Stochastic Zeroth-order Optimization in High Dimensions

Yining Wang *[1], Simon Du[1], Sivaraman Balakrishnan[2], and Aarti Singh[1]

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of Statistics, Carnegie Mellon University

October 31, 2017

Abstract

We consider the problem of optimizing a high-dimensional convex function using stochastic zeroth-order query oracles. Such problems arise naturally in a variety of practical applications, including optimizing experimental or simulation parameters with many variables. Under sparsity assumptions on the gradients or function values, we present a successive component/feature selection algorithm and a noisy mirror descent algorithm with Lasso gradient estimates and show that both algorithms have convergence rates depending only logarithmically on the ambient problem dimension. Empirical results verify our theoretical findings and suggest that our designed algorithms outperform classical zeroth-order optimization methods in the high-dimensional setting.

1 Introduction

We consider the prolblem of *stochastic zeroth order optimziation*, where one wishes to compute the minimizer of a function $f : \mathcal{X} \to \mathbb{R}$ defined on a known d-dimensional domain $\mathcal{X} \subseteq \mathbb{R}^d$. Unlike classical settings where the entire function f is accessible to the optimizer, in the stochastic zeroth-order optimization setting the target function f is unknown, and the optimizer gathers information about f only through *noisy* function evaluations at T adaptively chosen points $x_1, \cdots, x_T \in \mathcal{X}$:

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

where ξ_t represents stochastic noise.

The (stochastic) zeroth-order optimization problem is a classical problem in machine learning, optimization and related fields, and is also known under the names of *derivative-free optimization* and *black-box optimization*. Examples include applications where gradients are difficult to evaluate and/or communicate (e.g., distributed learning and parameter optimization of complicated desicion processes), and problems where the function f itself is unknown or inaccessible such as hyper-parameter selection in learning and search for optimal parameters in experimental or simulation studies (Snoek et al., 2012; Reeja-Jayan et al., 2012; Nakamura et al., 2017; Leeds et al., 2014).

*Correspondence to: Yining Wang (yiningwa@cs.cmu.edu)

The main focus of this paper is to understand the (convex) stochastic zeroth-order optimization problem under *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 more classical stochastic *first-order* optimization setting, the high dimensionality under the zeroth-order query model (1) poses unique challenges. 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, for the zeroth-order optimization problem Jamieson et al. (2012) establishes an information-theoretical lower bound showing that, under the same Lipschitz continuous gradient assumptions, any zeroth-order optimization algorithm requires sample complexity growing polynomially with dimension d. In addition, classical zeroth-order optimization algorithms such as the local averaging method (Flaxman et al., 2005; Agarwal et al., 2010) have variance scaling linearly with dimension d and are less likely to work under high-dimensional settings. Hence, we introduce additional sparsity assumptions motivated by real-world applications to make the problem feasible.

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 can be applied to obtain a minimizer of f restricted to S. We also propose a different method that combines stochastic mirror descent (Lan, 2012; Nemirovski et al., 2009; Duchi et al., 2010) and de-biased Lasso gradient estimates (Javanmard & Montanari, 2014; Van de Geer et al., 2014; Zhang & Zhang, 2014), which requires weaker assumptions and is shown to outperform the variable selection based methods in simulations.

1.1 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*): $\mathcal{X} = \mathbb{R}^d$ and f is convex, meaning that for all $x, x' \in \mathcal{X}$ and $\lambda \in [0, 1]$, $f(\lambda x + (1-\lambda)x') \le \lambda f(x) + (1-\lambda) f(x')$;
- A2 (*Minimizer of bounded ℓ_1 norm*): there exists $x^* \in \mathcal{X}$ such that $f(x^*) = f^* = \inf_{x \in \mathcal{X}} f(x)$ and $\|x^*\|_1 \le B$; x^* does not have to be unique;
- A3 (*Sparsity of gradients*): f is differentiable; furthermore, there exists $H > 0$ and $s \ll d$ such that
$$\|\nabla f(x)\|_0 \le s, \quad \|\nabla f(x)\|_1 \le H, \quad \forall x \in \mathcal{X},$$
where $\|z\|_0 := \sum_{i=1}^d 1[z_i \neq 0]$ and $\|z\|_1 := \sum_{i=1}^d |z_i|$ are the ℓ_0 and ℓ_1 vector norms; the support of $\nabla f(x)$ could be potentially different for different $x \in \mathcal{X}$;
- A4 (*Weak sparsity of Hessians*): f is twice differentiable; furthermore, 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 entrywise ℓ_1 norm of a matrix.

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

A5 *(Function sparsity)*: there exists $S \subseteq [d]$, $|S| \leq s$ 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 estimation of gradient at a specific point using $n \ll d$ noisy zeroth-order queries, which would be otherwise very difficult. On the other hand, (A4) 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 certain hyper-parameters, essentially implying the sparsity of the gradients and Hessians. Another example includes the optimization of visual stimuli so that certain types of neural responses are maximized, or optimizing the experimental parameters (pressure, temperature, etc.) so that the resulting synthesized material has the best quality (Reeja-Jayan et al., 2012; Nakamura et al., 2017). again, the neural responses under consideration might only be sensitive to some properties of the input stimuli. Finally, we remark that similar sparsity assumptions were considered in the literature (Bandeira et al., 2012; Lei et al., 2017) for improved convergence of optimization methods.

Evaluation measures: Let T be the number of queries an algorithm $\mathcal{A}$ is allowed to make in the model (1). Let also $x_1, \cdots, x_T, x_{T+1} \in \mathcal{X}$ be the points at which $\mathcal{A}$ makes queries. 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 the cumulative regret $R^{\mathsf{C}}_{\mathcal{A}}(T)$ (used extensively in online learning problems) concerns also 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 $\{\varepsilon_t\}_{t=1}^T$ and the intrinsic randomness in $\mathcal{A}$. Finally, we remark that the simple regret is always *easier* than the cumulative regret for convex problems, because for any algorithm $\mathcal{A}$ that has small $R^{\mathsf{C}}_{\mathcal{A}}(T)$, applying the averaging online-to-batch conversion $x_{T+1} = \frac{1}{T}\sum_{t=1}^T x_t$ achieves $R^{\mathsf{S}}_{\mathcal{A}}(T) \leq R^{\mathsf{C}}_{\mathcal{A}}(T)$.

1.2 Related work

The works of Bandeira et al. (2012, 2014) are perhaps the most relevant to our paper, which applied compressed sensing and sparse recovery techniques to estimate both gradients and Hessians and employed trust-region algorithms afterwards. While the ideas are similar, our focus is primarily gradient based only because recovery of Hessian matrices are not always computationally desirable, especially under high-dimensional settings. Furthermore, no explicit finite-sample convergence rates were given.

The zeroth-order optimization problem and its extension to bandit convex optimization have been extensively studied in the machine learning and optimization literature. Flaxman et al. (2005) considered a locally smoothed surrogate of f whose gradients can be unbiasedly estimated under the zeroth-order query model (1) and proved sub-linear regret for the bandit convex optimization

problem; its convergence rate was later improved by Agarwal et al. (2010); Shamir (2013); Hazan & Levy (2014) under additional smoothness and strong convexity assumptions. Using techniques beyond gradient based optimization, Jamieson et al. (2012); Agarwal et al. (2013); Bubeck et al. (2017) achieved tight dependency on T; however, the dependency on dimension d is usually worse.

There is a rich literature on sparse (high-dimensional) optimization under the traditional optimization context, where the objective function f is known a priori and its gradients can be exactly or approximately (unbiasedly) computed, such as by sampling when the objective is a finite sum (SGD or SVRG (Johnson & Zhang, 2013)). *Mirror descent* (Nemirovski & Yudin, 1983) is the classical approach for optimization under non-standard geometry and has been applied to problems subject to ℓ_1, sparsity or simplex constraints (Beck & Teboulle, 2003; Agarwal et al., 2012; Shalev-Shwartz & Tewari, 2011; Lan, 2012; Nemirovski et al., 2009; Ghadimi & Lan, 2012). Alternative methods such as coordinate descent (Shalev-Shwartz et al., 2010) and the homotopy method (Xiao & Zhang, 2013) were developed to achieve faster convergence. We also 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 (Snoek et al., 2012) 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 (Bull, 2011). While global optimization of non-convex high-dimensional functions are certainly of interest, such generalization is out of the scope of this paper.

Lasso and ℓ_1-penalized methods have seen great success in the fields of sparse signal recovery and high-dimensional statistical estimation (Tibshirani, 1996; Candès et al., 2006; Donoho, 2006). Theoretical properties of Lasso such as ℓ_p error bounds and model selection consistency are well understood (Knight & Fu, 2000; Zhao & Yu, 2006; Bickel et al., 2009; Wainwright, 2009; Raskutti et al., 2011). Recently, there has been growing interest in "de-biasing" or "densifying" the Lasso estimator for the purpose of building component-wise confidence intervals (Javanmard & Montanari, 2014; Zhang & Zhang, 2014; Van de Geer et al., 2014). We found that such de-biasing procedures were also useful for improving convergence rates in sparse zeroth-order optimization problems.

1.3 Additional notations

We write $f(n) \lesssim g(n)$ if there exists constant $C > 0$ such that $|f(n)| \leq C|g(n)|$ for all $n \in \mathbb{N}$. 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 with respect to the Euclidean norm 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 $|a| \leq 1/\alpha$.

Function: GRADIENTESTIMATE$(x_t, n, \delta, \lambda)$.
Sample i.i.d. Rademacher $z_1, \cdots, z_n \in \{1, -1\}^d$;
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

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: $\widehat{g}_t \leftarrow$ 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 (Flaxman et al., 2005) 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

2 Lasso gradient estimation

In this section we introduce the Lasso gradient estimator, which plays a central role in the development of 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 estimator (and its variants) can then be applied under certain sparsity structure assumptions on g_t.

Fix arbitrary $x_t \in \mathcal{X}$ and let $z_1, \cdots, z_n \in \{\pm 1\}^d$ be n samples of i.i.d. binary random vectors such that $\Pr[z_{ij} = 1] = \Pr[z_{ij} = -1] = 1/2$. Let $\delta > 0$ be a probing parameter which will be specified later, and $y_1 = f(x_t + \delta z_1) + \xi_1, \cdots, y_n = f(x_t + \delta z_n) + \xi_n$ be the n observations in Model (1) under random probings $z_1, \cdots, z_n$. Using first-order Taylor expansions with Lagrangian remainders, the normalized $\widetilde{y}_i := y_i/\delta$ can be written as

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

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 (Tibshirani, 1996) to

obtain an estimate of g_t and μ_t:

$$\textbf{The Lasso estimator:} \quad \min_{g\in\mathbb{R}^d,\mu\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(\widetilde{y}_i - g^\top z_i - \mu)^2 + \lambda\|g\|_1 + \lambda|\mu|. \tag{3}$$

Here in Eq. (3) $\lambda > 0$ is a regularization parameter that will be specified later. Let $\widehat{g}_t$ and $\widehat{\mu}_t$ be the optimizer of Eq. (3). 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* $n = \Omega(s^2 \log d)$, $n \le d$ *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* $\|\widehat{g}_t - g_t\|_1 \le 2s\|\widehat{g}_t - g_t\|_\infty$.

Lemma 1 follows by the standard ℓ_1 and ℓ_∞ error bound analysis of the Lasso estimator (Bickel et al., 2009; Lounici, 2008). 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. Hence, we decide to reproduce the proof of Lemma 1 in the appendix for completeness.

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 depends or requires 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.

A pseudocode description of the Lasso gradient estimator is given in Algorithm 1.

3 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 that 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 $\widehat{S}(\eta)$. For the remaining components (i.e., $i \in S$ with small $|[\nabla f(x_t)]_i|$), the assumption (A3) suggests that such components might not matter much in terms of optimization of f.

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 $T \leq d$. 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*

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

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

Remark 3. In the $\mathcal{O}(\cdot)$ 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. 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 5. The failure probability of Theorem 1 is at a constant level and does not go to 0 as d or T goes to infinity. This is a consequence of the $T^{-1/3}$ regret bound of (Flaxman et al., 2005) for low-dimensional zeroth-order optimization, which only holds in expectation. To the best of our knowledge, its "with high probability" version with exponential tail bounds remains an open question (Shamir, 2013).

4 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 (Javanmard & Montanari, 2014; Van de Geer et al., 2014; Zhang & Zhang, 2014) and apply stochastic mirror descent (Nemirovski & Yudin, 1983) to handle the entrywise error introduced by the de-biasing estimators.

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 $t = 0, \cdots, T' - 1$ **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

4.1 De-biased Lasso estimation

The de-biased Lasso estimator was introduced in (Zhang & Zhang, 2014) and generalized in (Javanmard & Montanari, 2014; Van de Geer et al., 2014) 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 eventually improve convergence rates.

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

$$\textbf{The de-biased Lasso}: \quad \widetilde{g}_t := \widehat{g}_t + \frac{1}{n} Z_t^\top (\widetilde{Y}_t - Z_t \widehat{g}_t - \widehat{\mu}_t \cdot 1_n). \tag{5}$$

Here $\widehat{g}_t, \widehat{\mu}_t$ is the Lasso estimator defined in Eq. (3) and $1_n = (1, \cdots, 1) \in \mathbb{R}^n$ 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*

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

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 $\alpha \lesssim \sigma \|a\|_2 / \delta n$; γ_t is a d-dimensional vector that satisfies

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

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 $\mathcal{O}(\delta H + s \log d / \delta^2 n)$. 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 ℓ_1 or ℓ_2 norm.

4.2 Bregman divergence and stochastic mirror descent

Mirror descent (MD, (Nemirovski & Yudin, 1983)) is a classical method in optimization with smoothness and domain geometry measured in non-standard metric other than the Euclidean distance. The MD algorithm was applied to stochastic optimization with noisy first-order oracles in (Nemirovski et al., 2009; Agarwal et al., 2012) and was also studied in (Lan, 2012) for strongly smooth composite functions with accelerated rates, and in (Ghadimi & Lan, 2012, 2013) 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 corresponding to ψ and $\|\cdot\|_{\psi^*}$ be its dual norm, defined as $\|z\|_{\psi^*} := \sup\{z^\top x : \|x\|_\psi \leq 1\}$. One important class of Bregman divergences is those that are *κ-strongly convex*: $\Delta_\psi(x, y) \geq \frac{\kappa}{2}\|x - y\|_\psi^2$. Many choices of ψ lead to 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 $1 < a \leq 2$. It was proved in (Agarwal et al., 2012; Srebro et al., 2011) that ψ_a leads to a valid Bregman divergence that satisfies 1-strong convexity with respect to $\|\cdot\|_a$. The MD then iteratively solves

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

where $\{\eta_t\}_{t=1}^T$ is a sequence of step sizes and $\widetilde{X} \subseteq \mathcal{X}$ 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.

4.3 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* $T = \Omega(s^3 \log^2 d + s(1+H)^2(1 + B^4 H^4 \log^2 d))$ *and* $T \leq d$. *Let* $\eta := B\sqrt{\frac{n \log d}{T}}$, $n := \left\lfloor (1+H)\sqrt{sT} \right\rfloor$ *and* $\delta := \sqrt{s \log d / n}$. *Then with probability* $1 - \mathcal{O}(d^{-1})$

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

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 much 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 holds with constant probability.

4.4 Improved rates with Hessian smoothness

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

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

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

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.

Consider again the de-biased Lasso gradient estimator $\widetilde{g}_t(\delta)$ in Eqs. (3,5) corresponding to a probing step size of δ. (The λ parameter in Eq. (3) should still be on the order of $\delta^{-1}\sigma\sqrt{\log d/n} + \delta H$.) Under the additional Hessian smooth 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 parameters $\nu \lesssim \delta H\|a\|_1/\sqrt{n}$; $\gamma_t(\delta)$ is a d-dimensional vector that satisfies

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

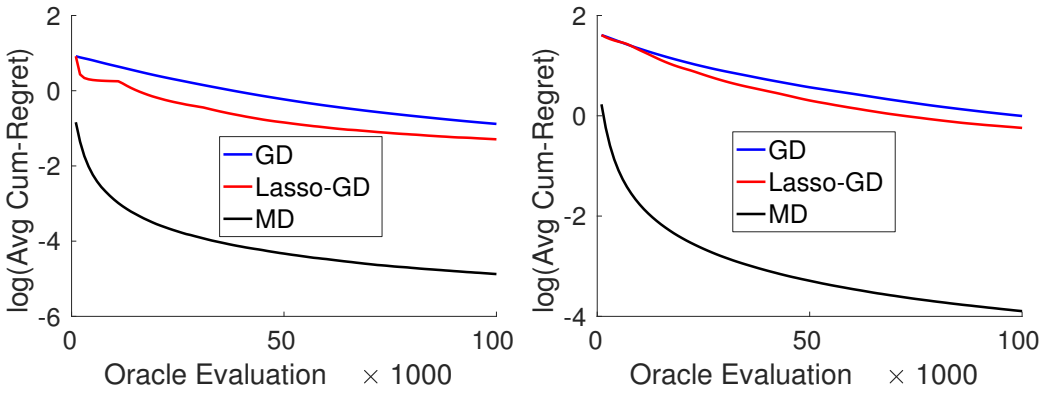
Note that $\widetilde{\zeta}_t(\delta)$ and $\widetilde{\beta}_t(\delta)$ might be correlated conditioned on x_t. Comparing Lemma 3 with Lemma 2, we observe that the bias term $\widetilde{\gamma}_t(\delta)$ is significantly smaller ($\mathcal{O}(\delta^2)$ instead of $\mathcal{O}(\delta)$); while the second term $\frac{\delta}{2}\mathbb{E}[z^\top H_t z]$ 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:

$$\textbf{The twice de-biased estimator:} \quad \widetilde{g}_t^{\mathsf{tw}} := 2\widetilde{g}_t(\delta/2) - \widetilde{g}_t(\delta). \tag{7}$$

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 $\widetilde{\gamma}_t = \widetilde{\gamma}_t(\delta/2) - \widetilde{\gamma}_t(\delta)$.



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

Figure 1: Sparse quadratic optimization with identity quadratic term.

The twice de-biased estimator is, in principle, similar to the "twicing" trick in nonparametric kernel smoothing (Newey et al., 2004) 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 estimation needs to be computed.

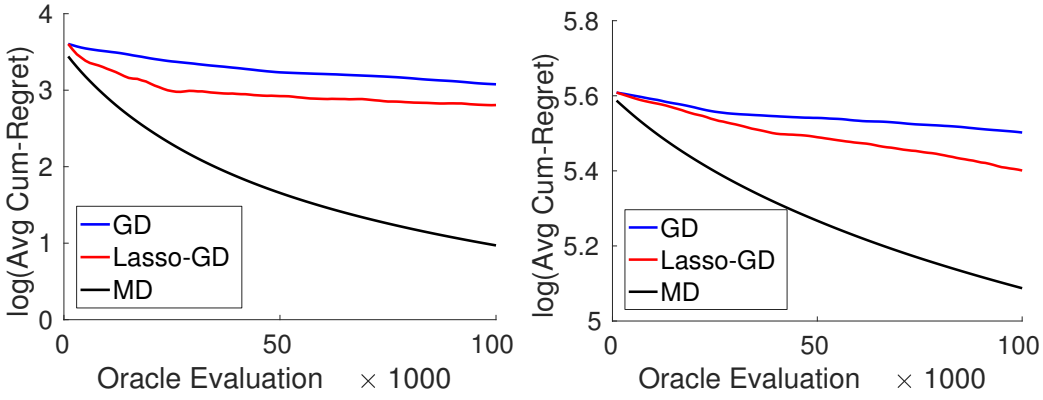
Plugging the "twice" de-biased gradient estimator $\widetilde{g}_t^{\mathsf{tw}}$ into the stochastic mirror descent procedure (Algorithm 3) and choosing tuning parameters n, λ, δ and η appropriately, we obtain the following improved convergence rate:

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* $T \leq d$. *Let* $\eta := Bn^{2/3}\sqrt{\frac{\log d}{T}}$, $n := \lfloor (1+L)s^{2/3}\sqrt{T} \rfloor$ *and* $\delta := (s \log d / n)^{1/3}$ *Let also* $x_{T+1} = \frac{1}{T'}\sum_{t=0}^{T'-1} x_t$ *be the online-to-batch averaging. Then the simple regret* $R_{\mathcal{A}}^{\mathsf{S}}(T) = f(x_{T+1}) - f^*$ *can be upper bounded with probability* $1 - \mathcal{O}(d^{-1})$ *as*

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

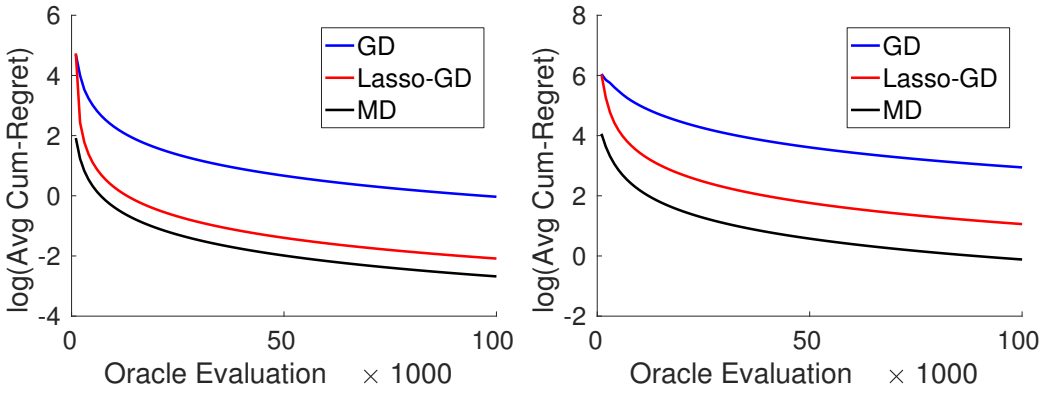
where $\widetilde{\xi}_{\sigma,s} = (1 + \sigma + \sigma^2/s^{2/3})$.

Compared to Theorem 2, the stochastic mirror descent algorithm with the twice de-biased gradient estimator ($\widetilde{g}_t^{\mathsf{tw}}$) 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 $R_{\mathcal{A}}^{\mathsf{S}}(T)$ of the averaging online-to-batch conversion $x_{T+1} = \frac{1}{T'}\sum_{t=0}^{T'-1} x_t$; we have yet to work out a similar bound for the cumulative regret $R_{\mathcal{A}}^{\mathsf{C}}(T) = \frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f^*$ for the particular choices of n and δ in Theorem 3.



(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.

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 $\nabla^2 f \equiv Q$. Subsequently, applying results in Theorem 3 we obtain a convergence rate of $\mathcal{O}(T^{-1/3})$ for the simple regret $R^{\mathsf{S}}_{\mathcal{A}}(T)$.

5 Simulations

We compare our two proposed algorithms with the baseline method for low-dimensional zeroth-order optimization (proposed in (Flaxman et al., 2005)) on synthetic function examples. We use GD to represent "zero-th order" gradient descent algorithm proposed in (Flaxman et al., 2005), 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 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 $R_{\mathcal{A}}^{\mathsf{C}}(t) = \frac{1}{t}\sum_{t'=0}^{t-1} f(x_t) - f^*$ is reported for all algorithms and selected time epochs $t \le T$.

We observe that in all our simulation settings, the vanilla gradient descent algorithm is dominated by our proposed algorithms. Therefore, if we know a priori that the function has sparsity structure, we should use specialized methods to exploit it. Our simulation results also suggest that the mirror descent algorithm is superior to the successive component selection algorithm. MD is also easy to deploy as it has fewer parameters. Thus, we recommend the mirror descent algorithm for practical usage.

6 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 convergence bounds that only depend logarithmically on the ambient domain dimension d are proved.

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 requires "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 the ℓ_1 norm of function properties to be bounded. Such results, if possible, would greatly expand the applicability of the problem, as few functions in practice are exactly sparse.

2. We show that 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 right convergence rate is $\mathcal{O}(\mathrm{poly}(d)T^{-1/2})$, and there are computationally efficient algorithms achieving such rates (Bubeck et al., 2017). 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.

7 Proofs

7.1 Proof of Lemma 1

We first prove a technical lemma that bounds the ℓ_∞ norm of error vectors.

Lemma 4. *For any $x \in \mathbb{R}^d$ and $z_i \in \{\pm 1\}^d$, with probability $1 - \mathcal{O}(d^{-3})$ (conditioned on x_t and*

z_i)

$$\left\|\sum_{i=1}^n \varepsilon_i z_i\right\|_\infty \lesssim \frac{\sigma}{\delta}\sqrt{\frac{\log d}{n}} + H\delta.$$

Proof. Let $\bar{\xi}_i = \xi_i/\delta \sim \mathcal{N}(0, \sigma^2/\delta^2)$. Consider the following decomposition:

$$\left\|\sum_{i=1}^n \varepsilon_i z_i\right\|_\infty \leq \frac{1}{n\delta}\left\|\sum_{i=1}^n \bar{\xi}_i z_i\right\|_\infty + \delta \cdot \sup_{1\leq i\leq n}\left|z_i^\top H_t(\kappa_i, z_i) z_i\right| \cdot \|z_i\|_\infty.$$

The second term on the right-hand side of the above inequality is upper bounded by $\mathcal{O}(H\delta)$ almost surely, because $\|z_i\|_\infty \leq 1$ and $|z_i^\top H_t(\kappa_i, z_i) z_i| \leq \|H_t(\kappa_i, z_i)\|_1 \|z_i\|_\infty^2 \leq H$. For the first term, because $\bar{\xi}_i$ are centered sub-Gaussian random variables independent of z_i and $\|z_i\|_\infty \leq 1$, we have that $1/n \cdot \|\sum_{i=1}^n \bar{\xi}_i z_i\|_\infty \lesssim \sqrt{\sigma^2 \log d/n}$ with probability $1 - \mathcal{O}(d^{-3})$, by invoking standard sub-Gaussian concentration inequalities. □

Now define $\widehat{\theta} = (\widehat{g}_t, \widehat{\mu}_t)$, $\theta_0 = (g_t, \delta^{-1} f(x_t))$ and $\bar{Z} = (\bar{z}_1, \cdots, \bar{z}_n)$ where $\bar{z}_i = (z_i, 1) \in \mathbb{R}^{d+1}$. Define also that $Y = (\widetilde{y}_1, \cdots, \widetilde{y}_n)$. The estimator can then be written as $\widehat{\theta} = \arg\min_{\theta\in\mathbb{R}^{d+1}} \frac{1}{n}\|\widetilde{Y} - \bar{Z}\theta\|_2^2 + \lambda\|\theta\|_1$ where $\widetilde{Y} = \bar{Z}\theta_0 + \varepsilon$, $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)$. We first establish a "basic inequality" type results that are essential in performance analysis of Lasso type estimators. By optimality of $\widehat{\theta}$, we have that

$$\frac{1}{n}\|Y - \bar{Z}\widehat{\theta}\|_2^2 + \lambda\|\widehat{\theta}\|_1 \leq \frac{1}{n}\|Y - \bar{Z}\theta_0\|_2^2 + \lambda\|\theta_0\|_1 = \frac{1}{n}\|\varepsilon\|_2^2 + \lambda\|\theta_0\|_1.$$

Re-organizing terms we obtain

$$\lambda\|\widehat{\theta}\|_1 \leq \lambda\|\theta_0\|_1 + \frac{2}{n}(\widehat{\theta} - \theta_0)^\top \bar{Z}^\top \varepsilon.$$

On the other hand, by Hölder's inequality and Lemma 4 we have, with probability $1 - \mathcal{O}(d^{-2})$,

$$\frac{2}{n}(\widehat{\theta} - \theta_0)^\top \bar{Z}^\top \varepsilon \leq 2\|\widehat{\theta} - \theta_0\|_1 \cdot \left\|\frac{1}{n}\bar{Z}^\top \varepsilon\right\|_\infty \lesssim \|\widehat{\theta} - \theta_0\|_1 \cdot \left(\frac{\sigma}{\delta}\sqrt{\frac{\log d}{n}} + H\delta\right).$$

Subsequently, if $\lambda \leq c_0(\sigma\delta^{-1}\sqrt{\log d/n} + H\delta)$ for some sufficiently small $c_0 > 0$, we have that $\|\widehat{\theta}\|_1 \leq \|\theta_0\|_1 + 1/2\|\widehat{\theta} - \theta_0\|_1$. Multiplying by 2 and adding $\|\widehat{\theta} - \theta_0\|_1$ on both sides of the inequality we obtain $\|\widehat{\theta} - \theta_0\|_1 \leq 2(\|\widehat{\theta} - \theta_0\|_1 + \|\widehat{\theta}_0\|_1 - \|\widehat{\theta}\|_1)$. Recall that θ_0 is sparse and let $\bar{S} = S \cup \{d+1\}$ be the support of θ_0. We then have $\|(\widehat{\theta} - \theta_0)_{\bar{S}^c} + \|(\theta_0)_{\bar{S}^c}\|_1 - \|\widehat{\theta}_{\bar{S}^c}\|_1 = 0$ and hence $\|(\widehat{\theta} - \theta_0)_{\bar{S}^c}\|_1 - \|(\widehat{\theta} - \theta_0)_{\bar{S}}\|_1 \leq \|\widehat{\theta} - \theta_0\|_1 \leq 2\|(\widehat{\theta} - \theta_0)_{\bar{S}}\|_1$. Thus,

$$\|(\widehat{\theta} - \theta_0)_{\bar{S}^c}\|_1 \leq 3\|(\widehat{\theta} - \theta_0)_{\bar{S}}\|_1. \tag{8}$$

Now consider $\widehat{\theta}$ that minimizes $\frac{1}{n}\|Y - \bar{Z}\theta\|_2^2 + \lambda\|\theta\|_1$. By KKT condition we have that

$$\left\|\frac{1}{n}\bar{Z}^\top(Y - \bar{Z}\widehat{\theta})\right\|_\infty \leq \frac{\lambda}{2}.$$

Define $\widehat{\Sigma} = \frac{1}{n}\bar{Z}^\top \bar{Z}$ and recall that $Y = \bar{Z}\theta_0 + \varepsilon$. Invoking Lemma 4 and the scaling of λ we have that, with probability $1 - \mathcal{O}(d^{-2})$

$$\|\widehat{\Sigma}(\widehat{\theta} - \theta_0)\|_\infty \leq \frac{\lambda}{2} + \left\|\frac{1}{n}\bar{Z}^\top \varepsilon\right\| \lesssim \frac{\sigma}{\delta}\sqrt{\frac{\log d}{n}} + \delta H. \tag{9}$$

By definition of $\{\bar{z}_i\}_{i=1}^n$, we know that $\widehat{\Sigma}_{jj} = 1$ for all $j = 1, \cdots, d+1$ and $\mathbb{E}[\widehat{\Sigma}_{jk}] = 0$ for $j \neq k$. By Hoeffding's inequality (Hoeffding, 1963) and union bound we have that with probability $1 - \mathcal{O}(d^{-2})$, $\|\widehat{\Sigma} - I_{(d+1)\times(d+1)}\|_\infty \lesssim \sqrt{\log d/n}$, where $\|\cdot\|_\infty$ denotes the maximum absolute value of matrix entries. Also note that $\widehat{\theta} - \theta_0$ satisfies $\|(\widehat{\theta} - \theta_0)_{\bar{S}^c}\|_1 \leq 3\|(\widehat{\theta} - \theta_0)_{\bar{S}}\|_1$ thanks to Eq. (8). Subsequently,

$$\begin{aligned}
\|\widehat{\theta} - \theta_0\|_\infty &\leq \|\widehat{\Sigma}(\widehat{\theta} - \theta_0)\|_\infty + \|(\widehat{\Sigma} - I)(\widehat{\theta} - \theta_0)\|_\infty \\
&\leq \|\widehat{\Sigma}(\widehat{\theta} - \theta_0)\|_\infty + \|\widehat{\Sigma} - I\|_\infty \|\widehat{\theta} - \theta_0\|_1 \\
&\leq \|\widehat{\Sigma}(\widehat{\theta} - \theta_0)\|_\infty + \|\widehat{\Sigma} - I\|_\infty \cdot 4\|(\widehat{\theta} - \theta_0)_{\bar{S}}\|_1 \\
&\leq \|\widehat{\Sigma}(\widehat{\theta} - \theta_0)\|_\infty + \|\widehat{\Sigma} - I\|_\infty \cdot 4(s+1)\|\widehat{\theta} - \theta_0\|_\infty \\
&\lesssim \frac{\sigma}{\delta}\sqrt{\frac{\log d}{n}} + \delta H + \sqrt{\frac{s^2 \log d}{n}} \cdot \|\widehat{\theta} - \theta_0\|_\infty.
\end{aligned} \tag{10}$$

Combining Eq. (10) together with the scaling $n = \Omega(s^2 \log d)$ we complete the proof of Lemma 1. Note that the statement on the ℓ_1 error $\|\widehat{\theta} - \theta_0\|_1$ is a simple consequence of the basic inequality Eq. (8).

7.2 Proof of Theorem 1

The basis of our algorithm is the analysis of the finite-difference algorithm proposed by Flaxman et al. (2005) under low dimensions. In particular, applying the analysis in (Agarwal et al., 2010) for low-dimensional strongly smooth functions, we have for every epoch $t < s$

$$\mathbb{E}[f(x_t)] - \inf_{x \in \widetilde{\mathcal{X}}, x_{\widehat{S}_t^c} = 0} f(x) \lesssim \mathrm{poly}(s, \sigma, H, \|x^*_{\widehat{S}_t}\|_1) \cdot T^{-1/3},$$

where x_t is the solution point at the tth epoch in Algorithm 2 and $\mathrm{poly}(\cdot)$ is any polynomial function of constant degrees. Recall that $\|x^*_{\widehat{S}_t}\|_1 \leq \|x^*\|_1 \leq B$ by Assumption (A2). Using Markov's inequality we have that with probability 0.9,

$$f(x_t) - \inf_{x \in \widetilde{\mathcal{X}}, x_{\widehat{S}_t^c} = 0} f(x) \lesssim \mathrm{poly}(s, \sigma, H, \|x^*_{\widehat{S}_t}\|_1) \cdot T^{-1/3}, \qquad \forall t = 0, \cdots, s. \tag{11}$$

We are now ready to prove Theorem 1. Let $\widehat{S} = \widehat{S}_t$ be the subset when Algorithm 2 terminates. In the rest of the proof we assume the conclusions in Corollary 1 and Lemma 1 hold, which happens with probability $1 - \mathcal{O}(d^{-1})$. Define $\Delta S = S \backslash \widehat{S}$, $x^* := \inf_{x \in \mathcal{X}} f(x)$ and $x_t^* = \inf_{x \in \widetilde{\mathcal{X}}, x_{\widehat{S}_t^c} = 0} f(x)$. Assumption (A5) implies that x^* can be chosen such that $x^*_{S^c} = 0$. Also, if $\Delta_S = \emptyset$ we know that $x_t^* = x^*$ and Theorem 1 automatically holds due to Eq. (11). Therefore in the rest of the proof we shall assume that $\Delta_S \neq \emptyset$.

Because $\Delta_S \neq \emptyset$ and $|S| = s$, we must have $|\widehat{S}_t| < s$. From the description of Algorithm 2, it can only happen with $\widehat{S}_t = \widehat{S}_{t-1}$. We then have that

$$\begin{aligned} f(x_{T+1}) - f(x^*) &= f(x^*_{t-1}) - f(x^*) + f(\widehat{x}_{t-1}) - f(x^*_{t-1}) \\ &\leq f(x^*_{t-1}) - f(x^*) + \mathrm{poly}(s, \sigma, H, \|x^*\|_1) \cdot T^{-1/3} \qquad (12) \\ &\leq \nabla f(x^*_{t-1})^\top (x^*_{t-1} - x^*) + \mathrm{poly}(s, \sigma, H, \|x^*\|_1) \cdot T^{-1/3}, \qquad (13) \end{aligned}$$

where Eq. (12) holds with probability at least 0.9, thanks to Eq. (11). Because x^*_{t-1} is the minimizer of f on vectors in $\widetilde{\mathcal{X}}$ that are supported on $\widehat{S} = \widehat{S}_{t-1} = \widehat{S}_t$, and that both x^*_{t-1} and x^* truncated on $\widehat{S}$ are feasible (i.e., in the restrained set $\widetilde{\mathcal{X}}$), it must hold that $\langle [\nabla f(x^*_{t-1})]_{\widehat{S}}, (x^*_{t-1} - x^*)_{\widehat{S}} \rangle \leq 0$ by first-order optimality conditions. On the other hand, by Corollary 1 and the definition of $\widehat{S}_t$, we have that $\|[\nabla f(x^*_{t-1})_{\Delta_S}]\|_\infty \leq 2\eta$. Also note that $(x^* - x^*_{t-1})_{S^c} = 0$ and $[x^*_{t-1}]_{\Delta_S} = 0$. Subsequently,

$$\nabla f(x^*_{t-1})^\top (x^*_{t-1} - x^*) \leq \left| \langle \nabla f(x^*_{t-1})_{\Delta_S}, x^*_{\Delta_S} \rangle \right| \leq \|[\nabla f(x^*_{t-1})]_{\Delta_S}\|_\infty \|x^*_{\Delta_S}\|_1 \leq 2\eta \|x^*\|_1. \quad (14)$$

Combining Eqs. (13,14) and the scalings of η, δ, λ and $T' = T/2s$ we complete the proof of Theorem 1.

7.3 Proof of Lemma 2

We use the "full-length" parameterization $\widetilde{\theta}_t = \widehat{\theta}_t + \frac{1}{n} \bar{Z}_t^\top (\widetilde{Y}_t - \bar{Z}_t \widehat{\theta}_t)$, where $\widehat{\theta}_t, \bar{Z}_t$ and $\widetilde{Y}_t$ are notations defined in the proof of Lemma 1 (with subscripts t added to emphasize that both Z_t and $\widetilde{Y}_t$ are specific to the tth epoch in Algorithm 3). Because $\widetilde{Y}_t = \bar{Z}_t \theta_{0t} + \varepsilon_t$ (where $\theta_{0t} = \nabla f(x_t)$ and $\varepsilon = (\varepsilon_{t1}, \cdots, \varepsilon_{tn})$, with ε_{ti} defined in Eq. (2)). we have

$$\widetilde{\theta}_t = \widehat{\theta}_t + \frac{1}{n} \bar{Z}_t^\top (\bar{Z}_t \theta_{0t} + \varepsilon_t - \bar{Z}_t \widehat{\theta}_t) = \theta_{0t} + \frac{1}{n} \bar{Z}_t^\top \varepsilon_t + (\widehat{\Sigma} - I_{(d+1)\times(d+1)})(\widehat{\theta}_t - \theta_{0t}),$$

where $\widehat{\Sigma} = \frac{1}{n} \bar{Z}_t^\top \bar{Z}_t$. Recall that $\varepsilon_{ti} = \xi_i / \delta + \delta z_i^\top H_t(\kappa_i, z_i) z_i$. Define $b_i = z_i^\top H_t(\kappa_i, z_i) z_i$ and $b = (b_1, \cdots, b_n)$. Also note that the first d components of $\widetilde{\theta}_t$ are identical to $\widetilde{g}_t$ defined in Eq. (5). Subsequently,

$$\widehat{g}_t = g_t + \underbrace{\frac{1}{n\delta} Z_t^\top \xi}_{:=\zeta_t} + \underbrace{\frac{\delta}{n} Z_t^\top b + \left[(\widehat{\Sigma} - I_{(d+1)\times(d+1)})(\widehat{\theta}_t - \theta_{0t}) \right]_{1:d}}_{:=\gamma_t}. \qquad (15)$$

In Eq. (15) we divide $\widehat{g}_t - g_t$ into two terms. We first consider the term $\zeta_t := \frac{1}{n\delta} Z_t^\top \xi$. It is clear that $\mathbb{E}[\zeta_t | x_t] = 0$ because $\mathbb{E}[\xi | x_t, Z_t] = 0$. Now consider any d-dimensional vector $a \in \mathbb{R}^d$, and to simplify notations all derivations below are conditioned on x_t. For any $i \in [n]$, $z_{ti}^\top a$ are i.i.d. sub-Gaussian random variables with common parameter $\nu^2 = \|a\|_2^2$. Also, $\bar{\xi}_i$ is a sub-Gaussian random variable with parameter σ^2 and is independent of $z_{ti}^\top a$. Thus, invoking Lemma 6 we have that $\xi_i z_{ti}^\top a$ is a sub-exponential random variable with parameters $\nu = \alpha/\sqrt{2} \lesssim \sigma \|a\|_2$. Consequently, $\langle \zeta_t, a \rangle = \frac{1}{n\delta} \sum_{i=1}^n \xi_i z_{ti}^\top a$ is a centered sub-exponential random variable with parameters $\nu = \sqrt{n/2} \cdot \alpha \lesssim \sigma \|a\|_2 / \delta \sqrt{n}$.

We next consider the term $\gamma_t = \frac{\delta}{n} Z_t^\top b + (\widehat{\Sigma} - I)(\widehat{\theta}_t - \theta_{0t})$. By Assumption (A3) we know that $\|b\|_\infty \leq \delta H$. Subsequently, by Hölder's inequality we have that

$$\|\gamma_t\|_\infty \leq \frac{\delta}{n} \|Z_t\|_{1,\infty} \|b\|_\infty + \|\widehat{\Sigma} - I\|_\infty \|\widehat{\theta}_t - \theta_{t0}\|_1$$

$$\lesssim H\delta + \sqrt{\frac{\log d}{n}}\left(\frac{\sigma s}{\delta}\sqrt{\frac{\log d}{n}} + s\delta H\right).$$

where the second inequality holds with probability $1 - \mathcal{O}(d^{-2})$ thanks to Lemma 1.

7.4 Proof of Theorem 2

We first note that the cumulative regret $R_{\mathcal{A}}^{\mathsf{C}}(T)$ can be upper bounded as

$$R_{\mathcal{A}}^{\mathsf{C}}(T) \lesssim \left[\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f^*\right] + \sup_t \sup_{z\in\{\pm 1\}^d} \left|f(x_t + \delta z) - f(x_t)\right|.$$

Because $\|\nabla f(x)\|_1 \le H$ for all $x \in \mathcal{X}$ and $z \in \{\pm 1\}^d$, using Hölder's inequality we have that

$$\left|f(x_t + \delta z) - f(x_t)\right| \le \delta H \lesssim B\left(\frac{s\log^2 d}{T}\right)^{1/4},$$

which is a second-order term. Thus, to prove upper bounds on $R_{\mathcal{A}}^{\mathsf{C}}(T)$ it suffices to consider only $\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f^*$.

We next cite the result in (Lan, 2012) that gives explicit cumulative regret bounds for mirror descent with approximate gradients:

Lemma 5 (Lan (2012), Lemma 3)**.** *Let $\|\cdot\|_\psi$ and $\|\cdot\|_{\psi^*}$ be a pair of conjugate norms, and let $\Delta_\psi(\cdot,\cdot)$ be a Bregman divergence that is κ-strongly convex with respect to $\|\cdot\|_\psi$. Suppose f is $\widetilde{H}$-smooth with respect to $\|\cdot\|_\psi$, meaning that $f(y) \le f(x) + \nabla f(x)^\top(y-x) + \frac{\widetilde{H}}{2}\|x-y\|_\psi^2$ for all $x, y \in \mathcal{X}$, and $\eta < \kappa/\widetilde{H}$. Define $g_t = \nabla f(x_t)$, and let $x_0, \cdots, x_{T'-1}$ be iterations in Algorithm 3. Then for every $0 \le t \le T'-1$ and any $x^* \in \widetilde{\mathcal{X}}$,*

$$\eta\left[f(x_{t+1}) - f(x^*)\right] + \Delta_\psi(x_{t+1}, x^*) \le \Delta_\psi(x_t, x^*) + \eta\langle \widetilde{g}_t - g_t, x^* - x_t\rangle + \frac{\eta^2\|\widetilde{g}_t - g_t\|_{\psi^*}^2}{2(\kappa - \widetilde{H}\eta)}. \quad (16)$$

Adding both sides of Eq. (16) from $t = 0$ to $t = T'-1$, telescoping and noting that $\Delta_\psi(x_{T'}, x^*) \ge 0$, we obtain

$$\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f(x^*) \le \frac{\Delta_\psi(x_0, x^*)}{\eta T'} + \frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t - g_t, x_t - x^*\rangle + \frac{\eta}{2(\kappa - H\eta)}\cdot \sup_{0\le t<T'}\|\widetilde{g}_t - g_t\|_{\psi^*}^2. \quad (17)$$

Set $\|\cdot\|_\psi = \|\cdot\|_a$ for $a = \frac{2\log d}{2\log d - 1}$. It is easy to verify that under Assumption (A3), the function f satisfies

$$\begin{aligned} f(y) &\ge f(x) + \nabla f(x)^\top(y-x) + H\|y-x\|_\infty^2 \\ &\ge f(x) + \nabla f(x)^\top(y-x) + \widetilde{H}\|y-x\|_\psi^2 \end{aligned}$$

for all $x, y \in \mathcal{X}$ with $\widetilde{H} \leq eH$, because $\|x-y\|_1^2 \leq d^{2(1-1/a)}\|x-y\|_a^2 \leq d^{1/\log d}\|x-y\|_1^2 = e\|x-y\|_1^2$ by Hölder's inequality. In addition, by definition of Bregman divergence we have that

$$\Delta_\psi(x_0, x^*) \leq \frac{1}{2(a-1)}\|x^*\|_a^2 \leq \frac{1}{2(a-1)}\|x^*\|_1^2 \leq \|x^*\|_1^2 \log d \leq B^2 \log d, \tag{18}$$

where the first inequality holds because $\psi_a(x_0) = \psi_a(0) = 0$ and $\nabla\psi_a(x_0) = \nabla\psi_a(0) = 0$ for $a > 1$.

We next upper bound the $\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t - g_t, x^* - x_t\rangle$ and $\|\widetilde{g}_t - g_t\|_{\psi^*}^2$ terms. By Lemma 2 and sub-exponential concentration inequalities (e.g., Lemma 7), we have that with probability $1 - \mathcal{O}(d^{-1})$

$$\|\widetilde{g}_t - g_t\|_\infty \leq \|\zeta_t\|_\infty + \|\gamma_t\|_\infty \lesssim \frac{\sigma}{\delta}\left(\sqrt{\frac{\log d}{n}} + \frac{\log d}{n}\right) + H\delta + \frac{\sigma s \log d}{\delta n} \lesssim \frac{\sigma}{\delta}\sqrt{\frac{\log d}{n}} + H\delta$$

uniformly over all $t' \in \{0, \cdots, T'-1\}$, where the last inequality holds because $n = \Omega(s^2 \log d)$. Subsequently, by Hölder's inequality we have that

$$\sup_{0\leq t<T'}\|\widetilde{g}_t - g_t\|_{\psi^*}^2 \leq d^{2(a-1)/a} \cdot \sup_{0\leq t<T'}\|\widetilde{g}_t - g_t\|_\infty^2 \lesssim \frac{\sigma^2 \log d}{\delta^2 n} + H^2\delta^2. \tag{19}$$

We now consider the first term $\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t - g_t, x^* - x_t\rangle \leq \frac{1}{T'}\sum_{t=0}^{T'-1} X_t + \sup_{0\leq t\leq T'-1}\|\gamma_t\|_\infty \|x^* - x_t\|_1$, where $X_t := \langle \zeta_t, x^* - x_t\rangle$. By Lemma 2, we know that $X_t | X_1, \cdots, X_{t-1}$ is a centered sub-exponential random variable with parameters $\nu = \sqrt{n/2}\cdot\alpha \lesssim \sigma\|x^* - x_t\|_2/\delta\sqrt{n} \lesssim \sigma\|x^*\|_1/\delta\sqrt{n}$. Invoking concentration inequalities for sub-exponential martingales ((Victor, 1999), also phrased as Lemma 8 for a simplified version in the appendix) and the definition that $T' = T/n$, we have with probability $1 - \mathcal{O}(d^{-1})$

$$\left|\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \zeta_t, x^* - x_t\rangle\right| \lesssim \frac{\sigma\|x^*\|_1}{\delta}\left(\sqrt{\frac{\log d}{T}} + \frac{\log d}{T}\right) \lesssim \frac{\sigma\|x^*\|_1}{\delta}\sqrt{\frac{\log d}{T}},$$

where the last inequality holds because $T \geq n = \Omega(s^2 \log d)$. Thus,

$$\left|\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t - g_t, x^* - x_t\rangle\right| \lesssim \frac{\sigma\|x^*\|_1}{\delta}\sqrt{\frac{\log d}{T}} + \|x^*\|_1\left(H\delta + \frac{\sigma s \log d}{\delta n}\right). \tag{20}$$

Combining Eqs. (18,19,20) with Eq. (17) and taking x^* to be a minimizer of f on $\mathcal{X}$ that satisfies $\|x^*\|_1 \leq B$, we obtain

$$\begin{aligned}
&\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f(x^*) \\
&\lesssim \frac{\|x^*\|_1^2 \log d}{\eta}\frac{n}{T} + \frac{\sigma\|x^*\|_1}{\delta}\sqrt{\frac{\log d}{T}} + \|x^*\|_1\left(H\delta + \frac{\sigma s \log d}{\delta n}\right) + \eta\left(\frac{\sigma^2 \log d}{\delta^2 n} + H^2\delta^2\right) \\
&\leq \frac{B^2 \log d}{\eta}\frac{n}{T} + \frac{\sigma B}{\delta}\sqrt{\frac{\log d}{T}} + B\left(H\delta + \frac{\sigma s \log d}{\delta n}\right) + \eta\left(\frac{\sigma^2 \log d}{\delta^2 n} + H^2\delta^2\right)
\end{aligned} \tag{21}$$

with probability $1 - \mathcal{O}(d^{-1})$, provided that $\eta < \kappa/2H = 1/2H$.

We are now ready to prove Theorem 2. By the conditions we impose on T and the choices of η and n, it is easy to verify that $\eta < 1/2H$, $n = \Omega(s^2 \log d)$ and $n = \mathcal{O}(T)$. Subsequently,

$$
\begin{aligned}
&\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f(x^*) \\
&\lesssim B\sqrt{\frac{n\log d}{T}} + \sigma B\sqrt{\frac{n}{sT}} + B(\sigma+H)\sqrt{\frac{s\log d}{n}} + B\sqrt{\frac{n\log d}{T}}\left(\frac{\sigma^2}{s} + \mathcal{O}(n^{-1})\right) \\
&\lesssim B\left(\frac{(1+H)^2 s\log^2 d}{T}\right)^{1/4} + \frac{\sigma B\sqrt{(1+H)}}{s^{1/4}T^{1/4}} + \frac{B(\sigma+H)}{\sqrt{1+H}}\left(\frac{s\log^2 d}{T}\right)^{1/4} \\
&+ B\left(\frac{(1+H)^2 s\log d}{T}\right)^{1/4}\left(\frac{\sigma^2}{s} + \mathcal{O}(T^{-1/2})\right) \\
&\lesssim \left(B\sqrt{\log d} + \frac{\sigma B}{\sqrt{s}} + \frac{\sigma^2 B}{s}\right)\left[\frac{(1+H)^2 s}{T}\right]^{1/4} + B(\sigma+\sqrt{H})\sqrt{\log d}\left[\frac{s}{T}\right]^{1/4} + \mathcal{O}(T^{-1/2}) \\
&\lesssim (1+\sigma+\sigma^2/s)B\sqrt{\log d}\left[\frac{(1+H)^2 s}{T}\right]^{1/4} + \mathcal{O}(T^{-1/2}).
\end{aligned}
$$

7.5 Proof of Lemma 3

Using the model Eq. (2) we can decompose $\widetilde{g}_t(\delta) - g_t$ as

$$
\begin{aligned}
\widetilde{g}_t(\delta) - g_t = \frac{\delta}{2}\mathbb{E}\left[(z^\top H_t z)z\right] + \underbrace{\frac{1}{n\delta}Z_t^\top \xi}_{:=\widetilde{\zeta}_t(\delta)} + \underbrace{\frac{\delta}{2n}\sum_{i=1}^n (z_i^\top H_t z_i)z_i - \mathbb{E}[(z^\top H_t z)z]}_{:=\widetilde{\beta}_t(\delta)} \\
+ \underbrace{\frac{\delta}{2n}\sum_{i=1}^n (z_i^\top (H_t(\delta z_i) - H_t)z_i)z_i + \left[(\widehat{\Sigma} - I)(\widehat{\theta}_t - \theta_{0t})\right]_{1:d}}_{:=\widetilde{\gamma}_t(\delta)},
\end{aligned}
$$

where $\widehat{\Sigma}$, $\widehat{\theta}_t$ and θ_{0t} are similarly defined as in the proof of Lemma 2. The sub-exponentiality of $\langle\widetilde{\zeta}_t(\delta), a\rangle$ for any $a \in \mathbb{R}^d$ is established in Lemma 2. We next consider $\widetilde{\beta}_t(\delta)$. For any $a \in \mathbb{R}^d$ consider $\langle\widetilde{\beta}_t(\delta), a\rangle = \frac{\delta}{2n}\sum_{i=1}^n X_i(a)$ where $X_i(a) = (z_i^\top H_t z_i)(z_i^\top a) - \mathbb{E}[(z_i^\top H_t z_i)(z_i^\top a)]$ are centered i.i.d. random variables conditioned on H_t and x_t. In addition, $|X_i(a)| \le 2\|H_t\|_1\|z_i\|_\infty^2 \cdot \|a\|_1\|z_i\|_\infty \lesssim H\|a\|_1$ almost surely. Therefore, $X_i(a)$ is a sub-Gaussian random variable with parameter $\nu = H\|a\|_1$, and hence $\langle\widetilde{\beta}_t(\delta), a\rangle$ is a sub-Gaussian random variable with parameter $\nu = \delta H\|a\|_1/\sqrt{n}$. Finally, for the deterministic term $\widetilde{\gamma}_t(\delta)$, we have that

$$
\begin{aligned}
\|\widetilde{\gamma}_t(\delta)\|_\infty &\le \frac{\delta}{2}\sup_{z\in\{\pm 1\}^d}\|H_t(\delta z) - H_t\|_1\|z\|_\infty^2 + \|(\widehat{\Sigma} - I)(\widehat{\theta}_t - \theta_{0t})\|_\infty \\
&\le \frac{\delta}{2}\sup_{z\in\{\pm 1\}^d} L\cdot\|\delta z\|_\infty\|z\|_\infty^2 + \|\widehat{\Sigma} - I\|_{\max}\|\widehat{\theta}_t - \theta_{0t}\|_\infty
\end{aligned}
$$

$$\lesssim L\delta^2 + \sqrt{\frac{\log d}{n}}\left(\frac{\sigma s}{\delta}\sqrt{\frac{\log d}{n}} + s\delta H\right)$$
$$\lesssim L\delta^2 + \frac{\sigma s \log d}{n\delta} + s\delta H\sqrt{\frac{\log d}{n}}.$$

7.6 Proof of Theorem 3

Because f is convex, $R^{\mathsf{S}}_{\mathcal{A}}(T) = f(x_{T+1}) - f^* \le \frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f^*$. Thus it suffices to upper bound $\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f(x^*)$, where $x^* \in \mathcal{X}$, $\|x^*\|_1 \le B$ is a minimizer of f over $\mathcal{X}$. Using the strategy in the proof of Theorem 2, this amounts to upper bound (with high probability) $\|\widetilde{g}_t^{\mathsf{tw}} - g_t\|_{\psi^*}^2$ and $\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t^{\mathsf{tw}} - g_t, x^* - x_t\rangle$.

For the first term, using sub-exponentiality of $\widetilde{\zeta}_t$ and sub-gaussianity of $\widetilde{\beta}_t$, we have with probability $1 - \mathcal{O}(d^{-1})$ uniformly over all $t \in \{0, \cdots, T'-1\}$,

$$\begin{aligned}\|\widetilde{g}_t^{\mathsf{tw}} - g_t\|_\infty &\le \|\widetilde{\zeta}_t\|_\infty + \|\widetilde{\beta}_t\|_\infty + \|\widetilde{\gamma}_t\|_\infty \\ &\lesssim \frac{\sigma}{\delta}\left(\sqrt{\frac{\log d}{n}} + \frac{\log d}{n}\right) + \delta H\sqrt{\frac{\log d}{n}} + L\delta^2 + H\delta\sqrt{\frac{s^2\log d}{n}} + \frac{\sigma s\log d}{\delta n} \\ &\lesssim \left(\frac{\sigma}{\delta} + s\delta H\right)\sqrt{\frac{\log d}{n}} + L\delta^2,\end{aligned}$$

where the last inequality holds because $n = \Omega(s^2 \log d)$. Subsequently, with probability $1 - \mathcal{O}(d^{-1})$

$$\sup_{0\le t\le T'-1}\|\widetilde{g}_t^{\mathsf{tw}} - g_t\|_{\psi^*}^2 \lesssim \left(\frac{\sigma^2}{\delta^2} + s^2\delta^2 H^2\right)\frac{\log d}{n} + L^2\delta^4. \tag{22}$$

For the other term $\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t^{\mathsf{tw}} - g_t, x^* - x_t\rangle$, again using concentration inequalities of sub-exponential/sub-Gaussian martingales and noting that $\|x^* - x_t\|_2 \le \|x^* - x_t\|_1 \le 2B$, we have

$$\begin{aligned}\frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{g}_t^{\mathsf{tw}} - g_t, x^* - x_t\rangle &= \frac{1}{T'}\sum_{t=0}^{T'-1}\langle \widetilde{\zeta}_t + \widetilde{\beta}_t + \widetilde{\gamma}_t, x^* - x_t\rangle \\ &\lesssim \left(\frac{\sigma}{\delta} + s\delta H\right)B\sqrt{\frac{\log d}{T}} + B\left(L\delta^2 + \frac{\sigma s\log d}{\delta n} + s\delta H\sqrt{\frac{\log d}{n}}\right).\end{aligned} \tag{23}$$

Subsequently, combining Eqs. (22,23) with Eq. (17) we have

$$\begin{aligned}\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f(x^*) \lesssim \frac{B^2\log d}{\eta}\frac{n}{T} + \left(\frac{\sigma}{\delta} + s\delta H\right)B\sqrt{\frac{\log d}{T}} + (B+\eta)\left(L\delta^2 + \frac{\sigma s\log d}{\delta n} + s\delta H\sqrt{\frac{\log d}{n}}\right) \\ + \eta\left(\frac{\sigma^2}{\delta^2} + s^2\delta^2 H^2\right)\frac{\log d}{n} + \eta L^2\delta^4.\end{aligned} \tag{24}$$

We are now ready to prove Theorem 3. It is easy to verify that with the condition imposed on T and the selection of η and n, it holds that $\eta < 1/2H$, $n = \Omega(s^2\log d)$ and $n \le T/10$. Subsequently,

$$\frac{1}{T'}\sum_{t=0}^{T'-1} f(x_t) - f(x^*)$$

$$
\begin{aligned}
&\lesssim Bn^{1/3}\sqrt{\frac{\log d}{T}} + \left[\sigma\left(\frac{n}{s\log d}\right)^{1/3} + \mathcal{O}(n^{-1/3})\right] B\sqrt{\frac{\log d}{T}} \\
&+ \left(B + \mathcal{O}\left(\frac{n^{2/3}}{\sqrt{T}}\right)\right)\left[(L+\sigma)\left(\frac{s\log d}{n}\right)^{2/3} + \mathcal{O}(n^{-5/6})\right] \\
&+ Bn^{2/3}\sqrt{\frac{\log d}{T}}\left(\sigma^2\left(\frac{n}{s\log d}\right)^{2/3} + \mathcal{O}(n^{-2/3})\right)\frac{\log d}{n} + Bn^{2/3}\sqrt{\frac{\log d}{T}}L^2\left(\frac{s\log d}{n}\right)^{4/3} \\
&\lesssim Bn^{1/3}\sqrt{\frac{\log d}{T}} + \sigma B\left(\frac{n}{s\log d}\right)^{1/3}\sqrt{\frac{\log d}{T}} + B(L+\sigma)\left(\frac{s\log d}{n}\right)^{2/3} \\
&+ \sigma^2 B\left(\frac{n}{s^2\log^2 d}\right)^{1/3}\sqrt{\frac{\log d}{T}} + \mathcal{O}(T^{-5/12}) \\
&\lesssim \left(B\sqrt{\log d} + \frac{\sigma B\sqrt{\log d}}{s^{1/3}} + \frac{\sigma^2 B\sqrt{\log d}}{s^{2/3}}\right)\left[\frac{(1+L)s^{2/3}}{T}\right]^{1/3} + \frac{B(L+\sigma)}{(1+L)^{2/3}}\left(\frac{s^{2/3}\log d}{T}\right)^{1/3} + \mathcal{O}(T^{-5/12}) \\
&\lesssim \left(B\sqrt{\log d} + \frac{\sigma B\sqrt{\log d}}{s^{1/3}} + \frac{\sigma^2 B\sqrt{\log d}}{s^{2/3}}\right)\left[\frac{(1+L)s^{2/3}}{T}\right]^{1/3} + B\sigma\sqrt{\log d}\left(\frac{(1+L)s^{2/3}}{T}\right)^{1/3} + \mathcal{O}(T^{-5/12}) \\
&\lesssim (1+\sigma+\sigma^2/s^{2/3})B\sqrt{\log d}\left(\frac{(1+L)s^{2/3}}{T}\right)^{1/3} + \mathcal{O}(T^{-5/12}).
\end{aligned}
$$

Appendix: tail inequalities

Lemma 6. *Suppose X and Y are centered sub-Gaussian random variables with parameters ν_1^2 and ν_2^2, respectively. Then XY is a centered sub-exponential random variable with parameter $\nu = \sqrt{2}v$ and $\alpha = 2v$, where $v = 2e^{2/e+1}\nu_1\nu_2$.*

Proof. XY is clearly centered because $\mathbb{E}XY = \mathbb{E}X \cdot \mathbb{E}Y = 0$, thanks to independence. We next bound $\mathbb{E}[|XY|^k]$ for $k \geq 3$ (i.e., verification of the Bernstein's condition). Because X and Y are independent, we have that $\mathbb{E}[|XY|^k] = \mathbb{E}|X|^k \cdot \mathbb{E}|Y|^k$. In addition, because X is a centered sub-Gaussian random variable with parameter ν_1^2, it holds that $(\mathbb{E}|X|^k)^{1/k} \leq \nu_1 e^{1/e}\sqrt{k}$. Similarly, $(\mathbb{E}|X|^k)^{1/k} \leq \nu_2 e^{1/e}\sqrt{k}$. Subsequently,

$$
\mathbb{E}|XY|^k \leq \left(e^{2/e}\nu_1\nu_2\right)^k \cdot k^k \leq \left(e^{2/e}\nu_1\nu_2\right)^k \cdot e^k k! \leq \frac{1}{2}k! \cdot \left(2e^{2/e+1}\nu_1\nu_2\right)^k.
$$

where in the second inequality we use the Stirling's approximation inequality that $\sqrt{2\pi k}k^k e^{-k} \leq k!$. The sub-exponential parameter of XY can then be determined. □

Lemma 7 (Bernstein's inequality)**.** *Suppose X is a sub-exponential random variable with parameters ν and α.*

$$
\Pr\left[\left|X - \mathbb{E}X\right| > t\right] \leq \begin{cases} 2\exp\left\{-t^2/2\nu^2\right\}, & 0 < t \leq \nu^2/\alpha; \\ 2\exp\left\{-t/2\alpha\right\}, & t > \nu^2/\alpha. \end{cases}
$$

The following lemma is a simplified version of Theorem 1.2A in (Victor, 1999) (note that the original form in (Victor, 1999) is one-sided; the two-sided version below can be trivially obtained by considering $-X_1, \cdots, -X_n$ and applying the union bound).

Lemma 8 (Bernstein's inequality for martingales)**.** *Suppose $X_1, \cdots, X_n$ are random variables such that $\mathbb{E}[X_j | X_1, \cdots, X_{j-1}] = 0$ and $\mathbb{E}[X_j^2 | X_1, \cdots, X_{j-1}] \leq \sigma^2$ for all $t = 1, \cdots, n$. Further assume that $\mathbb{E}[|X_j|^k | X_1, \cdots, X_{j-1}] \leq \frac{1}{2} k! \sigma^2 b^{k-2}$ for all integers $k \geq 3$. Then for all $t > 0$,*

$$\Pr\left[\left|\sum_{j=1}^{n} X_j\right| \geq t\right] \leq 2 \exp\left\{-\frac{t^2}{2(n\sigma^2 + bt)}\right\}.$$

References

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

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

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

Bandeira, A. S., Scheinberg, K., & Vicente, L. N. (2012). Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization. *Mathematical Programming (Series B)*, *134*, 223–257.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lan, G. (2012). An optimal method for stochastic composite optimization. *Mathematical Programming (Series A)*, *133*, 365–397.

Leeds, D. D., Pyles, J. A., & Tarr, M. J. (2014). Exploration of complex visual feature spaces for object perception. *Frontiers in computational neuroscience*, *8*.

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

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

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

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

Nemirovski, A., & Yudin, D. (1983). *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication.

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

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

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

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

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

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

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

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

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

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

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

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (Lasso). *IEEE transactions on information theory*, *55*(5), 2183–2202.

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

Zhang, C.-H., & Zhang, S. S. (2014). 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.

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