Stochastic Zeroth-order Optimization in High Dimensions

Yining Wang **Simon S. Du** **Sivaraman Balakrishnan** **Aarti Singh**
Carnegie Mellon University

Abstract

We consider the problem of optimizing a high-dimensional convex function using stochastic zeroth-order queries. Under sparsity assumptions on the gradients or function values, we present two algorithms: a successive component/feature selection algorithm and a noisy mirror descent algorithm using Lasso gradient estimates, and show that both algorithms have convergence rates that depend only logarithmically on the ambient dimension of the problem. Empirical results confirm our theoretical findings and show that the algorithms we design outperform classical zeroth-order optimization methods in the high-dimensional setting.

1 INTRODUCTION

We consider the problem of *stochastic zeroth order optimization*, where one wishes to compute the minimizer of a function $f : \mathcal{X} \to \mathbb{R}$ defined on a known d-dimensional domain $\mathcal{X} \subseteq \mathbb{R}^d$. In the stochastic zeroth-order optimization setting the target function f is unknown, and we obtain information about f only through *noisy* function evaluations at T adaptively chosen points $x_1, \ldots, x_T \in \mathcal{X}$. At each query point x_t we observe y_t where,

$$y_t = f(x_t) + \xi_t, \tag{1}$$

and ξ_t represents stochastic (zero-mean) noise.

The (stochastic) zeroth-order optimization problem is a classical problem in optimization, machine learning, statistics, and related fields, and is also known as *derivative-free optimization* or *black-box optimization*. Examples include applications where gradients are difficult to evaluate and/or communicate (e.g., distributed learning and parameter optimization of complicated decision processes), and where the function f is itself unknown or inaccessible such as hyperparameter tuning in machine learning and search for optimal parameters in experimental or simulation studies [36, 31, 26, 23].

The main focus of this paper is to understand the (convex) stochastic zeroth-order optimization problem in *high-dimensional* scenarios, where the dimension of the function to be optimized (d) is very large and may far exceed the sample budget T allowed. Compared to the classical stochastic *first-order* optimization setting, high dimensionality poses unique challenges in the zeroth-order query model (1). For example, if first-order information is available (exact or noisy) and the gradient of the function f is Lipschitz continuous with respect to the Euclidean distance, the iteration complexity of the classical (stochastic) gradient descent algorithm is independent of dimension d; on the other hand, the paper [18] establishes an information-theoretic lower bound for the zeroth-order optimization problem showing that, under the same Lipschitz continuous gradient assumptions, any zeroth-order optimization algorithm requires sample complexity growing polynomially with the dimension d. In addition, classical zeroth-order optimization algorithms such as the local averaging method [13, 2] have variance scaling linearly with d and are not directly feasible in the high-dimensional setting. Motivated by these observations and by real-world applications we introduce additional sparsity assumptions that enable tractable zeroth-order optimization in high dimensions.

We propose two methods for high-dimensional zeroth-order optimization: the first method uses a few noisy samples to select a small subset of "important variables" $S \subseteq [d]$; afterwards, existing low-dimensional zeroth-order optimization techniques are applied to obtain a minimizer of f restricted to S. We also propose a different method that combines stochastic mirror descent [22, 27, 12] and de-biased Lasso gradient estimates [19, 39, 43]. This stochastic mirror descent based method requires weaker assumptions and is seen

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

to outperform the variable selection based method in simulations.

1.1 Related work

We conclude this section with a discussion of related works. The zeroth-order optimization problem and its extension to bandit convex optimization have been extensively studied in the machine learning and optimization literature. The paper [13] considers a locally smoothed surrogate of f whose gradients can be unbiasedly estimated under the zeroth-order query model (1) and provides sub-linear regret bounds for the bandit convex optimization problem; the bounds in this setting were later improved by [2, 35, 16] under additional smoothness and strong convexity assumptions. Using techniques beyond gradient-based optimization, [18, 3, 8] achieved tight dependency on the sample-budget T, often obtaining worse dependency on the dimension d.

There is a rich literature on sparse (high-dimensional) optimization in the classical optimization setup, where the gradients of the objective function f can be exactly or approximately (unbiasedly) computed, such as by sampling when the objective is a finite sum [20]. *Mirror descent* [28] is the classical approach for optimization with non-standard geometry and has been applied to problems with ℓ_1, sparsity or simplex constraints [6, 1, 34, 22, 27, 14]. Alternative methods such as coordinate descent [33] and the homotopy method [42] were developed to achieve faster convergence. We remark that the first-order settings, even with noisy/stochastic gradient oracles, are very different from zeroth-order optimization because in first-order optimization the approximate gradient estimation is usually assumed to be *unbiased* with respect to the gradient of the original function, which is generally not possible in zeroth-order settings.

Bayesian optimization [36] considers the same problem of optimizing an unknown function through zeroth-order query points. Typically, in Bayesian optimization the objective function f is not assumed to be convex, and the convergence rate generally scales *exponentially* with problem dimension d [9, 32].

The papers [4, 5] consider the zeroth-order optimization problem and apply compressed sensing and sparse recovery techniques to estimate both gradients and Hessians, and incorporate these in a trust-region algorithm. While the high-level ideas are similar, our algorithms are gradient-based because recovery of Hessian matrices are not always computationally desirable, especially in the high-dimensional settings. Furthermore, no explicit finite-sample convergence rates were established in [4, 5].

Lasso and ℓ_1-penalized methods have seen great success in the fields of sparse signal recovery and high-dimensional statistical estimation [38, 10, 11]. Theoretical properties of Lasso such as ℓ_p error bounds and model selection consistency are well understood [21, 44, 7, 41, 30]. Recently, there has been growing interest in "de-biasing" the Lasso estimator in order to build component-wise confidence intervals [19, 43, 39]. We build on such de-biasing procedures in order to obtain improved rates of convergence in sparse zeroth-order optimization problems.

2 PROBLEM SETUP

In this section we introduce some notation that we use throughout the paper before formally introducing the structural assumptions we work under and the loss metric we consider.

2.1 Additional notation

We write $f(n) \lesssim g(n)$ or $f(n) = \mathcal{O}(g(n))$ if there exists a constant $C > 0$ such that $|f(n)| \leq C|g(n)|$ for all $n \in \mathbb{N}$. We use $f(n) \asymp g(n)$ if $f(n) \lesssim g(n)$ and $g(n) \lesssim f(n)$. We also use $\widetilde{\mathcal{O}}(\cdot)$ to suppress poly-logarithmic dependency on n or d.

For $1 \leq p \leq \infty$, the ℓ_p norm of a vector $x \in \mathbb{R}^d$ is defined as $\|x\|_p := (\sum_{i=1}^d |x_i|^p)^{1/p}$ for $p < \infty$, and $\|x\|_\infty := \max_{1\leq i\leq d} |x_i|$ for $p = \infty$. For two vectors $x, y \in \mathbb{R}^d$, the inner product $\langle \cdot, \cdot \rangle$ is defined as $\langle x, y\rangle := \sum_{i=1}^d x_i y_i$. A univariate random variable X is *sub-Gaussian* with parameter ν^2 if $\mathbb{E}[\exp(a(X - \mathbb{E}X))] \leq \exp\{\nu^2 a^2/2\}$ for all $a \in \mathbb{R}$. A d-dimensional random vector X is sub-Gaussian with parameter ν^2 if $\langle X - \mathbb{E}X, a\rangle$ is sub-Gaussian with parameter $\nu^2\|a\|_2^2$ for all $a \in \mathbb{R}^d$. A random variable X is *sub-exponential* with parameters (ν, α) if $\mathbb{E}[\exp(a(X-\mathbb{E}X))] \leq \exp\{\nu^2 a^2/2\}$ for all $|a| \leq 1/\alpha$.

2.2 Assumptions and evaluation measures

We make the following assumptions on the target function $f : \mathcal{X} \to \mathbb{R}$ to be optimized:

A1 (*Unconstrained convex optimization*): We take $\mathcal{X} = \mathbb{R}^d$ and assume that f is convex, i.e. for all $x, x' \in \mathcal{X}$ and $\lambda \in [0, 1]$, $f(\lambda x + (1 - \lambda)x') \leq \lambda f(x) + (1 - \lambda)f(x')$.

A2 (*Minimizer of bounded ℓ_1-norm*): We assume there exists $x^* \in \mathcal{X}$ such that $f(x^*) = f^* = \inf_{x\in\mathcal{X}} f(x)$ and $\|x^*\|_1 \leq B$; x^* does not have to be unique.

A3 (*Sparsity of gradients*): We assume that f is differentiable and that there exist $H > 0$, $s \ll d$

such that

$$\|\nabla f(x)\|_0 \le s, \quad \|\nabla f(x)\|_1 \le H, \quad \forall x \in \mathcal{X},$$

where $\|z\|_0$ and $\|z\|_1$ are the ℓ_0 and ℓ_1 vector norms; the support of $\nabla f(x)$ could potentially vary with $x \in \mathcal{X}$.

A4 (*Weak sparsity of Hessians*): We assume that f is twice differentiable and there exists $H > 0$ such that

$$\|\nabla^2 f(x)\|_1 \le H, \quad \forall x \in \mathcal{X},$$

where $\|A\|_1 := \sum_{i,j=1}^d |A_{ij}|$ is the entry-wise ℓ_1 norm of matrix A.

(A3) and (A4) are key assumptions in our paper, which assumes the gradients of f are *sparse*, and places a weaker sparsity assumption on the Hessian matrices that constrains their ℓ_1 norm rather than ℓ_0 norm.

We also note that, assuming $\|\nabla f(x)\|_\infty$ and $\|\nabla^2 f(x)\|_\infty$ are both bounded, both (A3) and (A4) are implied by the following stronger but more intuitive "function sparsity" assumption:

A5 *(Function sparsity)*: there exists $S \subseteq [d]$, $|S| \le s$ and $f_S : \mathbb{R}^{|S|} \to \mathbb{R}$ such that $f(x) \equiv f_S(x_S)$, where $x_S \in \mathbb{R}^{|S|}$ is the restriction of $x \in \mathbb{R}^d$ on S.

We motivate Assumptions (A3), (A4) and (A5) from both theoretical and practical perspectives. Theoretically, the sparsity assumption allows us to estimate the gradient at a specific point using $n \ll d$ noisy zeroth-order queries. On the other hand, (A5) is at least approximately satisfied in many practical applications of zeroth-order optimization. For example, in hyper-parameter tuning problems of learning systems, it is usually the case that the performance of the system is insensitive to some hyper-parameters, essentially implying the sparsity of the gradients and Hessians. Other examples include the optimization of visual stimuli so that certain types of neural responses are maximized or optimizing experimental parameters (pressure, temperature, etc.) so that the resulting synthesized material has optimal quality [31, 26]. For the visual stimuli optimization example, it is well known that the hierarchical organization of the human visual system in the brain into regions such as V1, V4, LO, IT etc. is precisely based on the neural response in these regions being sensitive to specific subsets of low-level and higher-level features such as edges and curves. This in turn implies that the underlying function to be optimized satisfies (A5). Finally, we remark that similar sparsity assumptions have been considered in past work [4, 24] to obtain improved rates of convergence for optimization methods.

Function: GradientEstimate$(x_t, n, \delta, \lambda)$.
Sample i.i.d. Rademacher $z_1, \ldots, z_n \in \{1, -1\}^d$;
Observe $\widetilde{y}_i = y_i/\delta$, where $y_i = f(x_t + \delta z_i) + \xi_i$;
Let $(\widehat{g}_t, \widehat{\mu}_t)$ be the solution to Eq. (3);
Output: the Lasso gradient estimate $\widehat{g}_t$ and $\widehat{\mu}_t$.

Algorithm 1: Lasso gradient estimate

Evaluation measures: Let T be the number of queries an algorithm $\mathcal{A}$ is allowed to make in the model (1), and denote by $x_1, \ldots, x_T \in \mathcal{X}$ the points at which $\mathcal{A}$ makes queries, before producing a final estimate x_{T+1}. The performance of an optimization algorithm $\mathcal{A}$ can be measured in two ways:

- simple regret $R^{\mathsf{S}}_{\mathcal{A}}(T) := f(x_{T+1}) - f^*$;

- cumulative regret $R^{\mathsf{C}}_{\mathcal{A}}(T) := \frac{1}{T}\sum_{t=1}^T f(x_t) - f^*$.

The simple regret $R^{\mathsf{S}}_{\mathcal{A}}(T)$ coincides with the classical definition of optimization error and depends only on x_{T+1}, while the cumulative regret $R^{\mathsf{C}}_{\mathcal{A}}(T)$ (used extensively in online learning problems) is also affected by the quality of intermediate query points $\{x_t\}_{t=1}^T$. Note that both $R^{\mathsf{S}}_{\mathcal{A}}(T)$ and $R^{\mathsf{C}}_{\mathcal{A}}(T)$ are random variables, with randomness in measurement error $\{\xi_t\}_{t=1}^T$ and the intrinsic randomness in $\mathcal{A}$. Finally, we remark that the simple regret can always be upper bounded by the cumulative regret for convex problems, since for any algorithm $\mathcal{A}$ that has small $R^{\mathsf{C}}_{\mathcal{A}}(T)$, taking $x_{T+1} = \frac{1}{T}\sum_{t=1}^T x_t$ achieves a simple regret $R^{\mathsf{S}}_{\mathcal{A}}(T) \le R^{\mathsf{C}}_{\mathcal{A}}(T)$.

3 LASSO GRADIENT ESTIMATION

In this section we introduce the Lasso gradient estimator, which plays a central role in both our algorithms. More specifically, for any $x_t \in \mathcal{X}$, the Lasso gradient estimator uses $n \ll d$ samples to estimate the unknown gradient $g_t := \nabla f(x_t)$. The high-level idea is to consider $n \ll d$ random samples near the point x_t, and to then formulate the gradient estimation problem as a biased linear regression system. The Lasso procedure (and its variants) can then be applied to obtain a consistent estimator under certain sparsity assumptions on $\{g_t\}_{t=1}^T$.

Fix an arbitrary $x_t \in \mathcal{X}$ and let $z_1, \ldots, z_n \in \{\pm 1\}^d$ be n samples of i.i.d. binary random vectors such that $\Pr[z_{ij} = 1] = \Pr[z_{ij} = -1] = 1/2$, where $i \in [n]$ and $j \in [d]$. Let $\delta > 0$ be a probing parameter which will be specified later, and $y_1 = f(x_t + \delta z_1) + \xi_1, \ldots, y_n = f(x_t + \delta z_n) + \xi_n$ be the n observations (1) under random perturbations (scaled by δ) $z_1, \ldots, z_n$ of x_t. Using

Input: sample budget T, parameters η, δ, λ, sparsity level s, minimizer norm upper bound B
Initialization: $x_0 = 0$, $T' = \lfloor T/2s \rfloor$; $\widehat{S}_0 = \emptyset$, $\widehat{S}_{-1} \neq \emptyset$, $t = 0$; $\widetilde{\mathcal{X}} = \{x \in \mathcal{X} : \|x\|_1 \leq B\}$;
while $|\widehat{S}_t| < s$ *and* $t < s$ *and* $\widehat{S}_t \neq \widehat{S}_{t-1}$ **do**
  $t \leftarrow t + 1$;
  Gradient estimation: $\widehat{g}_t \leftarrow \textsc{GradientEstimate}(x_{t-1}, T', \delta, \lambda)$;
  Thresholding: $\widehat{S}_t \leftarrow \widehat{S}_{t-1} \cup \{i \in [d] : |[\widehat{g}_t]_i| \geq \eta\}$;
  Run finite-difference algorithm from [13] on $f_{\widehat{S}_t}$ with T' queries, feasible region $\widetilde{\mathcal{X}}$ and starting point x_{t-1}; suppose the output is x_t;
end
Output: $x_{T+1} = x_t$ if $|\widehat{S}_t| = s$ and x_{t-1} otherwise.

Algorithm 2: The successive component selection algorithm

first-order Taylor expansions with Lagrangian remainders, the normalized $\widetilde{y}_i := y_i/\delta$ can be written as

$$\begin{aligned}\widetilde{y}_i &= \frac{f(x_t + \delta z_i) + \xi_i}{\delta} \\ &= \delta^{-1} f(x_t) + g_t^\top z_i + \frac{\delta}{2} z_i^\top H_t(\kappa_i, z_i) z_i + \delta^{-1}\xi_i \\ &:= \mu_t + g_t^\top z_i + \varepsilon_i, \end{aligned} \tag{2}$$

where $\mu_t = \delta^{-1} f(x_t)$, $\varepsilon_i = \frac{\delta}{2} z_i^\top H_t(\kappa_i, z_i) z_i + \delta^{-1}\xi_i$ and $H_t(\kappa_i, z_i) = \nabla^2 f(x_t + \kappa_i \delta z_i)$ for some $\kappa_i \in (0, 1)$.

Eq. (2) shows that, essentially, the question of estimating $g_t = \nabla f(x_t)$ can be cast as a linear regression model with design $\{z_i\}_{i=1}^n$, unknown parameters $(\mu_t, g_t) \in \mathbb{R}^{d+1}$ and noise variables $\{\varepsilon_i\}_{i=1}^n$ whose bias (i.e., $\mathbb{E}[\varepsilon_i | z_i, x_t]$) goes to 0 as $\delta \to 0$, at the expense of increasing variance. Since g_t is a sparse vector as a consequence of (A3), one can use the Lasso [38] to obtain an estimate of g_t and μ_t:

$$(\widehat{g}_t, \widehat{\mu}_t) = \underset{g \in \mathbb{R}^d, \mu \in \mathbb{R}}{\arg\min} \frac{1}{n} \sum_{i=1}^n (\widetilde{y}_i - g^\top z_i - \mu)^2 + \lambda \|g\|_1 + \lambda |\mu|, \tag{3}$$

where $\lambda > 0$ is a regularization parameter that will be specified later. A pseudocode description of the Lasso gradient estimator is given in Algorithm 1. The following lemma shows that with a carefully chosen λ, $\widehat{g}_t$ is a good estimate of g_t in both ℓ_∞ and ℓ_1 norms.

Lemma 1. *Suppose (A1) through (A4) hold. Suppose also that* $n = \Omega(s^2 \log d)$, $n \leq d$ *and* $\lambda \asymp \delta^{-1}\sigma\sqrt{\log d/n} + \delta H$. *Then with probability* $1 - \mathcal{O}(d^{-2})$

$$\max\{|\widehat{\mu}_t - \mu_t|, \|\widehat{g}_t - g_t\|_\infty\} \lesssim \frac{\sigma}{\delta}\sqrt{\frac{\log d}{n}} + \delta H.$$

Furthermore, with probability $1 - \mathcal{O}(d^{-2})$ *it holds that* $\|\widehat{g}_t - g_t\|_1 \leq 2s\|\widehat{g}_t - g_t\|_\infty$.

Lemma 1 follows by the standard ℓ_1 and ℓ_∞ error bound analyses of the Lasso estimator [7, 25]. However, our model has a subtle difference from the standard high-dimensional regression model in that $\mathbb{E}[\varepsilon_i | z_i, x_t]$ are not exactly zero. and we provide a detailed proof in the Appendix.

Remark 1. The penalization of μ in Eq. (3) is in general unnecessary as it is a single component; however, we decide to keep this penalization term to simplify our analysis. Neither the estimation error nor the selection of the tuning parameter λ depend on knowledge of μ_t.

Remark 2. Lemma 1 reveals an interesting bias-variance tradeoff controlled by the "probing" parameter $\delta > 0$. When δ is close to 0, the bias (reflected by $\mathbb{E}[\varepsilon_i | z_i, x_t]$) resulting from the second-order Lagrangian remainder term $\frac{\delta}{2} z_i^\top H_t(\kappa_i, z_i) z_i$ is small; however, the variance of $\widehat{g}_t$ is large because the variance of the "stochastic" noise term ξ_i/δ increases as $\delta \to 0$; on the other hand, for large δ the stochastic variance is reduced but the bias from first-order approximation of $f(x_t)$ increases.

4 COMPONENT SELECTION

Given the estimation error bound of the Lasso gradient estimator and the stronger "function sparsity" assumption (A5), our first attempt is to use $\widehat{g}_t$ to select a few "relevant" components $\widehat{S} \subseteq [d]$, $|\widehat{S}| \ll d$ and perform classical low-dimensional zeroth-order optimization restricted to $\widehat{S}$. The following corollary shows that, the components in S whose gradients have large absolute values can be detected by a thresholding Lasso estimator:

Corollary 1. *Suppose the conditions in Lemma 1 hold and let* $\eta = \omega\lambda$ *depending on some sufficiently large constant* $\omega > 1$. *Let* $\widehat{S}(\eta) := \{i \in [d] : |[\widehat{g}_t]_i| > \eta\}$ *be the selected components by thresholding the Lasso estimate* $\widehat{g}_t$. *Then with probability* $1 - \mathcal{O}(d^{-2})$

$$\{i \in S : |[\nabla f(x_t)]_i| > 2\eta\} \subseteq \widehat{S}(\eta) \subseteq S.$$

Corollary 1 can be proved by directly applying the $\|\widehat{g}_t - g_t\|_\infty$ bound in Lemma 1. It shows that with threshold $\eta = \omega\lambda$ depending on some sufficiently large constant $\omega > 1$, the thresholding estimator $\widehat{S}(\eta)$ with high probability will not include components that do not belong to S (i.e., no false positives). On the other hand, all components in S that have a sufficiently large partial derivative (at x_t) will be detected by $\widehat{S}(\eta)$.

Algorithm 2 describes the pseudo-code of a "successive" component selection algorithm inspired by the above observations. The following theorem provides a convergence analysis for Algorithm 2:

Theorem 1. *Suppose (A1) through (A5) hold. Suppose also that $T = \Omega(s^3 \log d)$ and $T \le d$. Let parameters δ, λ, η be set as $\delta \asymp \left(\frac{\sigma^2 s \log d}{H^2 T}\right)^{1/4}$, $\lambda \asymp \frac{\sigma}{\delta}\sqrt{\frac{s \log d}{T}} + \delta H$ and $\eta = \omega\lambda$ depending on some sufficiently large constant $\omega > 1$. Then with probability at least 0.9*

$$R^{\mathsf{S}}_{\mathcal{A}}(T) \lesssim B\left(\frac{\sigma^2 H^2 s \log d}{T}\right)^{1/4} + \widetilde{\mathcal{O}}(T^{-1/3}), \qquad (4)$$

The proof of Theorem 1 is essentially a repeated application of Corollary 1, which we defer to the appendix.

Remark 3. In the $\widetilde{\mathcal{O}}(\cdot)$ notation in Eq. (4) we suppress polynomial dependency on σ, s, H, B and $\log d$. The $\lesssim$ notation does not suppress dependency on any problem dependent constants.

Remark 4. The choices of λ and δ differ by factors depending on s from the choices suggested by Lemma 1. This is due to the fact that we divide the sample budget over s rounds of component selection.

Remark 5. Theorem 1 only upper bounds the simple regret of the successive component selection algorithm $\mathcal{A}$. However, it is clear that Algorithm 2 cannot achieve consistent cumulative regret bounds, because the gradient estimation step already consumes a constant fraction of sample points (up to $\mathcal{O}(s)$ factors).

Remark 6. The failure probability of Theorem 1 is at a constant level and does not go to 0 as d or T go to infinity. This is a consequence of the fact that the $T^{-1/3}$ regret bound of the paper [13] for low-dimensional zeroth-order optimization only holds in expectation. To the best of our knowledge, exponential tail bounds remain an open question [35].

5 MIRROR DESCENT

Another possibility of applying the Lasso gradient estimator $\widehat{g}_t$ for optimizing f is to consider classical or sparse first-order methods (e.g., SGD or mirror descent), with the true gradients $g_t = \nabla f(x_t)$ at each iteration replaced by their estimates $\widehat{g}_t$. However, directly plugging in the Lasso estimator leads to poor convergence properties due to the inherent estimation bias in $\widehat{g}_t$. To overcome such difficulties, we consider the recent work on *de-biased* Lasso estimators [19, 39, 43] and apply stochastic mirror descent [28] to handle the entrywise error introduced by the de-biasing estimators.

5.1 De-biased Lasso estimation

The de-biased Lasso estimator was introduced in [43] and generalized in [19, 39] to reduce bias of the Lasso estimator for the purpose of constructing confidence intervals for low-dimensional model components. In our application, the bias-reduced gradient estimate allows stochastic noise to concentrate across epochs and leads to improved convergence rates.

Let $\widetilde{Y}_t = (\widetilde{y}_1, \ldots, \widetilde{y}_n) \in \mathbb{R}^n$ and $Z_t = (z_1, \ldots, z_n) \in \mathbb{R}^{n \times d}$ be the vector forms of $\{\widetilde{y}_i\}_{i=1}^n$ and $\{z_i\}_{i=1}^n$. Since the design points z_i are i.i.d. Rademacher variables, the de-biased gradient estimator $\widetilde{g}_t$ takes a particularly simple form:

The de-biased Lasso :

$$\widetilde{g}_t := \widehat{g}_t + \frac{1}{n} Z_t^\top (\widetilde{Y}_t - Z_t \widehat{g}_t - \widehat{\mu}_t \cdot 1_n). \qquad (5)$$

Here $(\widehat{g}_t, \widehat{\mu}_t)$ is the Lasso estimator defined in Eq. (3) and $1_n = (1, \ldots, 1) \in \mathbb{R}^n$ is the n-dimensional vector of all ones.

Lemma 2. *Suppose $n = \Omega(s^2 \log d)$. With probability $1 - \mathcal{O}(d^{-2})$ it holds that*

$$\widetilde{g}_t = g_t + \zeta_t + \gamma_t;$$

where ζ_t is a d-dimensional random vector such that, for any $a \in \mathbb{R}^d$, $\langle \zeta_t, a \rangle$ conditioned on x_t is a centered sub-exponential random variable with parameters $\nu = \sqrt{n/2} \cdot \alpha$ and $\alpha \lesssim \sigma \|a\|_2 / \delta n$; and γ_t is a d-dimensional vector that satisfies

$$\|\gamma_t\|_\infty \lesssim H\delta + \frac{\sigma s \log d}{\delta n} \qquad \text{almost surely.}$$

Comparing Lemma 2 with the error bound obtained for the Lasso estimator $\widehat{g}_t$ in Lemma 1, it is clear that the entry-wise bias (i.e., $\|\gamma_t\|_\infty$) is reduced from $\mathcal{O}(\delta H + \sqrt{\log d / \delta n})$ to $\mathcal{O}(\delta H + s \log d / \delta n)$. Such de-biasing is at the cost of inflated stochastic error ζ_t, which means that unlike $\widehat{g}_t$, $\widetilde{g}_t$ is not a good estimator of g_t in the ℓ_1 or ℓ_2 norm.

Input: minimizer norm B, sample budget T, gradient estimate budget n, potential ψ, parameters η, δ, λ.
Initialization: $x_0 = 0$, $T' := \lfloor T/2n \rfloor$, $\widetilde{\mathcal{X}} := \{x : \|x\|_1 \le B\}$;
for $t = 0, \dots, T' - 1$ **do**
  Lasso gradient estimation: $(\widehat{g}_t, \widehat{\mu}_t) \leftarrow \textsc{GradientEstimate}(x_t, 2n, \delta, \lambda)$;
  De-biasing: $\widetilde{g}_t \leftarrow \widehat{g}_t + \frac{1}{n} Z_t^\top (\widetilde{Y}_t - Z_t \widehat{g}_t - \widehat{\mu}_t \cdot 1_n)$;
  MD update: $x_{t+1} \leftarrow \arg\min_{x \in \widetilde{\mathcal{X}}} \{\eta \widetilde{g}_t^\top (x - x_t) + \Delta_\psi(x, x_t)\}$;
end

Algorithm 3: First-order mirror descent with estimated gradients

5.2 Bregman divergence and stochastic mirror descent

Mirror descent (MD) [28] is a classical method in optimization when smoothness and the domain geometry are measured in (possibly) non-Euclidean metrics. The MD algorithm was applied to stochastic optimization with noisy first-order oracles in the papers [27, 1] and was also studied in the work [22] for strongly smooth composite functions with accelerated rates, and in the works [14, 15] for strongly convex composite functions.

Let $\psi : \mathcal{X} \to \mathbb{R}$ be a continuously differentiable, strictly convex function. The *Bregman divergence* $\Delta_\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined as

$$\Delta_\psi(x, y) := \psi(y) - \psi(x) - \langle \nabla\psi(x), y - x \rangle. \tag{6}$$

Let $\|\cdot\|_\psi$ be a norm and $\|\cdot\|_{\psi^*}$ be its dual norm, defined as $\|z\|_{\psi^*} := \sup\{z^\top x : \|x\|_\psi \le 1\}$. One important class of Bregman divergences is those that are κ-*strongly convex* with respect to the chosen norm, i.e. they satisfy $\Delta_\psi(x, y) \ge \frac{\kappa}{2}\|x - y\|_\psi^2$. Many choices of ψ lead to a strongly convex Bregman divergence. In this paper we consider the ℓ_a norm as choice of ψ: $\psi_a(x) := \frac{1}{2(a-1)}\|x\|_a^2$ for $1 < a \le 2$. It was proved in [1, 37] that ψ_a leads to a valid Bregman divergence that satisfies 1-strong convexity with respect to $\|\cdot\|_a$. With this setup, the MD method iteratively computes

$$x_{t+1} := \arg\min_{x \in \widetilde{\mathcal{X}}} \left\{ \eta_t \nabla f(x_t)^\top (x - x_t) + \Delta_\psi(x, x_t) \right\},$$

where $\{\eta_t\}_{t=1}^T$ is a sequence of step sizes and $\widetilde{\mathcal{X}} \subseteq \mathcal{X}$ is a subset of the domain $\mathcal{X}$ of f. In our problem where $\nabla f(x_t)$ is not accessible, the de-biased Lasso gradient estimate $\widetilde{g}_t$ is used to replace the exact gradient $\nabla f(x_t)$. A pseudo-code description of our method is given in Algorithm 3.

5.3 Rates of convergence

We present the following convergence rate for Algorithm 3, which is proved in the appendix:

Theorem 2. *Suppose (A1) through (A4) hold. Suppose also that* $T = \Omega(s^3 \log^2 d + s(1+H)^2(1 + B^4 H^4 \log^2 d))$, $T \le d$ *and that we choose the parameters* $n := \left\lfloor (1+H)\sqrt{sT} \right\rfloor$, $\eta := B\sqrt{\frac{n \log d}{T}}$, *and* $\delta := \sqrt{s \log d / n}$. *Then with probability* $1 - \mathcal{O}(d^{-1})$

$$R_{\mathcal{A}}^{\mathsf{C}}(T) \lesssim \xi_{\sigma,s} B \sqrt{\log d} \left[\frac{(1+H)^2 s}{T} \right]^{1/4} + \widetilde{\mathcal{O}}(T^{-1/2}).$$

where $\xi_{\sigma,s} = 1 + \sigma + \sigma^2/s$.

Theorem 2 shows that Algorithm 3 has similar convergence rate as the successive component selection algorithm, but operates under weaker conditions (i.e., without the function sparsity assumption (A5)). There are also two additional differences between results in Theorems 1 and 2. First, Theorem 2 upper bounds the cumulative regret $R_{\mathcal{A}}^{\mathsf{C}}(T)$, while the error bound in Theorem 1 only applies to the simple regret $R_{\mathcal{A}}^{\mathsf{S}}(T)$. Furthermore, the error bound in Theorem 2 holds with high-probability $(1 - \mathcal{O}(d^{-1}))$, while the results in Theorem 1 only hold with constant probability.

5.4 Improved rates with Hessian smoothness

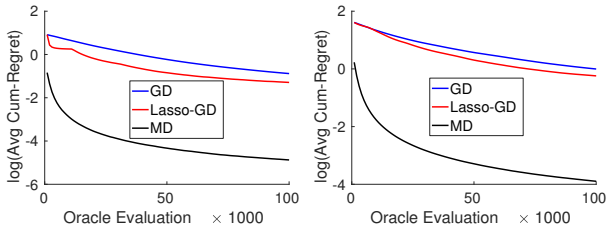
We show an extension of our algorithm that greatly improves the convergence rate under additional smoothness conditions on $\nabla^2 f$, with a small loss in computational efficiency. Formally, we assume:

A6 (*Hessian smoothness*). There exists $L > 0$ such that for all $x, x' \in \mathcal{X}$,

$$\|\nabla^2 f(x) - \nabla^2 f(x')\|_1 \le L \|x - x'\|_\infty$$

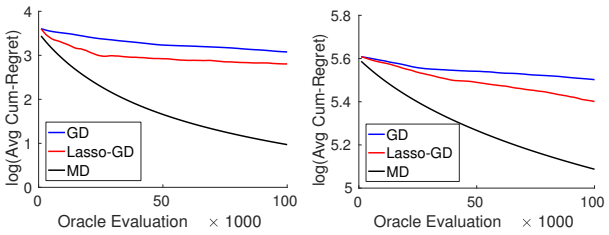
Recall that $\|A\|_1 = \sum_{i,j} |A_{ij}|$ denotes the entry-wise ℓ_1 norm of a matrix A.

If f is three-times differentiable, then (A6) is implied by the condition that $\|\nabla^3 f(x)\|_1 \le L$ for all $x \in \mathcal{X}$, where $\|A\|_1 := \sum_{i,j,k} |A_{ijk}|$ is the entry-wise ℓ_1 norm of a third order tensor. However, (A6) in general does not require third-order differentiability of f.



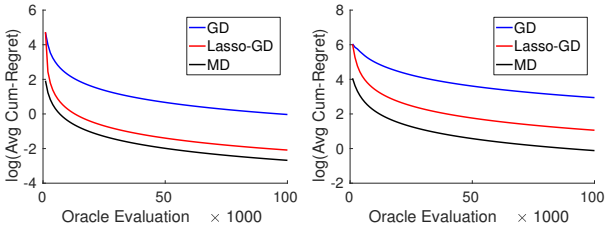
(a) $s = 10, d = 100$ (b) $s = 20, d = 100$

Figure 1: Sparse quadratic optimization with identity quadratic term.



(a) decay rate = 1.5 (b) decay rate = 3

Figure 2: Sparse quadratic optimization with polynomial decay of eigenvalues.



(a) $s = 10, d = 100$ (b) $s = 20, d = 100$

Figure 3: Sparse fourth-degree polynomial optimization with identity quadratic term.

Recall the de-biased Lasso gradient estimator $\widetilde{g}_t(\delta)$ in Eqs. (3,5) corresponding to a probing step size of δ. Under the additional condition (A6), the analysis in Lemma 2 can be strengthened as below:

Lemma 3. *Suppose (A1) through (A4) and (A6) hold. Suppose also that $n = \Omega(s^2 \log d)$, $n \le d$ and $\lambda \asymp \delta^{-1}\sigma\sqrt{\log d/n} + \delta H$. Then with probability $1 - \mathcal{O}(d^{-2})$*

$$\widetilde{g}_t(\delta) = g_t + \frac{\delta}{2}\mathbb{E}\left[(z^\top H_t z)z\right] + \widetilde{\zeta}_t(\delta) + \widetilde{\beta}_t(\delta) + \widetilde{\gamma}_t(\delta),$$

where $g_t = \nabla f(x_t)$, $H_t = \nabla^2 f(x_t)$; for any $a \in \mathbb{R}^d$, $\langle \widetilde{\zeta}_t(\delta), a\rangle$ conditioned on x_t is a centered d-dimensional sub-exponential random variable with parameters $\nu^2 = \sqrt{n/2}\cdot\alpha$ and $\alpha \lesssim \sigma\|a\|_2/\delta n$; $\langle \widetilde{\beta}_t(\delta), a\rangle$ conditioned on x_t is a centered d-dimensional sub-Gaussian random variable with parameter $\nu \lesssim \delta H\|a\|_1/\sqrt{n}$; $\gamma_t(\delta)$ is a d-dimensional vector that satisfies

$$\|\widetilde{\gamma}_t(\delta)\|_\infty \lesssim L\delta^2 + \frac{\sigma s \log d}{n\delta} + s\delta H\sqrt{\frac{\log d}{n}}.$$

Note that $\widetilde{\zeta}_t(\delta)$ and $\widetilde{\beta}_t(\delta)$ might be correlated conditioned on x_t. Comparing Lemma 3 with Lemma 2, we observe that the bias term $\widetilde{\gamma}_t(\delta)$ is significantly smaller ($\mathcal{O}(\delta^2)$ instead of $\mathcal{O}(\delta)$); while the second term $\frac{\delta}{2}\mathbb{E}[(z^\top H_t z)z]$ is still a bias term with non-zero mean, it only depends on δ and can be easily removed. This motivates the following definition of a "twice de-biased" gradient estimator:

The twice de-biased estimator:

$$\widetilde{g}_t^{\mathsf{tw}} := 2\widetilde{g}_t(\delta/2) - \widetilde{g}_t(\delta). \tag{7}$$

Corollary 2. *Suppose the conditions in Lemma 3 are satisfied. Then with probability $1 - \mathcal{O}(d^{-2})$,*

$$\widetilde{g}_t^{\mathsf{tw}} - g_t = \widetilde{\zeta}_t + \widetilde{\beta}_t + \widetilde{\gamma}_t,$$

where $\widetilde{\zeta}_t = 2\widetilde{\zeta}_t(\delta/2) - \widetilde{\zeta}_t(\delta)$, $\widetilde{\beta}_t = 2\widetilde{\beta}_t(\delta/2) - \widetilde{\beta}_t(\delta)$ and $\widetilde{\gamma}_t = \widetilde{\gamma}_t(\delta/2) - \widetilde{\gamma}_t(\delta)$.

The twice de-biased estimator is, in principle, similar to the "twicing" trick in nonparametric kernel smoothing [29] that reduces estimation bias. In particular, Corollary 2 shows that the $\frac{\delta}{2}\mathbb{E}[(z^\top H_t z)z]$ bias term is cancelled by the "twicing" trick, and the remaining bias term $\widetilde{\gamma}$ is an order of magnitude smaller than γ in the bias term before twicing (e.g., Lemma 2). We also remark that the twice de-biased estimator $\widetilde{g}_t^{\mathsf{tw}}$ does *not* significantly increase the computational burden, because the method remains first-order and only (two copies of) the de-biased gradient estimate needs to be computed.

Plugging the "twice" de-biased gradient estimator $\widetilde{g}_t^{\mathsf{tw}}$ into the stochastic mirror descent procedure (Algorithm 3) and choosing tuning parameters n, λ, δ and η appropriately, we obtain the following improved convergence rate:

Theorem 3. *Suppose (A1) through (A4) and (A6) hold. Suppose also that $T = \Omega(s^3 \log^2 d + (1+L)^2 s^2 + H^2 B^2 (1+L) s \log d)$ and $T \le d$. Let $\eta := Bn^{2/3}\sqrt{\frac{\log d}{T}}$, $n := \lfloor (1+L)s^{2/3}\sqrt{T}\rfloor$ and $\delta := (s\log d/n)^{1/3}$. Then the simple regret $R^{\mathsf{S}}_{\mathcal{A}}(T)$ can be upper bounded with probability $1 - \mathcal{O}(d^{-1})$ as*

$$R^{\mathsf{S}}_{\mathcal{A}}(T) \lesssim \widetilde{\xi}_{\sigma,s} B\sqrt{\log d}\left(\frac{(1+L)s^{2/3}}{T}\right)^{1/3} + \widetilde{\mathcal{O}}(T^{-5/12}),$$

where $\widetilde{\xi}_{\sigma,s} = (1 + \sigma + \sigma^2/s^{2/3})$.

As a simple illustration consider the following example:

Example 1. Consider a quadratic function $f(x) = \frac{1}{2}(x - x^*)^\top Q(x - x^*)$ with (unknown) $Q \succeq 0$ being positive semi-definite and supported on $S \subseteq [d]$ with $|S| \leq s$, meaning that $Q_{ij} = 0$ if $i \notin S$ or $j \notin S$. It is easy to verify that f satisfies (A1) through (A5), and also (A6) with $L = 0$ because $\nabla^2 f(x) \equiv Q$, independent of x. Subsequently, applying results in Theorem 3 we obtain a convergence rate of $\mathcal{O}(T^{-1/3})$ for the simple regret $R^{\mathsf{S}}_{\mathcal{A}}(T)$.

More broadly, compared to Theorem 2, the stochastic mirror descent algorithm with the twice de-biased gradient estimator ($\widehat{g}^{\mathsf{tw}}_t$) has the convergence rate of $\mathcal{O}(T^{-1/3})$, which is a strict improvement over the $\mathcal{O}(T^{-1/4})$ rate in Theorem 2. Such improvement is at the cost of the additional assumption of Hessian smoothness (A6); however, the optimization algorithm remains almost unchanged and no second-order information is required at runtime. Finally, we remark that Theorem 3 only applies to the simple regret $R^{\mathsf{S}}_{\mathcal{A}}(T)$; we have yet to work out a similar bound for the cumulative regret for the particular choices of n and δ in Theorem 3.

6 SIMULATIONS

We compare our two proposed algorithms with the baseline method for low-dimensional zeroth-order optimization (proposed in [13]) on synthetic function examples. We use GD to represent "zeroth order" gradient descent algorithm proposed in [13], Lasso-GD to represent Algorithm 2 and MD to represent Algorithm 3. For our synthetic function examples, we first construct a convex low-dimensional function $f_S : \mathbb{R}^{|S|} \to \mathbb{R}$ on a uniformly chosen subset $S \subseteq [d]$ with size s, and then "extend" f_S to f defined on the high-dimensional domain $\mathbb{R}^d$ by $f(x) \equiv f_S(x_S)$. Functions constructed as such naturally satisfy the sparsity assumptions (A3), (A4) and (A5). In all plots we start at the 1000th iterations (oracle evaluations) of all algorithms to avoid clutter caused by the volatile burn-in phases. Thus, the starting points in the plots are slightly different for different algorithms.

In Figure 1 we consider sparse quadratic optimization problem with $f_S(x_S) = x_S^\top Q x_S + b^\top x_S$ where we set $Q_{ii} = 1$ and $b_i = 1$ for $i \in S$ and other entries to 0. In Figure 2 we consider sparse quadratic optimization problem with $f_S(x_S) = x_S^\top Q x + b^\top x_S$ where we set $Q_{ii} = i^{-\gamma}$ where γ is the eigenvalue decay rate and $b_i = 1$ for $i \in S$ and other entries to 0. In Figure 3 we consider sparse degree-4 polynomial optimization problem with $f_S(x) = |(x_S - b)^\top Q(x_S - b)|^2 + (x_S - b)^\top Q(x_S - b)$ where we set $Q_{ii} = 1$ and $b_i = 1$ for $i \in S$ and other entries to 0. All hyper-parameters are tuned by grid search. The cumulative regret $R^{\mathsf{C}}_{\mathcal{A}}(t) = \frac{1}{t}\sum_{t'=0}^{t-1} f(x_t) - f^*$ is reported for all algorithms and selected time epochs $t \leq T$.

We observe that in all our simulation settings, the vanilla gradient descent algorithm is dominated by our proposed algorithms. Our simulation results also suggest that the mirror descent algorithm is superior to the successive component selection algorithm. MD is also easier to use in practice as it has fewer parameters. Thus, we recommend mirror descent algorithm for practical use.

7 CONCLUDING REMARKS

In this paper we consider the problem of optimizing high-dimensional functions with noisy zeroth-order oracles. Two algorithms are proposed that work under sparsity assumptions on the gradients/Hessians or the functions themselves, and we provide convergence bounds that only depend logarithmically on the ambient domain dimension d.

We view our work as a first step towards rather than the resolution of this problem. In particular, in future work we hope to address the following questions:

1. Both Algorithms 2 and 3 require "strong" sparsity conditions on the input function or its gradients, meaning that they have to be *exactly* sparse. It is an important question whether near dimension-independent convergence can be achieved with only "weak" sparsity assumptions, which only assume, for instance, that the ℓ_1 norm of the function gradient and Hessian are bounded. Such results, if possible, would greatly expand the applicability of the problem, as few functions in practice are exactly sparse.

2. Our proposed algorithms have an $\mathcal{O}(T^{-1/4})$ convergence rate, and the mirror descent algorithm converges at $\mathcal{O}(T^{-1/3})$ with additional Hessian smoothness conditions. On the other hand, in low-dimensional zeroth-order optimization it is well-understood that the optimal convergence rate is $\mathcal{O}(\mathrm{poly}(d)T^{-1/2})$, and there are computationally efficient algorithms achieving such rates [8]. Thus, an interesting open question is whether, under additional strong or weak sparsity conditions, a similar convergence rate of $\mathcal{O}(\mathrm{poly}\log(d)T^{-1/2})$ can be achieved, with only poly-logarithmic dependency on the ambient dimension d.

ACKNOWLEDGEMENTS

We would like to thank Renbo Zhao, Qi Lei and Yuanzhi Li for helpful discussions. This work is supported by CMU ProSEED/BrainHub seed grant, AFRL FA8750-17-2-0212, NSF CAREER IIS1252412, and NSF DMS1713003.

References

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

[2] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the annual Conference on Learning Theory (COLT)*, 2010.

[3] A. Agarwal, D. Foster, D. Hsu, S. Kakade, and A. Rakhlin. Stochastic convex optimization with bandit feedback. *SIAM Journal on Optimization*, 23(1):213–240, 2013.

[4] A. S. Bandeira, K. Scheinberg, and L. N. Vicente. Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization. *Mathematical Programming*, 134:223–257, 2012.

[5] A. S. Bandeira, K. Scheinberg, and L. N. Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization*, 24(3):1238–1264, 2014.

[6] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[7] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[8] S. Bubeck, R. Eldan, and Y. T. Lee. Kernel-based methods for bandit convex optimization. In *Proceedings of the annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2017.

[9] A. D. Bull. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(Oct):2879–2904, 2011.

[10] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[11] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[12] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the annual Conference on Learning Theory (COLT)*, 2010.

[13] A. D. Flaxman, A. T. Kalai, and H. B. McHanan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2005.

[14] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[15] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

[16] E. Hazan and K. Levy. Bandit convex optimization: Towards tight bounds. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2014.

[17] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[18] K. G. Jamieson, R. Nowak, and B. Recht. Query complexity of derivative-free optimization. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2012.

[19] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[20] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Proceedings of Advances in neural information processing systems (NIPS)*, 2013.

[21] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378, 2000.

[22] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133:365–397, 2012.

[23] D. D. Leeds, J. A. Pyles, and M. J. Tarr. Exploration of complex visual feature spaces for object perception. *Frontiers in computational neuroscience*, 8, 2014.

[24] Q. Lei, I. E.-H. Yen, C.-y. Wu, I. S. Dhillon, and P. Ravikumar. Doubly greedy primal-dual coordinate descent for sparse empirical risk minimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2017.

[25] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

[26] N. Nakamura, J. Seepaul, J. B. Kadane, and B. Reeja-Jayan. Design for low-temperature microwave-assisted crystallization of ceramic thin films. *Applied Stochastic Models in Business and Industry*, 2017.

[27] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[28] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication, 1983.

[29] W. K. Newey, F. Hsieh, and J. M. Robins. Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72(3):947–962, 2004.

[30] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over Lq-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

[31] B. Reeja-Jayan, K. L. Harrison, K. Yang, C.-L. Wang, A. Yilmaz, and A. Manthiram. Microwave-assisted low-temperature growth of thin films in solution. *Scientific reports*, 2, 2012.

[32] J. Scarlett, I. Bogunovic, and V. Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. In *Proceedings of the annual conference on Learning Theory (COLT)*, 2017.

[33] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6):2807–2832, 2010.

[34] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l1-regularized loss minimization. *Journal of Machine Learning Research*, 12(Jun):1865–1892, 2011.

[35] O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the annual Conference on Learning Theory (COLT)*, 2013.

[36] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, 2012.

[37] N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. In *Proceedings of Advanced in Neural Information Processing Systems (NIPS)*, 2011.

[38] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 58(1):267–288, 1996.

[39] S. Van de Geer, P. Bühlmann, Y. Ritov, R. Dezeure, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[40] H. Victor. A general class of exponential inequalities for martingales and ratios. *The Annals of Probability*, 27(1):537–564, 1999.

[41] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using L1-constrained quadratic programming (Lasso). *IEEE Transactions on InformationTtheory*, 55(5):2183–2202, 2009.

[42] L. Xiao and T. Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.

[43] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[44] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.