On the Convergence of Prior-Guided Zeroth-Order Optimization Algorithms

Anonymous Author(s)
Affiliation
Address
email

Abstract

Zeroth-order (ZO) optimization is widely used to handle challenging tasks, such as query-based black-box adversarial attacks and reinforcement learning. Various attempts have been made to integrate prior information into the gradient estimation procedure based on finite differences, with promising empirical results. However, their convergence properties are not well understood. This paper makes an attempt to fill this gap by analyzing the convergence of prior-guided ZO algorithms under a greedy descent framework with various gradient estimators. We provide a convergence guarantee for the prior-guided random gradient-free (PRGF) algorithms. Moreover, to further accelerate over greedy descent methods, we present a new accelerated random search (ARS) algorithm that incorporates prior information, together with a convergence analysis. Finally, our theoretical results are confirmed by experiments on several numerical benchmarks as well as adversarial attacks.

1 Introduction

Zeroth-order (ZO) optimization [17] provides powerful tools to deal with challenging tasks, such as query-based black-box adversarial attacks [6, 12], reinforcement learning [21, 16, 8], meta-learning [1], and hyperparameter tuning [22], where the access to gradient information is either not available or too costly. ZO methods only assume an oracle access to *the function value* at any given point, instead of gradients as in first-order methods. The primary goal is to find a solution with as few queries to the function value oracle as possible. Recently, various ZO methods are proposed in two main categories. One type is to obtain a gradient estimator and plug in some gradient-based methods. [20] analyzes the convergence of such methods with a random gradient-free (RGF) estimator obtained via finite difference along a random direction. The other type is directed search, where the update only depends on comparison between function values at different points. Such methods are robust against monotone transform of the objective function [23, 9, 3]. However, as they do not directly utilize function values, their query complexity is often larger than that of finite difference methods. Therefore, we focus on the first type of methods in this paper. Other methods exist such as CMA-ES [10] which is potentially better on objective functions with a rugged landscape, but typically lacks a convergence guarantee.

ZO methods are usually less efficient than first-order algorithms, as they typically take $O(d)$ queries to reach a given precision, where d is the input dimension (see Table 1 in [9]). Specifically, for the methods in [20], the oracle query count is at least d times larger than that of their corresponding schemes using gradients. This inefficiency stems from random search along uniformly distributed directions. To improve, various attempts have been made to augment random search with (extra) prior

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

information. For instance, [13, 18] use a time-dependent prior (i.e., the gradient estimated in the last step), while [15, 7, 4] use surrogate gradients obtained from other sources.[1] Among these methods, [13, 15, 7, 18] propose an objective function for designing a good gradient estimator to justify their choice or to choose optimal hyperparameters in the gradient estimator; and [18] uses a subspace estimator that maximizes the squared cosine similarity with the true gradient and finds a better descent direction than both the prior direction and the randomly sampled direction. However, all these methods treat gradient estimation and the optimization algorithm separately, and *it remains unclear whether these gradient estimators and the corresponding prior-exploiting optimization methods are theoretically sound, and what role the prior plays in accelerating convergence.*

In this paper, we attempt to answer these questions by establishing a formal connection between convergence rates and the quality of gradient estimates. Further, we develop a more efficient ZO algorithm with prior inspired by a theoretical analysis. First, we present a greedy descent framework of ZO methods and provide its convergence rate under smooth convex optimization, which is positively related to the squared cosine similarity between the gradient estimate and true gradient. As shown by [18] and our results, given some finite-difference queries, the optimal estimator maximizing the squared cosine similarity is the projection of true gradient on the subspace spanned by those queried directions. In the case with prior information, a natural such estimator is in the same form as the one in [18] (but we allow different priors). We call it PRGF estimator and analyze its convergence rate.[2] Our results show that no matter what the prior is, the convergence rate of PRGF is at least the same as that of the RGF baseline[3] [20, 14], and it could be significantly improved if given a useful prior. Such results shed light on exploring prior information in ZO optimization to accelerate convergence.

Then, as a concrete example, we apply the analysis to History-PRGF [18], which uses historical information (i.e., gradient estimate in the last iteration) as the prior. [18] presents analysis on linear functions, yet still lacking a convergence rate. We analyze on general L-smooth functions, and find that when the learning rate is smaller than the optimal value $1/L$ in smooth optimization, the expected squared cosine similarity could converge to a larger value, which compensates for the slowdown of convergence brought by the inappropriate choice of learning rate. We also show that History-PRGF admits a convergence rate independent of learning rate as long as it is in a fairly wide range.

Finally, to further accelerate greedy descent methods, we present Prior-Guided ARS (PARS), a new variant of Accelerated Random Search (ARS) [20] to utilize prior information. Technically, the extension from ARS to PARS is non-trivial, as directly replacing gradient estimator in ARS by PRGF estimator would lose the convergence analysis. Thus, we present necessary extensions to ARS, and show that PARS has a convergence rate no worse than that of ARS and admits potential acceleration given a good prior. In particular, when the prior is chosen as the historical information, the resulting History-PARS is robust to learning rate in experiments. To our knowledge, History-PARS is the first ARS-based method that is empirically robust while retaining the convergence rate as ARS.

Our experiments on numerical benchmarks and adversarial attacks confirm the theoretical results.

2 Setups

Assumptions on the problem class We consider unconstrained optimization, where the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-smooth for $L \geq 0$. Optionally, we require f to be τ-strongly convex for $\tau > 0$. We leave definitions of these concepts to Appendix A.1.

Directional derivative oracle In ZO optimization, we follow the finite difference approach, which makes more use of the queried function values than direct search and provides better gradient

[1] Some work [6, 13, 24] restricts the random search to a more effective subspace reflecting the prior knowledge. But, this eliminates the possibility of convergence to the optimal solution.

[2] Note that the estimator is different from the P-RGF estimator in [7].

[3] In this paper, RGF and PRGF could refer to either a gradient estimator or the greedy descent algorithm with the corresponding estimator, depending on the context.

Algorithm 1 Greedy descent framework

Input: L-smooth convex function f; initialization x_0; upper bound $\hat{L}$ ($\hat{L} \geq L$); iteration number T.
Output: x_T as the approximate minimizer of f.

```
for t = 0 to T − 1 do
    Let v_t be a random vector s.t. ‖v_t‖ = 1;
    x_{t+1} ← x_t − (1/L̂) g_t, where g_t ← ∇f(x_t)^⊤ v_t · v_t;
end for
return x_T.
```

approximations than alternatives [2]. In particular, we consider the forward difference method:

$$g_\mu(v; x) := \frac{f(x + \mu v) - f(x)}{\mu} \approx \nabla f(x)^\top v, \tag{1}$$

where v is a vector with unit ℓ_2 norm $\|v\| = 1$ and μ is a small positive step. As long as the objective function is smooth, the error between the finite difference and the directional derivative could be uniformly bounded, as shown in the following proposition (see Appendix A.2 for its proof).

Proposition 1. *If f is L-smooth, then for any (x, v) with $\|v\| = 1$, $|g_\mu(v; x) - \nabla f(x)^\top v| \leq \frac{1}{2} L\mu$.*

Thus, in smooth optimization, the error brought by finite differences to the convergence bound can be analyzed in a principled way, and its impact tends to zero as $\mu \to 0$. We also choose μ as small as possible in practice. Hence, in the following analysis we directly assume the **directional derivative oracle**: suppose that we can obtain $\nabla f(x)^\top v$ for any (x, v) in which $\|v\| = 1$ with one query.

3 Greedy descent framework and PRGF algorithm

We now introduce a greedy descent framework in ZO optimization which can be implemented with various gradient estimators. We first provide a general analysis, followed by a concrete example. All the proofs in this section are presented in Appendix B.

3.1 The greedy descent framework and general analysis

In first-order smooth convex optimization, a sufficient single-step decrease of the objective can guarantee convergence (see Sec. 3.2 in [5]). Inspired by this fact, we design the update in an iteration to greedily seek for maximum decrease. Suppose we are currently at x, and want to update along the direction v. Without loss of generality, assume $\|v\| = 1$ and $\nabla f(x)^\top v > 0$. To choose a suitable step size η that minimizes $f(x - \eta v)$, we note that

$$f(x - \eta v) \leq f(x) - \eta \nabla f(x)^\top v + \frac{1}{2} L\eta^2 := F(\eta) \tag{2}$$

by smoothness of f. For the r.h.s, we have $F(\eta) = f(x) - \frac{(\nabla f(x)^\top v)^2}{2L}$ when $\eta = \frac{\nabla f(x)^\top v}{L}$, which minimizes $F(\eta)$. Thus, choosing such η could guarantee maximum decrease of $f(x - \eta v)$ from $f(x)$. In practice, the value of L is often unknown, but we can verify that as long as $0 < \eta \leq \frac{\nabla f(x)^\top v}{L}$, then $f(x - \eta v) \leq F(\eta) < f(x)$, i.e., we can guarantee decrease of the objective (regardless of the direction of v if $\nabla f(x)^\top v > 0$). Based on the above discussion, we further allow v to be random and present the greedy descent framework in Alg. 1.

Remark 1. *If v_t is uniformly sampled from the unit sphere $\mathbb{S}^{d-1}$, then Alg. 1 is similar to the simple random search in [20] except that v_t is sampled from a Gaussian distribution there.*

Remark 2. *The distribution of v_t could depend on the history (i.e., the randomness before step t). For example, v_t can be biased towards a vector p_t that corresponds to prior information, and p_t depends on x_t or the optimization trajectory. Specifically, v_t can be expressed as $v_t = f_t(p_t, \xi_t)$ where p_t depends on the history, $\{\xi_1, \xi_2, \cdots, \xi_T\}$ are independent random vectors, and f_t is a deterministic mapping (which could be dependent on the history).*

By Remark 2, we introduce $\mathbb{E}_t[\cdot]$ to denote the conditional expectation given the history. Note $\mathbb{E}_t[\cdot]$ only takes expectation w.r.t. the randomness in step t. We let $\overline{v} := v/_{\|v\|}$ denote the ℓ_2 normalization

of vector v; $x^* := \operatorname{argmin}_x f(x)$ denote the minimizer of f; and $\delta_t := f(x_t) - f(x^*)$. Thanks to the descent property in Alg. 1, we have the following lemma on single step progress.

Lemma 1. *Let* $C_t := \left(\overline{\nabla f(x_t)}^\top v_t\right)^2$ *and* $L' := \frac{L}{1-(1-\frac{L}{\hat{L}})^2}$*, then in Alg. 1, we have*

$$\mathbb{E}_t[\delta_{t+1}] \leq \delta_t - \frac{\mathbb{E}_t[C_t]}{2L'}\|\nabla f(x_t)\|^2. \tag{3}$$

We note that $L' = \hat{L}/(2-L/\hat{L})$, so $\hat{L}/2 \leq L' \leq \hat{L}$ and $L' \geq L$.

To obtain a bound on $\mathbb{E}[\delta_T]$, one of the classical proofs (see e.g., Theorem 1 in [19]) requires us to take expectation on both sides of Eq. (3). Allowing the distribution of v_t to be dependent on the history leads to additional technical difficulty: in the r.h.s of Eq. (3), $\mathbb{E}_t[C_t]$ becomes a random variable that is not independent of $\|\nabla f(x_t)\|^2$. Thus, we cannot simplify the term $\mathbb{E}[\mathbb{E}_t[C_t]\|\nabla f(x_t)\|^2]$ if we take expectation. By using some techniques, we obtain the following main theorems on convergence rate.

Theorem 1 (Alg. 1, smooth and convex)**.** *Let* $R := \max_{x:f(x)\leq f(x_0)} \|x - x^*\|$ *and suppose* $R < \infty$*. Then, in Alg. 1, we have*

$$\mathbb{E}[\delta_T] \leq \frac{2L'R^2 \sum_{t=0}^{T-1} \mathbb{E}\left[\frac{1}{\mathbb{E}_t[C_t]}\right]}{T(T+1)}. \tag{4}$$

Theorem 2 (Alg. 1, smooth and strongly convex)**.** *If f is also τ-strongly convex, then we have*

$$\mathbb{E}\left[\frac{\delta_T}{\exp\left(-\frac{\tau}{L'}\sum_{t=0}^{T-1}\mathbb{E}_t[C_t]\right)}\right] \leq \delta_0. \tag{5}$$

Remark 3. *In the strongly convex case, currently we cannot directly obtain a final bound of* $\mathbb{E}[\delta_T]$*. However, the form of Eq.* (5) *is still useful since it roughly tells us that* δ_T *converges as the denominator* $\exp\left(-\frac{\tau}{L'}\sum_{t=0}^{T-1}\mathbb{E}_t[C_t]\right)$*. In Sec. 3.2, when we analyze History-PRGF, a special implementation in the greedy-descent framework, we will turn this intuition into a rigorous theorem since in that case we can prove that the denominator has a nice concentration property.*

Remark 4. *If we have a lower bound of* $\mathbb{E}_t[C_t]$*, e.g.* $\mathbb{E}_t[C_t] \geq a > 0$*, then we directly obtain that* $\mathbb{E}[\delta_T] \leq \frac{2L'R^2}{a(T+1)}$ *for Thm. 1, and* $\mathbb{E}[\delta_T] \leq \delta_0 \exp(-\frac{\tau}{L'}aT)$ *for Thm. 2. These could recover the ideal convergence rate for RGF estimator and the worst-case convergence rate for PRGF estimator, as explained in Exp. 1 and 2.*

From Theorems 1 and 2, we see that a larger value of $\mathbb{E}_t[C_t]$ would lead to a better bound. To find a good choice of v_t in Alg. 1, it becomes natural to discuss the following problem. Suppose in an iteration in Alg. 1, we query the directional derivative oracle at x_t along q directions $\{u_i\}_{i=1}^q$ (maybe randomly chosen) and obtain the values of $\{\nabla f(x_t)^\top u_i\}_{i=1}^q$. We could use this information to construct a random vector v_t. We note that $\mathbb{E}_t[C_t] = \mathbb{E}_{u_1,\ldots,u_q}[\mathbb{E}_{v_t}[C_t]]$, so what is the v_t that maximizes $\mathbb{E}_{v_t}[C_t|u_1,\ldots,u_q]$? To answer this question, we prove the following proposition based on Prop. 1 in [18] and additional justification.

Proposition 2 (Optimality of subspace estimator)**.** *In one iteration of Alg. 1, if we have queried* $\{\nabla f(x_t)^\top u_i\}_{i=1}^q$*, then the optimal* v_t *maximizing* $\mathbb{E}_t[C_t]$ *should be in the following form:* $v_t = \overline{\nabla f(x_t)_A}$*, where* $A := \operatorname{span}\{u_1, u_2, \ldots, u_q\}$ *and* $\nabla f(x_t)_A$ *denote the projection of* $\nabla f(x_t)$ *onto* A*.*

Note that in Line 3 of Alg. 1, we have $g_t = \nabla f(x_t)^\top \overline{\nabla f(x_t)_A} \cdot \overline{\nabla f(x_t)_A} = \nabla f(x_t)_A$. Therefore, the gradient estimator g_t is equivalent to the projection of the gradient to the subspace A, which justifies its name of subspace estimator. Next we discuss some special cases of the subspace estimator. We leave detailed derivation in following examples to Appendix B.3.

Example 1 (RGF)**.** $u_i \sim \mathcal{U}(\mathbb{S}^{d-1})$ *for* $i = 1, 2, \ldots, q$*. Without loss of generality, we assume they are orthonormal (e.g., via Gram-Schmidt orthogonalization). The corresponding estimator* $g_t = \sum_{i=1}^q \nabla f(x_t)^\top u_i \cdot u_i$ *(*$v_t = \overline{g_t}$*). When* $q = 1$*, the estimator is similar to the random gradient-free oracle in [20]. When* $q \geq 1$*, it is the same as the stochastic subspace estimator [14] and similar*

to the orthogonal ES estimator [8]. We can prove that $\mathbb{E}_t[C_t] = \frac{q}{d}$. By Theorems 1 and 2, the convergence rate is $\mathbb{E}[\delta_T] \leq \frac{2L'R^2\frac{d}{q}}{T+1}$ for smooth convex case, and $\mathbb{E}[\delta_T] \leq \delta_0 \exp(-\frac{\tau}{L'}\frac{q}{d}T)$ for smooth and strongly convex case. The bound is the same as that in [14]. Since the query complexity in each iteration is proportional to q, the bound for total query complexity is indeed independent of q.

Example 2 (PRGF)**.** *With slight notation abuse, we assume the subspace in Prop. 2 is spanned by $\{p_1, \ldots, p_k, u_1, \ldots, u_q\}$, so each iteration takes $q + k$ queries. Let $p_1, \cdots, p_k$ be k vectors corresponding to the prior message (e.g. the historical update, or the gradient of a surrogate model) which depends on the history, and $u_i \sim \mathcal{U}(\mathbb{S}^{d-1})$ for $i = 1, 2, \ldots, q$. We note that intuitively we cannot construct a better subspace to maximize $\mathbb{E}_t[C_t]$ since the only additional information we know is the k priors. In our analysis, we assume $k = 1$ for convenience, and we change the original notation p_1 to p_t to explicitly show the dependence of p_t on the history. Without loss of generality, we assume $\{p_t, u_1, \ldots, u_q\}$ are orthonormal (e.g., via Gram-Schmidt orthogonalization). The corresponding estimator $g_t = \nabla f(x_t)^\top p_t \cdot p_t + \sum_{i=1}^q \nabla f(x_t)^\top u_i \cdot u_i$, which is similar to the estimator in [18]. By [18] (the expected drift of X_t^2 in its Thm. 1), we have*

Lemma 2. *In Alg. 1 with PRGF estimator, $\mathbb{E}_t[C_t] = D_t + \frac{q}{d-1}(1 - D_t)$ where $D_t := (\overline{\nabla f(x_t)}^\top p_t)^2$.*

Hence $\mathbb{E}_t[C_t] \geq \frac{q}{d}$ holds. By Remark 4, PRGF admits a guaranteed convergence rate of RGF and is potentially better given a good prior (if D_t is large), but it costs an additional query per iteration.

3.2 Analysis on the PRGF algorithm with the historical prior

We apply the above analysis to a concrete example of the History-PRGF estimator [18], which considers the historical prior in the PRGF estimator. In this case, Lem. 1, Thm. 1 and Thm. 2 will manifest themselves by clearly stating the convergence rate which is robust to the learning rate.

Specifically, History-PRGF considers historical prior as follows: we choose $p_t = \overline{g_{t-1}}$, i.e., we let the prior be the direction of the previous gradient estimate.[4] This is equivalent to letting $p_t = v_{t-1}$. Thus, in Alg. 1, $v_t = \overline{\nabla f(x_t)_A} = \overline{\nabla f(x_t)^\top v_{t-1} \cdot v_{t-1} + \sum_{i=1}^q \nabla f(x_t)^\top u_i \cdot u_i}$. In this form we require $\{v_{t-1}, u_1, \ldots, u_q\}$ to be orthonormal, so we first determine v_{t-1}, and then sample $\{u_i\}_{i=1}^q$ in $A_\perp$, the $(d-1)$-dimensional subspace of $\mathbb{R}^d$ perpendicular to v_{t-1}, and then do Gram-Schmidt orthonormalization on $\{u_i\}_{i=1}^q$.

To study the convergence rate, we first study evolution of C_t under a general L-smooth function. This extends the analysis on linear functions (corresponding to $L = 0$) in [18]. Under the framework of Alg. 1, intuitively, the change of the gradient should be smaller when the objective function is very smooth (L is small) or the learning rate is small ($\hat{L}$ is large). Since we care about the inner product between the gradient and the prior, we prove the following lemma:

Lemma 3. *Let $D_t := (\overline{\nabla f(x_t)}^\top v_{t-1})^2$. Then $D_t \geq (1 - L/\hat{L})^2 C_{t-1}$.*

When $\hat{L} = L$, i.e., using the optimal learning rate, Lem. 3 does not provide a useful bound, since an optimal learning rate in smooth optimization could find an approximate minimizer along the update direction, so the update direction may be useless in next iteration. In this case, the historical prior does not provide acceleration. Hence, Lem. 3 explains the empirical findings in [18] that past directions can be less useful when the learning rate is larger. However, in practice we often use a conservative learning rate, for the following reasons: 1) We usually do not know L, and the cost of tuning learning rate could be large; 2) Even if L is known, it only provides a global bound, so a fixed learning rate could be too conservative in the smoother local regions. In scenarios where $\hat{L}$ is too conservative ($\hat{L} > L$), History-PRGF could bring more acceleration over RGF.

By Lem. 3, we can assume $D_t = (1 - L/\hat{L})^2 C_{t-1}$ to obtain a lower bound of quantities about C_t. Meanwhile, since D_t means quality of the prior, the construction in Exp. 2 tells us relationship between C_t and D_t. Then we have full knowledge of the evolution of C_t, and thus $\mathbb{E}_t[C_t]$. In Appendix B.4, we discuss about evolution of $\mathbb{E}[C_t]$ and show that $\mathbb{E}[C_t] \to O(\frac{q}{d}\frac{L'}{L})$ if $\frac{q}{d} \leq \frac{L}{\hat{L}} \leq 1$.

[4] One can also utilize multiple historical priors (e.g. the last k updates with $k > 1$, as proposed and in experiments of [18]), but here we only analyze the $k = 1$ case.

Therefore, by Lem. 3, assuming $\mathbb{E}_t[C_t] \approx \mathbb{E}[C_t]$ and hence $\frac{\mathbb{E}_t[C_t]}{L'} \approx \frac{q}{dL}$, PRGF could recover the single step progress with optimal learning rate ($\hat{L} = L$), since Eq. (3) only depends on $\frac{\mathbb{E}_t[C_t]}{L'}$ which is constant w.r.t. L' now. While the above discussion is informal, based on Thm. 1 and Thm. 2, we prove following theorems which show that convergence rate of History-PRGF is robust to choice of learning rate.[5]

Theorem 3 (History-PRGF, smooth and convex)**.** *In the setting of Thm. 1, assuming $d \geq 4$, $\frac{q}{d-1} \leq \frac{L}{\hat{L}} \leq 1$ and $T > \left\lceil \frac{d}{q} \right\rceil$ ($\lceil \cdot \rceil$ denotes the ceiling function), we have*

$$\mathbb{E}[\delta_T] \leq \left(\frac{32}{q} + 2\right) \frac{2L\frac{d}{q}R^2}{T - \left\lceil \frac{d}{q} \right\rceil + 1}. \tag{6}$$

Remark 5. *As stated in Exp. 2, using RGF with optimal learning rate, we have $\mathbb{E}[\delta_T] \leq \frac{2L\frac{d}{q}R^2}{T+1}$. Therefore, History-PRGF with a suboptimal learning rate $\frac{1}{\hat{L}}$ under the condition $\frac{q}{d-1}\frac{1}{L} \leq \frac{1}{\hat{L}} \leq \frac{1}{L}$ could reach similar convergence rate to RGF with optimal learning rate (up to constant factors), which indicates that History-PRGF is more robust to learning rate than RGF.*

Theorem 4 (History-PRGF, smooth and strongly-convex)**.** *Under the same conditions as in Thm. 2, then assuming $d \geq 4$, $\frac{q}{d-1} \leq \frac{L}{\hat{L}} \leq 1$, $\frac{q}{d} \leq 0.2\frac{L}{\tau}$, and $T \geq 5\frac{d}{q}$, we have*

$$\mathbb{E}[\delta_T] \leq 2\exp(-0.1\frac{q}{d}\frac{\tau}{L}T)\delta_0. \tag{7}$$

Remark 6. *As in Exp. 2, using RGF with optimal learning rate, we have $\mathbb{E}[\delta_T] \leq \exp\left(-\frac{q}{d}\frac{\tau}{L}T\right)\delta_0$. Similar to Rem. 5, the theorem shows that History-PRGF is robust to learning rate.*

Note that if we adopt line search in Alg. 1, then one can reach convergence guarantee with robustness against a larger range of the choice of learning rate. However, it requires at least several more queries per iteration, while History-PRGF only requires one. Of course, when q is large, this practical advantage decreases. Nevertheless, it is still an intriguing fact that History-PRGF could reach the similar performance with methods adapting the learning rate, while its mechanism is very different. When there are cases in which a small q is required or line search does not work (the latter arises in stochastic optimization), History-PRGF could be practically better than line search.

4 Extension of ARS framework and PARS algorithm

To further accelerate greedy descent methods, we extend our analysis to a new variant of Accelerated Random Search (ARS) [20] by incorporating prior information, under the smooth and convex setting.[6] All the proofs in this section are presented in Appendix C.

By delving into the proof in [20], we present our extension to ARS in Alg. 2, state its convergence guarantee in Thm. 5 and explain its design in the proof sketch in Appendix C.1.

Theorem 5. *In Alg. 2, we have*

$$\mathbb{E}\left[(f(x_T) - f(x^*))\left(1 + \frac{\sqrt{\gamma_0}}{2}\sum_{t=0}^{T-1}\sqrt{\theta_t}\right)^2\right] \leq f(x_0) - f(x^*) + \frac{\gamma_0}{2}\|x_0 - x^*\|^2. \tag{8}$$

Remark 7. *If we let $g_1(y_t)$ be the RGF estimator in Exp. 1 and let $g_2(y_t) = d \cdot g_1(y_t)$, then we can show that $\mathbb{E}[g_2(y_t)] = \nabla f(y_t)$ and θ_t could be chosen as $\frac{q^2}{\hat{L}d^2}$ since $\frac{\mathbb{E}_t\left[\left(\nabla f(y_t)^\top v_t\right)^2\right]}{\hat{L}\cdot\mathbb{E}_t[\|g_2(y_t)\|^2]} = \frac{q^2}{\hat{L}d^2}$. Then roughly, the convergence rate $\propto q$, so the total query complexity is independent of q. When $q = 1$, ARS baseline is recovered. For convenience we call the algorithm ARS regardless of the value of q.*

[5] The detailed proofs are technically complicated (see Appendix B.5 for the proof sketch). Besides, we note that the constants in the bound are loose and have a large potential to be improved in future work.

[6] The procedure of ARS requires knowledge of the strong convexity parameter τ (τ can be 0), but for clarity we only discuss the case $\tau = 0$ here (i.e., assume the objective function is not strongly convex), and leave the strongly convex case to Appendix C.6.

Algorithm 2 Extended accelerated random search framework[7]

Input: L-smooth black-box function f; initialization x_0; $\hat{L} \geq L$; iteration number T; $\gamma_0 > 0$.
Output: x_T as the approximate minimizer of f.

1: $m_0 \leftarrow x_0$;
2: **for** $t = 0$ to $T-1$ **do**
3: Find a $\theta_t > 0$ such that $\theta_t \leq \frac{\mathbb{E}_t\left[\left(\nabla f(y_t)^\top v_t\right)^2\right]}{\hat{L}\cdot\mathbb{E}_t[\|g_2(y_t)\|^2]}$ where y_t, v_t and $g_2(y_t)$ is defined in following 3 steps:
4: Step 1: $y_t \leftarrow (1-\alpha_t)x_t + \alpha_t m_t$, where $\alpha_t \geq 0$ is a positive root of the equation $\alpha_t^2 = \theta_t(1-\alpha_t)\gamma_t$; $\gamma_{t+1} \leftarrow (1-\alpha_t)\gamma_t$;
5: Step 2: Let v_t be a random vector s.t. $\|v_t\| = 1$; $g_1(y_t) \leftarrow \nabla f(y_t)^\top v_t \cdot v_t$;
6: Step 3: Let $g_2(y_t)$ be an unbiased estimator of $\nabla f(y_t)$, i.e., $\mathbb{E}_t[g_2(y_t)] = \nabla f(y_t)$;
7: $x_{t+1} \leftarrow y_t - \frac{1}{\hat{L}} g_1(y_t)$, $m_{t+1} \leftarrow m_t - \frac{\theta_t}{\alpha_t} g_2(y_t)$;
8: **end for**
9: **return** x_T.

Remark 8. *If we have a uniform constant lower bound $\theta > 0$ such that $\forall t, \theta_t \geq \theta$, then we have*

$$\mathbb{E}\left[f(x_T) - f(x^*)\right] \leq \frac{f(x_0) - f(x^*) + \frac{\gamma_0}{2}\|x_0 - x^*\|^2}{\left(1 + \frac{\sqrt{\gamma_0}}{2} T\sqrt{\theta}\right)^2}. \tag{9}$$

Next we present Prior-Guided ARS (PARS) by specifying the choice of $g_1(v_t)$ and $g_2(v_t)$ in Alg. 2 when prior information $p_t \in \mathbb{R}^d$ is available. Since we want to maximize the value of θ_t, regarding $g_1(y_t)$ we want to maximize $\mathbb{E}_t\left[\left(\nabla f(y_t)^\top v_t\right)^2\right]$. By Prop. 2 and Exp. 2, it is natural to let $g_1(y_t)$ be the PRGF estimator for $\nabla f(y_t)$. Then by Lem. 2, we have $\mathbb{E}_t\left[\left(\nabla f(y_t)^\top v_t\right)^2\right] = \|\nabla f(y_t)\|^2(D_t + \frac{q}{d-1}(1-D_t))$, where $D_t := (\overline{\nabla f(y_t)}^\top p_t)^2$. The remaining problem is to construct $g_2(y_t)$, an unbiased estimator of $\nabla f(y_t)$ ($\mathbb{E}_t[g_2(y_t)] = \nabla f(y_t)$), and make $\mathbb{E}_t[\|g_2(v_t)\|^2]$ as small as possible. We leave the construction of $g_2(v_t)$ in Appendix C.2. Finally, we calculate the following expression which appears in Line 3 of Alg. 2 to complete the description of PARS:

$$\frac{\mathbb{E}_t\left[\left(\nabla f(y_t)^\top v_t\right)^2\right]}{\hat{L}\cdot\mathbb{E}_t[\|g_2(y_t)\|^2]} = \frac{D_t + \frac{q}{d-1}(1-D_t)}{\hat{L}\left(D_t + \frac{d-1}{q}(1-D_t)\right)}. \tag{10}$$

Since $D_t \geq 0$, the right-hand side is larger than $q^2/\hat{L}d^2$ (by Remark 7, this value corresponds to the value of θ_t in ARS), so by Remark 8 PARS is guaranteed a convergence rate of ARS.

In implementation of PARS, we note that there remain two problems to solve. The first is that D_t is not accessible through one oracle query, since $D_t = (\overline{\nabla f(y_t)}^\top p_t)^2 = \left(\nabla f(y_t)^\top p_t / \|\nabla f(y_t)\|\right)^2$, and $\|\nabla f(y_t)\|$ requires estimation. Fortunately, the queries used to construct $g_1(y_t)$ and $g_2(y_t)$ can also be used to estimate D_t. With a moderate value of q, we can prove that considering the error brought by estimation of D_t, a modified version of PARS is guaranteed to converge with high probability. We leave related discussion to Appendix C.3. The second is that Line 3 of Alg. 2 has a subtlety that y_t depends on θ_t, so we cannot directly determine an optimal θ_t satisfying $\theta_t \leq \frac{\mathbb{E}_t\left[\left(\nabla f(y_t)^\top v_t\right)^2\right]}{\hat{L}\cdot\mathbb{E}_t[\|g_2(y_t)\|^2]}$. Theoretically, we can guess a conservative estimate of θ_t and verify this inequality, but in practice we adopt a more aggressive strategy to find an approximate solution of θ_t. We leave the actual implementation, named PARS-Impl, in Appendix C.4.

In PARS, if we adopt the historical prior as in Sec. 3.2, i.e., letting p_t be the previous PRGF gradient estimator $\overline{g_1(y_{t-1})}$, then we arrive at a novel algorithm named History-PARS. Here, we note that unlike the case in History-PRGF, it is more difficult to derive the evolution of θ_t theoretically, so we currently cannot prove theorems corresponding to Thm. 3. However, History-PARS can be

[7]Keys in this extension are: 1) We require $\mathbb{E}_t[g_2(y_t)] = \nabla f(y_t)$. Thus, $g_2(y_t)$ could not be the PRGF estimator as it is biased towards the prior; 2) To accelerate convergence, we need to find an appropriate θ_t since Eq. (8) depends on it. If we set θ_t to the value in ARS baseline, then no potential acceleration is guaranteed.

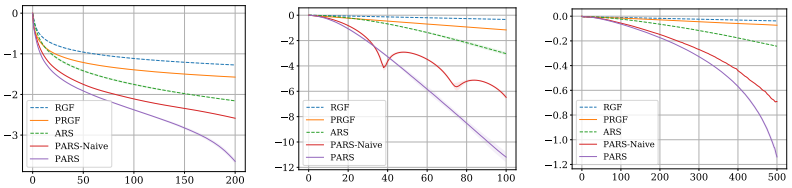
guaranteed the convergence rate of ARS, which is desirable since if we apply line search in ARS to reach robustness against learning rate (e.g. in [23]), currently there is no convergence guarantee. We present the actual implementation History-PARS-Impl in Appendix C.5 and empirically verify that History-PARS-Impl is robust to learning rate in Sec. 5.

5 Experiments

5.1 Numerical benchmarks

We first experiment on several closed-form test functions to support our theoretical claims. We leave more details of experimental settings to Appendix D.

First, we present experimental results when a general useful prior is provided. The prior-guided methods include PRGF, PARS (refer to PARS-Impl) and PARS-Naive (simply replacing the RGF estimator in ARS with the PRGF estimator). We adopt the setting in Sec. 4.1 of [15] in which the prior is a biased version of the true gradient. Our test functions are as follows: 1) f_1 is the "worst-case smooth convex function" used to construct the lower bound complexity of first-order optimization, as in [20]; 2) f_2 is a simple smooth and strongly-convex function with a worst-case initialization: $f_2(x) = \frac{1}{d}\sum_{i=1}^{d}\left(i \cdot (x^{(i)})^2\right)$, where $x_0^{(1)} = d, x_0^{(i)} = 0$ for $i \geq 2$; and 3) f_3 is the Rosenbrock function (f_8 in [11]) which is a well-known non-convex function used to test the performance of optimization problems. For f_1 and f_2, we set $\hat{L}$ to ground truth value L; for f_3, we search $\hat{L}$ for best performance for each algorithm.We set $d = 256$ for all test functions and set q such that each iteration of these algorithms costs 11 queries[8] to the directional derivative oracle.[9] We plot experimental results in Fig. 1, where the horizontal axis represents the number of iterations divided by $\lfloor d/11 \rfloor$, and the vertical axis represents $\log_{10} \frac{f(x_{\text{current}})-f(x^*)}{f(x_0)-f(x^*)}$. Methods without using the prior information are shown with dashed lines. We also plot the 95% confidence interval in the colored region.



(a) f_1 (b) f_2 (c) f_3

Figure 1: Experimental results using biased gradient as the prior (best viewed in color).

The results show that for these functions (which have ill-conditioned Hessians), ARS-based methods perform better than the methods based on greedy-descent. Importantly, the utilization of the prior could significantly accelerate convergence for both greedy descent and ARS. We note that the performance of our proposed PARS algorithm is better than PARS-Naive which naively replaces the gradient estimator in the original ARS with the PRGF estimator, demonstrating the value of our algorithm design with convergence analysis.

Next, we verify the properties of History-PRGF and History-PARS, i.e., the historical-prior-guided algorithms. In this part we set $d = 500$. We first verify that they are robust against learning rate on f_1 and f_2, and plot the results in Fig. 2(a)(b). In the legend, for example, 'RGF' means RGF using the optimal learning rate ($\hat{L} = L$), and 'RGF-0.02' means that the learning rate is set to 0.02 times of the optimal one ($\hat{L} = 50L$). We note that for PRGF and PARS, $q = 10$, so $\frac{q}{d} = 0.02$. From Fig. 2(a)(b), we see that: 1) when using the optimal learning rate, the performance of prior-guided algorithms is not worse than that of its corresponding baseline; and 2) the performance of prior-guided algorithms under the sub-optimal learning rate such that $\frac{q}{d} \leq \frac{L}{\hat{L}} \leq 1$ is at least comparable to that of its corresponding baseline with optimal learning rate. However, for baseline algorithms (RGF and ARS), the convergence rate significantly degrades if a smaller learning rate is set. To this end, we

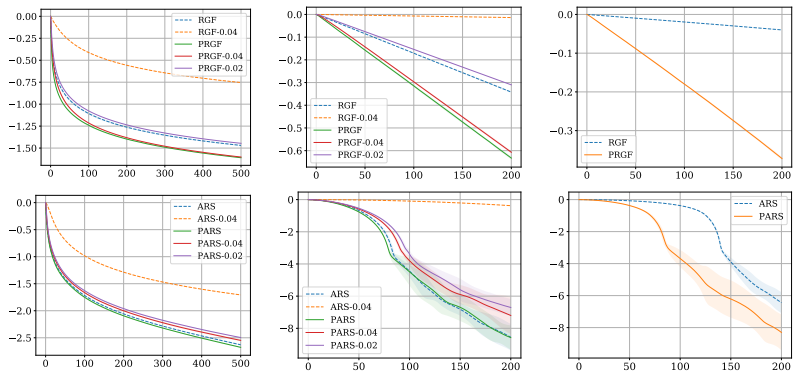
[8] In PARS, 2 additional queries of directional derivative oracle per iteration are required to find θ_t.

[9] The directional derivative is approximated by finite differences.

verify our claims that History-PRGF and History-PARS are robust to learning rate if $\frac{q}{d} \leq \frac{L}{\tilde{L}} \leq 1$. Moreover, we show that they can provide acceleration over baselines with optimal learning rate on functions with varying local smoothness. We design a new test function as follows:

$$f_4(x) = \begin{cases} \frac{1}{2}r^2, & r \leq 1 \\ r - \frac{1}{2}, & r > 1 \end{cases}, r = \sqrt{f_2(x)}, \text{ where } x_0^{(1)} = 5\sqrt{d}, x_0^{(i)} = 0 \text{ for } i \geq 2. \tag{11}$$

We note that f_4 in regions far away from the origin is more smooth than in the region near the origin, and the global smoothness parameter is determined by the worst-case situation (the region near the origin). Therefore, baseline methods using an optimal learning rate could also manifest sub-optimal performance. Fig. 2(c) shows the results. We can see that when utilizing the historical prior, the algorithm could show behaviors of adapting to the local smoothness, thus accelerating convergence when the learning rate is locally too conservative.



(a) f_1 (b) f_2 (c) f_4

Figure 2: Experimental results using the historical prior.

5.2 Black-box adversarial attacks

In this section, we perform ZO optimization on real-world problems. We conduct score-based black-box targeted adversarial attacks on 500 images from MNIST. In view of optimization, this corresponds to performing maximization over $\{f_i\}_{i=1}^{500}$ respectively, where f_i denotes the loss function to maximize in attacks. For each image i, we record the number of queries used to optimize f_i until the attack succeeds (when using C&W loss function, this means $f_i > 0$). For each optimization method, we report the median query number over these images (smaller is better).

We find that in this task, ARS-based methods do not perform better than RGF-based ones. Since it has been shown that PRGF with transfer-based prior works well [7], we only report the results of RGF and History-PRGF in Tab. 1, where the subscript of the method name indicates the learning rate. History-PRGF are more robust to learning rate than RGF. However, a small learning rate could still lead to its deteriorated performance due to non-smoothness of the objective function.

Table 1: Attack results on MNIST.

METHOD	MEDIAN QUERY
$\text{RGF}_{0.2}$	778
$\text{RGF}_{0.1}$	1597
$\text{PRGF}_{0.2}$	485
$\text{PRGF}_{0.1}$	573
$\text{PRGF}_{0.05}$	705

6 Conclusion and Discussion

In this paper, we theoretically conduct convergence analysis on existing prior-guided ZO optimization algorithms including PRGF and History-PRGF. Furthermore, we propose a novel prior-guided ARS algorithm and give its convergence guarantee. Experimental results confirm our theoretical analysis.

Our limitation lies in 1) we adopt the directional derivative oracle in our analysis, so the error on the convergence bound brought by finite-difference approximation has not been clearly stated; 2) our implementation of PARS in practice requires an approximate solution of θ_t, and the influence of this approximation is not well studied yet. We leave these as future work.

References

[1] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. *arXiv preprint arXiv:1606.04474*, 2016.

[2] Albert S Berahas, Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg. A theoretical and empirical comparison of gradient approximations in derivative-free optimization. *Foundations of Computational Mathematics*, pages 1–54, 2021.

[3] El Houcine Bergou, Eduard Gorbunov, and Peter Richtarik. Stochastic three points method for unconstrained smooth minimization. *SIAM Journal on Optimization*, 30(4):2726–2749, 2020.

[4] Thomas Brunner, Frederik Diehl, Michael Truong Le, and Alois Knoll. Guessing smart: Biased sampling for efficient black-box adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4958–4966, 2019.

[5] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

[6] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

[7] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *arXiv preprint arXiv:1906.06919*, 2019.

[8] Krzysztof Choromanski, Mark Rowland, Vikas Sindhwani, Richard Turner, and Adrian Weller. Structured evolution with compact architectures for scalable policy optimization. In *International Conference on Machine Learning*, pages 970–978. PMLR, 2018.

[9] Daniel Golovin, John Karro, Greg Kochanski, Chansoo Lee, Xingyou Song, and Qiuyi Zhang. Gradientless descent: High-dimensional zeroth-order optimization. *arXiv preprint arXiv:1911.06317*, 2019.

[10] Nikolaus Hansen and Andreas Ostermeier. Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of IEEE international conference on evolutionary computation*, pages 312–317. IEEE, 1996.

[11] Nikolaus Hansen, Steffen Finck, Raymond Ros, and Anne Auger. *Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions*. PhD thesis, INRIA, 2009.

[12] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pages 2137–2146. PMLR, 2018.

[13] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. *arXiv preprint arXiv:1807.07978*, 2018.

[14] David Kozak, Stephen Becker, Alireza Doostan, and Luis Tenorio. A stochastic subspace approach to gradient-free optimization in high dimensions. *Computational Optimization and Applications*, 79(2): 339–368, 2021.

[15] Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, and Jascha Sohl-Dickstein. Guided evolutionary strategies: Augmenting random search with surrogate gradients. In *International Conference on Machine Learning*, pages 4264–4273. PMLR, 2019.

[16] Horia Mania, Aurelia Guy, and Benjamin Recht. Simple random search of static linear policies is competitive for reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1805–1814, 2018.

[17] J Matyas. Random optimization. *Automation and Remote control*, 26(2):246–253, 1965.

[18] Florian Meier, Asier Mujika, Marcelo Matheus Gauy, and Angelika Steger. Improving gradient estimation in evolutionary strategies with past descent directions. *arXiv preprint arXiv:1910.05268*, 2019.

[19] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[20] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[21] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[22] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *arXiv preprint arXiv:1206.2944*, 2012.

[23] Sebastian U Stich, Christian L Muller, and Bernd Gartner. Optimization of convex functions with random pursuit. *SIAM Journal on Optimization*, 23(2):1284–1309, 2013.

[24] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the supplemental material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5 and the supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 5.1.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]