optim. Appl., 79 (2021), pp. 339–368.

[34] H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems*, vol. 26 of Applied Mathematical Sciences, Springer-Verlag, New York-Berlin, 1978.

[35] H. Mania, A. Guy, and B. Recht, *Simple random search of static linear policies is competitive for reinforcement learning*, in Advances in Neural Information Processing Systems 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., Curran Associates, Inc., 2018, pp. 1800–1809.

[36] P.-G. Martinsson and J. A. Tropp, *Randomized numerical linear algebra: foundations and algorithms*, Acta Numer., 29 (2020), pp. 403–572.

[37] J. Matyas, *Random optimization*, Automation and Remote control, 26 (1965), pp. 246–253.

[38] F. Mezzadri, *How to generate random matrices from the classical compact groups*, Notices Amer. Math. Soc., 54 (2007), pp. 592–604.

[39] Y. Nesterov, *Random gradient-free minimization of convex functions*, tech. rep., Université catholique de Louvain, Center for Operations Research and . . . , 2011.

[40] Y. Nesterov and V. Spokoiny, *Random gradient-free minimization of convex functions*, Foundations of Computational Mathematics, 17 (2017), pp. 527–566. First appeard as CORE discussion paper 2011.

[41] Z. a. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591–597.

[42] C. Paquette and K. Scheinberg, *A stochastic line search method with expected complexity analysis*, SIAM J. Optim., 30 (2020), pp. 349–376.

[43] B. T. Polyak, *Introduction to optimization.*, vol. 1, Optimization Software Inc., New York, 1987.

[44] L. A. Rastrigin, *About convergence of random search method in extremal control of multi-parameter systems*, Avtomat. i Telemekh, 24 (1963), pp. 1467–1473.

[45] E. S. Riis, M. J. Ehrhardt, G. Quispel, and C.-B. Schönlieb, *A geometric integration approach to nonsmooth, nonconvex optimisation*, arXiv preprint arXiv:1807.07554, (2018).

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

[47] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Optimizing methods in statistics (Proc. Sympos., Ohio State Univ., Columbus, Ohio, 1971), 1971, pp. 233–257.

[48] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, *Evolution strategies as a scalable alternative to reinforcement learning*, arXiv preprint arXiv:1703.03864, (2017).

[49] S. Salzo and S. Villa, *Parallel random block-coordinate forward-backward algorithm: A unified convergence analysis*, arXiv preprint arXiv:1906.07392, (2019).

[50] M. Schmidt, N. L. Roux, and F. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, arXiv preprint arXiv:1109.2415, (2011).

[51] J. C. Spall, *Multivariate stochastic approximation using a simultaneous perturbation gradient approximation*, IEEE Trans. Automat. Control, 37 (1992), pp. 332–341.

[52] R. Tappenden, M. Takáč, and P. Richtárik, *On the complexity of parallel coordinate descent*, Optimization Methods and Software, 33 (2018), pp. 372–395.

[53] R. J. Williams, *Simple statistical gradient-following algorithms for connectionist reinforcement learning*, Machine learning, 8 (1992), pp. 229–256.

[54] D. P. Woodruff, *Sketching as a tool for numerical linear algebra*, Found. Trends Theor. Comput. Sci., 10 (2014), pp. iv+157.

[55] S. J. Wright, *Coordinate descent algorithms*, Math. Program., 151 (2015), pp. 3–34.

[56] S. K. Zavriev, *On the global optimization properties of finite-difference local descent algorithms*, J. Global Optim., 3 (1993), pp. 67–78.

9 Supplementary material and auxiliary lemmas

Here we collect the main auxiliary results used in the convergence analysis of Algorithm 1.

Proof of Lemma 4.1:   Let $k \in \mathbb{N}$, denoting by $p^{(j)}$ the j-th column of P_k, we want to estimate

$$\left\| \nabla_{(P_k, h_k)} f(x) - P_k^\top \nabla f(x) \right\| = \sqrt{\sum_{j=1}^{\ell} \left(\left[\nabla_{(P_k, h_k)} f(x) \right]_j - \langle \nabla f(x), p^{(j)} \rangle \right)^2}. \tag{9.1}$$

To get an upper-bound for the term in parenthesis, we use **H.1**. As ∇f is λ-Lipschitz, the Descent Lemma 9.2 holds: for every $x \in \mathbb{R}^d$,

$$\left| f(x + h_k p_j) - f(x) - h_k \langle \nabla f(x), p^{(j)} \rangle \right| \leq \frac{\lambda h_k^2}{2} \left\| p^{(j)} \right\|^2. \tag{9.2}$$

Then, rearranging,

$$\left| \left[\nabla_{(P_k, h_k)} f(x) \right]_j - \langle \nabla f(x), p^{(j)} \rangle \right| \leq \frac{\lambda h_k}{2} \left\| p^{(j)} \right\|^2 \quad \text{a.s.} \tag{9.3}$$

Note that **P.1** implies $\|p^{(j)}\|^2 \stackrel{a.s.}{=} d/\ell$ and so (9.1) and (9.3) yield

$$\left\| \nabla_{(P_k, h_k)} f(x) - P_k^\top \nabla f(x) \right\| \leq \sqrt{\sum_{j=1}^{\ell} \left(\frac{\lambda h_k d}{2\ell} \right)^2} = \frac{\lambda h_k d}{2\sqrt{\ell}} \quad \text{a.s..}$$

$\square$

Lipschitz smooth functions   We start with two well-known lemmas on Lipschitz smooth functions, namely differentiable functions with Lipschitz continuous gradient.

Lemma 9.1 (Baillon-Haddad Theorem [4]). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and Fréchet differentiable function with λ-Lipschitz continuous gradient for some $\lambda > 0$. Then ∇f is $1/\lambda$ co-coercive; namely, for every $x, y \in \mathbb{R}^d$,*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \|\nabla f(y) - \nabla f(x)\|^2 / \lambda.$$

Lemma 9.2 (Descent Lemma [43, Sec 1.1.2]). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a Fréchet differentiable function with λ-Lipschitz continuous gradient. Then, for every $x, y \in \mathbb{R}^d$,*

$$|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| \leq \lambda \|y - x\|^2 / 2.$$

Real sequences   In this section, we first recall Opial's Lemma in the deterministic setting. In the next, we collect some results regarding convergence and convergence rates for real sequences.

Lemma 9.3 (Opial, deterministic version [41]). *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ be a non-empty subset and $(x_k) \subseteq \mathbb{R}^d$ a sequence. Assume that*
- *for every $z \in \mathcal{Z}$,*

$$\exists \lim_k \|x_k - z\|;$$

- *every cluster point of x_k belongs to $\mathcal{Z}$; namely,*

$$x_{k_j} \to x_\infty \quad \implies \quad x_\infty \in \mathcal{Z}.$$

Then there exists $\bar{x} \in \mathcal{Z}$ such that $x_k \to \bar{x}$.

Lemma 9.4 (Chung's Lemma [16]). *Let (a_k) be a non-negative sequence and c, d and p strictly-positive constants.*
First suppose that, for every $k \in \mathbb{N}$,

$$a_{k+1} \leq \left(1 - \frac{c}{k}\right) a_k + \frac{d}{k^{p+1}}.$$

Then

$$a_k \leq \begin{cases} \frac{d}{(c-p)k^p} + o\left(\frac{1}{k^p}\right) & \text{if } p < c; \\ O\left(\frac{\log k}{k^c}\right) & \text{if } p = c; \\ O\left(\frac{1}{k^c}\right) & \text{if } p > c. \end{cases} \tag{9.4}$$

Now suppose that, for some $0 < s < 1$ and $s < t$ and every $k \in \mathbb{N}$,

$$a_{k+1} \leq \left(1 - \frac{c}{k^s}\right) a_k + \frac{d}{k^t}.$$

Then

$$a_k \leq \frac{d}{c} \frac{1}{k^{t-s}} + o\left(\frac{1}{k^{t-s}}\right).$$

Lemma 9.5. *Let (a_k) and (c_k) be non-negative sequences with $\lim_k c_k = 0$ and let $\eta \in (0,1)$. If for every $k \in \mathbb{N}$*

$$a_{k+1} \leq (1 - \eta) a_k + c_k, \tag{9.5}$$

then $\lim_k a_k = 0$.
Moreover, if $c_k = c/k^t$ for some $c \geq 0$ and $t > 0$, then

$$\limsup_k k^t a_k \leq \frac{c}{\eta}.$$

Proof. Iterating the inequality in (9.5), we get

$$a_k \leq (1 - \eta)^k a_0$$
$$+ \left(c_{k-1} + (1 - \eta) c_{k-2} + (1 - \eta)^2 c_{k-3} + (1 - \eta)^3 c_{k-4} + \ldots + (1 - \eta)^{k-1} c_0 \right)$$
$$\leq a_0 + \left(\sup_{j \geq 0} c_j \right) \sum_{j \geq 0} (1 - \eta)^j =: \tilde{C} < +\infty,$$

where we used the fact that (c_k) is convergent (and thus bounded) and that $\eta \in (0,1)$. In particular, the sequence (a_k) is bounded and so its $\limsup$ is a real number. Then, again from the hypothesis that $a_{k+1} \leq (1 - \eta) a_k + c_k$ and the existence of $\lim_k c_k$, we have

$$\limsup_k a_k \leq \limsup_k \left[(1 - \eta) a_k + c_k \right] = (1 - \eta) \limsup_k a_k + \lim_k c_k.$$

So, as $\lim_k c_k = 0$ and $\limsup_k a_k \in [0, +\infty)$, $\eta \limsup_k a_k \le 0$. Finally, as $\eta \in (0, 1)$,

$$\limsup_k a_k \le 0$$

and so, as $a_k \ge 0$, $\lim_k a_k = 0$. Now assume that $c_k = c/k^t$ for some $c \ge 0$ and $t > 0$. By $k^t \le (k+1)^t$, we have

$$\frac{c}{k^t} \le \frac{c}{\eta}\left[\frac{1}{(k+1)^t} - (1-\eta)\frac{1}{k^t}\right].$$

Using the latter inequality (9.5), we get

$$a_{k+1} - \frac{c}{\eta(k+1)^t} \le (1-\eta)\left[a_k - \frac{c}{\eta k^t}\right]. \tag{9.6}$$

First suppose that there exists $\bar{k} \in \mathbb{N}$ such that $a_{\bar{k}} \le \frac{c}{\eta \bar{k}^t}$. Using (9.6), it is easy to see by recursion that, for every $k \ge \bar{k}$,

$$a_k \le \frac{c}{\eta k^t}$$

and so that the claim holds. Now suppose the opposite; namely, that for every $k \in \mathbb{N}$

$$a_k - \frac{c}{\eta k^t} > 0.$$

Then, iterating (9.6), we get

$$a_k - \frac{c}{\eta k^t} \le (1-\eta)^k\left[a_1 - \frac{c}{\eta}\right].$$

Finally,

$$\limsup_k\, k^t a_k \le \limsup_k\left\{\frac{c}{\eta} + k^t(1-\eta)^k\left[a_1 - \frac{c}{\eta}\right]\right\} = \frac{c}{\eta}.$$

$\square$

We conclude this part with the following three well-known results. The proof of the first can be found in [31, Theorem 3.3.1], while the second is just the deterministic version of Lemma 9.11. For the third, related to estimates with errors, see [50].

Lemma 9.6. *Let (a_k) be a non-negative, non-increasing and summable sequence. Then $a_k = o(k^{-1})$.*

Lemma 9.7. *Let (r_k), (β_k), (y_k) and (w_k) be non-negative real sequences with (β_k) and (w_k) in ℓ^1. Suppose that, for every $k \in \mathbb{N}$,*

$$r_{k+1} - (1 + \beta_k)r_k + y_k \le w_k.$$

Then (r_k) is convergent and (y_k) belongs to ℓ^1.

Lemma 9.8. *[Discrete Bihari's Lemma] Assume that (u_k) is a non-negative real sequence, that (S_k) is a non-decreasing sequence such that $S_0 \ge u_0^2$ and that (ρ_j) is a non-negative sequence. If, for every $k \in \mathbb{N}$,*

$$u_k^2 \le S_k + \sum_{j=0}^{k}\rho_j u_j,$$

then, for every $k \in \mathbb{N}$,

$$u_k \le \frac{1}{2}\sum_{j=0}^{k}\rho_j + \left[S_k + \left(\frac{1}{2}\sum_{j=0}^{k}\rho_j\right)^2\right]^{1/2}.$$

Random sequences   In this section, we recall the extension of Opial's Lemma 9.3 and Lemma 9.7 to the stochastic setting (see Lemma 9.10 and 9.11 - respectively). For completeness, we show the proof of Lemma 9.10, starting with the auxiliary Lemma 9.9. In the next, $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space and we say that $\tilde{\Omega}$ is full-measure (f.m.) if $\tilde{\Omega} \subseteq \Omega$ and $\mathbb{P}(\tilde{\Omega}) = 1$. The proof of the Lemma 9.9 can be found in [17, Proposition 2.3] as part of a result about Fejér monotonicity. We repeat the reasoning for clarity.

Lemma 9.9. *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ be a non-empty subset and (x_k) a random sequence on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $\mathbb{R}^d$. Assume that, for every $z \in \mathcal{Z}$, there exists Ω_z f.m. such that, for every $\omega \in \Omega_z$, the sequence $(\|x_k(\omega) - z\|)$ converges. Then there exists $\tilde{\Omega}$ f.m. such that, for every $\omega \in \tilde{\Omega}$ and every $z \in \mathcal{Z}$, $\exists \lim_k \|x_k(\omega) - z\|$.*

Proof. By separability of $\mathcal{Z} \subseteq \mathbb{R}^d$, let $W \subseteq \mathcal{Z}$ be a countable subset such that $\overline{W} = \mathcal{Z}$ and define $\tilde{\Omega} := \bigcap_{w \in W} \Omega_w$. As W is countable and $\mathbb{P}(\Omega_w) = 1$ for every $w \in W$, $\tilde{\Omega}$ is f.m. Moreover, for every $\omega \in \tilde{\Omega}$ and every $w \in W$, there exists

$$\lim_k \|x_k(\omega) - w\|.$$

We want to show that, for every $\omega \in \tilde{\Omega}$ and every $z \in \mathcal{Z}$, there exists

$$\lim_k \|x_k(\omega) - z\|.$$

Fix $\omega \in \tilde{\Omega}$ and $z \in \mathcal{Z}$. As W is dense in $\mathcal{Z}$, there exists a sequence $w_j \subseteq W$ such that $w_j \to z$. As $w_j \in W$ for each $j \in \mathbb{N}$, we know that there exists

$$\lim_k \|x_k(\omega) - w_j\| =: \tau_j(\omega). \tag{9.7}$$

Notice that

$$-\|w_j - z\| \le \|x_k(\omega) - z\| - \|x_k(\omega) - w_j\| \le \|w_j - z\|. \tag{9.8}$$

Then,

$$-\|w_j - z\| \le$$
$$(9.8) \quad \le \liminf_k \left[\|x_k(\omega) - z\| - \|x_{(}\omega) - w_j\| \right]$$
$$(9.7) \quad = \liminf_k \|x_k(\omega) - z\| - \tau_j(\omega)$$
$$\le \limsup_k \|x_k(\omega) - z\| - \tau_j(\omega)$$
$$(9.7) \quad = \limsup_k \left[\|x_k(\omega) - z\| - \|x_k(\omega) - w_j\| \right]$$
$$(9.8) \quad \le \|w_j - z\|.$$

Taking the limit for $j \to +\infty$ and recalling that $w_j \to z$,

$$\liminf_k \|x_k(\omega) - z\| = \limsup_k \|x_k(\omega) - z\|$$

and so that there exists $\lim_k \|x_k(\omega) - z\|$. □

Lemma 9.10 (Opial, stochastic version). *Let $\mathcal{Z} \subseteq \mathbb{R}^d$ a non-empty subset and (x_k) a random sequence on $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $\mathbb{R}^d$. Assume that*

- *for every $z \in \mathcal{Z}$, there exists Ω_z f.m. such that, for every $\omega \in \Omega_z$,*

$$\exists \lim_k \|x_k(\omega) - z\|;$$

  (i.e., for every $z \in \mathcal{Z}$, the random variable $\|x_k - z\|$ converges a.s.)
- *there exists $\hat{\Omega}$ f.m. such that, for every $\omega \in \hat{\Omega}$, every cluster point of $x_k(\omega)$ belongs to $\mathcal{Z}$; namely,*

$$x_{k_j}(\omega) \to x_\infty \implies x_\infty \in \mathcal{Z}.$$

Then there exists a $\mathcal{Z}$-valued random variable $\bar{x}$ such that $x_k \to \bar{x}$ a.s.; namely, there exists $\bar{\Omega}$ f.m. such that, for every $\omega \in \bar{\Omega}$, $x_k(\omega) \to \bar{x}(\omega)$ with $\bar{x}(\omega) \in \mathcal{Z}$.

Proof. From the assumptions and Lemma 9.9, there exists $\tilde{\Omega}$ f.m. such that, for every $\omega \in \tilde{\Omega}$ and every $z \in \mathcal{Z}$,

$$\exists \lim_k \|x_k(\omega) - z\|.$$

Let $\bar{\Omega} := \tilde{\Omega} \cap \hat{\Omega}$. Then $\mathbb{P}(\bar{\Omega}) = 1$ and, for every $\omega \in \bar{\Omega}$, we have both that
- for every $z \in \mathcal{Z}$, $\exists \lim_k \|x_k(\omega) - z\|$;
- every cluster point of $x_k(\omega)$ belongs to $\mathcal{Z}$.

We conclude by the deterministic version of Opial's Lemma 9.3 that, for every $\omega \in \bar{\Omega}$, there exists $\bar{x}(\omega) \in \mathcal{Z}$ such that $x_k(\omega) \to \bar{x}(\omega)$. $\qquad\square$

Lemma 9.11 (Robbins-Siegmund [47]). *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a measure space, and let $(\mathcal{F}_k)$ be a filtration of $\mathcal{A}$. Let (r_k), (y_k), (w_k) and (β_k) be sequences of non-negative random variables adapted to $(\mathcal{F}_k)$. Let (β_k) and (w_k) belong to ℓ^1 a.s. and suppose that, for every $k \in \mathbb{N}$,*

$$\mathbb{E}\left[r_{k+1} \mid \mathcal{F}_k\right] - (1 + \beta_k)r_k + y_k \overset{a.s.}{\leq} w_k.$$

Then (r_k) converges a.s. to a random variable with non-negative values and $(y_k) \in \ell^1$ a.s.

Remark 9.12. Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ with λ-Lipschitz gradient and at least one minimizer. Applying the recursion of Algorithm 1 to such a function with an arbitrary starting point $x_0 \in \mathbb{R}^d$ we get that $\|x_k - x_*\|^2$, $\langle P_k P_k^\top \nabla f_k, x_k - x_* \rangle$, and $\|\nabla f_k\|^2$ are bounded for all $k > 0$. In particular, $\|\nabla f_k\|^2$, $\|x_k - x_*\|^2$, and $\langle P_k P_k^\top \nabla f_k, x_k - x_* \rangle$ are all integrable. To see this note first that ∇f_k and x_k are both measurable. Recall that for finite x_k, λ-Lipschitz gradient of f implies that $\|\nabla f_k\|^2 \leq \lambda^2 \|x_k - x_*\|^2$. Choose an arbitrary finite x_0 to begin the recursion. Then,

$$\|x_1 - x_*\|^2 \leq 2\|x_0 - x_*\|^2 + 2\alpha_0^2 \|P_0 \nabla_{P_0} f_0\|^2$$
$$\leq C_1 + C_2 \left\| P_0 \nabla_{P_0} f_0 - P_0 P_0^\top \nabla f_0 + P_0 P_0^\top \nabla f_0 \right\|^2$$
$$\text{Lemma 4.1} \leq C_1 + C_2 \left(C_3 + \left(\frac{d}{\ell}\right)^2 \|\nabla f_0\|^2 \right),$$

where C_1, C_2, C_3 are fixed and finite, the values are unimportant but can be calculated. Repeated recursion reveals that $\|x_k - x_*\|^2$ is bounded for all $k > 0$. Further, by λ-Lipschitz gradient, $\|\nabla f_k\|^2 \leq \lambda^2 \|x_k - x_*\|^2$. The claim on the inner product follows from Young's inequality and the previous results. Finally, since $\|\nabla f_k\|^2$, $\|x_k - x_*\|^2$, and $\langle P_k P_k^\top \nabla f_k, x_k - x_* \rangle$ are bounded, and measurable, they are integrable.

9.1 Special cases of the algorithm

We present several well-known special cases of Algorithm 1 (among which are: discrete gradient descent, discrete coordinate descent, spherical smoothing, and more), and provide some historical perspective on the development of these black-box algorithms.

Assumptions **P.1** and **P.2** describe matrices P that are generalizations of a few well-known cases. It is important that P is comprised of ℓ orthonormal columns, and in the special case that $d = \ell$, P is an orthogonal matrix. Specifically, when $d = \ell$, $PP^\top \nabla f(x) = \nabla f(x)$ and $\nabla_{(P,h)} f(x)$ is a forward finite difference estimate of the gradient along d orthogonal directions. Therefore gradient descent and discrete gradient descent can be viewed as special cases of (2.2) and Algorithm 1 respectively. Though this is of little significance in practice, it provides a means for verifying our analysis: by setting $\ell = d$ we ought to recover previously stated results for the discrete gradient method.

In our analysis we do not differentiate between various choices of P, all of our proofs hold whenever **P.1** and **P.2** are satisfied. It is clear that the specific choice of P does impact the performance of the algorithm (see, e.g., [7, 33]), but the purpose of this work is to present a unified convergence analysis rather than to investigate the nuances of each particular case. Therefore, in this section we present several choices for P that satisfy **P.1** and **P.2** with an emphasis on choices that correspond to previously described methods. Our results hold for all of the special cases described in this section, and in many of the cases our results represent an advancement over the current theoretical understanding of the special cases described.

Coordinate descent Suppose $P = (\sqrt{d/\ell})S$ where S consists of ℓ columns of the identity matrix chosen uniformly at random. In this case it is straightforward to see that $PP^\top \nabla f(x)$ corresponds to ℓ coordinate directions of the gradient scaled by a constant d/ℓ; that is, (2.2) with this choice of P is a scaling of block coordinate descent with uniform sampling of the coordinates. Analogously, $P\nabla_{(P,h)} f(x)$ is (up to a scaling constant) a forward finite difference estimate of the gradient along ℓ coordinate directions and we recover discretized coordinate descent.

Coordinate descent with a change of basis More generally, if P consists of ℓ columns selected uniformly at random from an orthogonal matrix in $\mathbb{R}^{d \times d}$, and scaled by $\sqrt{d/\ell}$ then P satisfies **P.1** and **P.2**. For instance, suppose that one selects ℓ columns uniformly at random from scaled versions of a discrete cosine transform matrix or the Hadamard matrix. The resulting $d \times \ell$ matrix defines a valid matrix P for our analysis. Any particular choice of fixed orthogonal matrix for P amounts to discretized coordinate descent in a different basis, but if little is known a priori about the structure of f then there is no reason to select one basis over another.

Random orthogonal matrices The orthogonal matrices mentioned in the previous paragraph have been coupled with a random component and used extensively to "sketch" problems. Typically, sketching entails approximating a problem by representing the data in a lower dimensional (random) subspace, and solving the approximate problem (see, e.g., [3, 19] for the least squares case, or [54, 36] for a more general overview of sketching algorithms in numerical linear algebra). The properties of the matrix that projects the data onto a subspace allow for guarantees on the quality of the approximated solution as compared to the true solution. These same matrices can instead be used to sketch the gradient using our method. For an example of such a sketching matrix consider a Hadamard matrix $H \in \mathbb{R}^{d \times d}$, a diagonal matrix $D \in \mathbb{R}^{d \times d}$ with equiprobable entries $\{1, -1\}$ along the diagonal, and a matrix $S \in \mathbb{R}^{d \times \ell}$ independent of D with columns chosen uniformly at random from the identity. Let $P = (\sqrt{d/\ell})DHS$, with D and S re-sampled at each iteration. This is

similar to the sketching matrix described in [19] and its properties are well known, it is simple to verify that it satisfies (**P.1**) and (**P.2**). This method is described in [15], however they provide only empirical results, making no claims about the convergence properties. To our knowledge our analysis is the first to provide convergence analysis for these types of matrices used in a derivative-free optimization setting.

Spherical smoothing   Consider instead $P = (\sqrt{d/\ell})\,Q\mathbb{I}_{d\times\ell}$, where Q is as in the QR-decomposition of a matrix $Z = QR \in \mathbb{R}^{d\times d}$ with $R_{ii} > 0$, and the entries of Z_k are iid $\mathcal{N}(0,1)$. The matrix $\mathbb{I}_{d\times\ell}$ truncates Q to its first ℓ columns so $Q\mathbb{I}_{d\times\ell}$ corresponds to ℓ columns of the random orthogonal matrix distributed according to the Haar measure on orthogonal matrices [38]. In other words, the columns $p^{(j)}$ are orthogonal and distributed uniformly on the sphere for all j. Thus, when $\ell = 1$, $P\nabla_{(P,h)}f(x)$ is a spherical smoothing estimate of the gradient, as described in, e.g., [22, 7]. The matrix Z is re-sampled at each iteration so, as with the matrices described in the previous paragraph, the basis changes with each iteration. In fact, sampling from the Haar measure on the set of orthogonal matrices corresponds to sampling uniformly from the set of orthogonal matrices.

For the case $\ell > 1$ it is more common in the literature [7] to sample $p^{(j)}$ independently and uniformly on the sphere, but in our case, to satisfy Assumptions (**P.1**) and (**P.2**), the columns of P must be orthonormal, consistent with [33]. The advantage of a matrix P with orthonormal columns is discussed at length in [33], we remark here merely that this property is required to obtain our results and to connect Algorithm 1 with discrete gradient descent when $\ell = d$; indeed consider that when $p^{(j)}$ are sampled independently and uniformly on the sphere, which we denote as $p^{(j)} \overset{iid}{\sim} \mathcal{U}(S(0,1))$, the gradient estimate is

$$\nabla f(x) \approx \frac{1}{\ell}\sum_{j=1}^{\ell} p^{(j)}\frac{f(x + p^{(j)}h) - f(x)}{h}, \qquad p^{(j)} \overset{iid}{\sim} \mathcal{U}(S(0,1)),$$

with the discrete gradient recovered only as $\ell \to \infty$. In contrast, the approximation

$$\nabla f(x) \approx P\begin{pmatrix} \frac{f(x_k + p^{(1)}h) - f(x)}{h} \\ \vdots \\ \frac{f(x_k + p^{(\ell)}h) - f(x)}{h} \end{pmatrix}, \qquad p^{(j)} \text{ columns of Haar}$$

and the discrete gradient is recovered whenever $\ell = d$ due to the orthogonality of the Haar distributed random matrix.

Gaussian smoothing   The Gaussian smoothing framework first described in the technical report [39] and later in [40] *does not* fit into our framework because the columns do not have unit norm. However, spherical smoothing, which is covered by our framework, can be thought of as a normalized version of Gaussian smoothing since a Haar distributed random matrix is generated by orthonormalizing a Gaussian random matrix. It is shown in [7] that spherical smoothing provides better approximations to the gradient than does Gaussian smoothing.

The intuition behind this statement is both illuminating and simple to provide. In Gaussian smoothing, $P = \sqrt{d}z$ where $z \sim \mathcal{N}(0,1)$. Then, $\nabla f(x) \approx (1/h)(f(x + zh) - f(x))z$. Since z has infinite support, the finite difference stepsize varies with each iteration irrespective of the value of h. Thus the approximation of the gradient has a positive probability of being arbitrarily bad even when the direction chosen is near to the direction of the gradient! With spherical smoothing, the directions chosen are identical, but the finite difference stepsize is always of length h which leads to more consistent and predictable results.

Remark 9.13. The literature is scarce but growing when it comes to convergence results for many of the above-mentioned methods in the finite difference setting. There has been plenty of attention to analyzing these methods when exact directional derivatives are available (i.e., the setting of (2.2), see, e.g., [55, 33, 5, 56]), but implementing the algorithms they analyze requires access to exact directional derivatives (e.g., via forward-mode automatic differentiation). Generally speaking, practical implementations of these algorithms often do not use automatic differentiation software – either because it is not feasible, or because it is too restrictive and time consuming – relying instead on function evaluations and finite difference approximations of the gradient.

9.2 Previous works

The limit definition of the derivative makes it natural to estimate the gradient via finite differences, the method of finite difference gradient descent goes back to Cauchy [14]. For a more modern treatment we can look to the seminal paper of Kiefer and Wolfowitz [30] which extends the results of Robbins and Monro [46] on stochastic approximation to the case where the gradient is approximated by a central finite difference. In [30] it is shown that with sufficiently fast decaying stepsize and finite difference step, the iterates converge asymptotically to the minimizer of a function under regularity conditions on the specified function. The setting of [30] differs from that of this paper by accessing only stochastic approximations of the function that is being minimized whereas we assume the function can be queried exactly, allowing for much stronger results.

Kushner and Clark [34, pg. 59-61] explore the asymptotic properties of what is now known as spherical smoothing, a special case of our algorithm. They work in a slightly different setting, adopting the noisy function evaluations case of Kiefer and Wolfowitz and making more assumptions on the objective function such as twice-differentiability. Again, the results are weaker and less general than those we provide. For their algorithm and under their regularity assumptions, Kushner and Clark are able to show that the iterates of their algorithm converge to a minimizer. Their analysis provided the basis for much subsequent work. Spall [51] compares his work to that of Kushner and Clark with the notable difference being that Spall does not choose directions uniformly on the sphere, but from a more general, unspecified, mean-zero distribution; this is perhaps the clearest intellectual predecessor to Gaussian smoothing, which is discussed below, because it includes Gaussian smoothing as a special case. The generality of Spall's results requires him to assume thrice-differentiability of the objective function. A main beneficiary of the work of Kushner and Clark is the reinforcement learning community which has adopted their method and renamed it *evolutionary strategies*. These evolutionary strategies were first described by Williams [53], who called them REINFORCE algorithms. Williams was apparently unaware of the work of Kushner and Clark and suggested that "While there is a clear need for an analytical characterization of the asymptotic behavior of REINFORCE algorithms, such results are not yet available, leaving simulation studies as our primary source of understanding of the behaviour of these algorithms". Subsequent literature that makes use of evolutionary strategies (e.g., [15, 48]) typically mention Williams as the forebear for these methods, while some mention Spall's work in providing convergence properties, and many mention the work of Nesterov [39] discussed below. Of particular note is [22], which works in the setting of Kushner and Clark but frames it as a reinforcement learning problem and provides finite-time results in expectation. See also [1].

The asymptotic behavior of discrete gradient descent with exact function queries is investigated in [56] which provides an upper bound on the level sets of the limiting function values, as well as a radius of convergence for the iterates. The setting is somewhat restrictive, with the objective f assumed to be γ-strongly convex with λ Lipschitz gradient, and only asymptotic properties are provided. A method for performing finite difference coordinate descent is described and analyzed in [56], however the method described therein

uses coordinates only as directions, no approximate derivative (i.e., finite difference) information is used.

Nesterov published a technical note [39] in 2011, and later a peer-reviewed article with Spokoiny [40] analyzing the convergence behaviour of a finite difference optimization algorithm in which the direction of descent is chosen according to a Gaussian distribution in the following manner. Let $U \sim \mathcal{N}(0, \mathbb{I}_d)$, then the direction of descent is $U(f(x + Uh) - f(x))/h \approx \nabla f(x)$.

Though they do not analyze any particular algorithm, Berahas et al. [7] provide a thorough comparison of the quality of various derivative-free approximations of the gradient. Included in their analysis are finite difference (and by simple corollary, coordinate descent), Gaussian smoothing, and spherical smoothing. Their work provides the framework required for analysis of any gradient-based descent algorithm based on finite difference approximations, analysis that is particularly useful in the related paper [8] which describes a line search method appropriate for gradient descent algorithms when the gradient is known only approximately. Expected rates of convergence are provided in [8] under a variety of convexity assumptions when the gradient is approximated using any of the methods discussed in [7].

The algorithm analyzed in [33] is identical to (2.2). They discuss but do not analyze Algorithm (1), which is the primary focus of our work. Furthermore, we provide stronger results in the convex case using (2.2): we prove almost sure convergence of the iterates to a minimizer (cf. Theorem 5.12) whereas in [33] only expected results are provided, and only for the function values.

Trust-region methods are a class of derivative-free optimization algorithms that we do not explore here, we note merely that recent work analyzes a stochastic subspace method that is analogous to ours for the trust-region framework, specifically for non-linear least squares problems [12] .

Acknowledgements