
Sparse Polyak: an adaptive step size rule for high-dimensional M-estimation

Tianqi Qiao

Texas A&M University
College Station, TX, USA
tianqi.qiao@tamu.edu

Marie Maros

Texas A&M University
College Station, TX, USA
mmaros@tamu.edu

Abstract

We propose and study Sparse Polyak, a variant of Polyak’s adaptive step size, designed to solve high-dimensional statistical estimation problems where the problem dimension is allowed to grow much faster than the sample size. In such settings, the standard Polyak step size performs poorly, requiring an increasing number of iterations to achieve optimal statistical precision—even when, the problem remains well conditioned and/or the achievable precision itself does not degrade with problem size. We trace this limitation to a mismatch in how smoothness is measured: in high dimensions, it is no longer effective to estimate the Lipschitz smoothness constant. Instead, it is more appropriate to estimate the smoothness restricted to specific directions relevant to the problem (restricted Lipschitz smoothness constant). Sparse Polyak overcomes this issue by modifying the step size to estimate the restricted Lipschitz smoothness constant. We support our approach with both theoretical analysis and numerical experiments, demonstrating its improved performance.

1 Introduction

Consider the high-dimensional statistical estimation problem

$$\min_{\mathbb{R}^d \ni \theta: \|\theta\|_0 \leq s} f(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(z_i, \theta), \quad (1)$$

with data points $z_i \in \mathbb{R}^d, i = 1 \dots n$. We focus on the regime in which the dimensionality grows much faster than the sample size, i.e. $\frac{d}{n} \rightarrow \infty$. To obtain consistent estimates in this setting, it is necessary to assume that the true solution exhibits some low-dimensional structure—such as sparsity. In (1) sparsity is enforced through the ℓ_0 constraint, which guarantees that θ will have at most s non-zero elements. This constraint renders the problem in (1) non-convex and, in general, NP-hard, regardless of the objective function f Natarajan (1995). Nevertheless, under certain assumptions on the data, various algorithms have been developed to efficiently find approximate solutions to (1). Notably, under suitable assumptions that hold for a variety of statistical models, the Iterative Hard Thresholding (IHT) algorithm has been shown to efficiently find sufficiently accurate solutions to (1). The IHT algorithm results from applying projected gradient descent to (1) and reads

$$\theta_{t+1} = \text{HT}_s(\theta_t - \gamma \nabla f(\theta_t)),$$

where HT_s denotes the hard thresholding operator. HT_s retains the s largest-magnitude components of its input and sets the remaining to zero. Here, $\gamma > 0$ denotes the step-size which ought to be chosen as $\gamma = 2/(3\bar{L})$ Jain et al. (2014), where $\bar{L}$ denotes the *restricted* Lipschitz smoothness (RSS) constant, and can be interpreted as the Lipschitz smoothness constant of f when restricted to sparse directions.

For a variety of statistical models, such as Generalized Linear Models (GLMs), $\bar{L}$ remains constant as long as $\frac{s \log(d)}{n}$ remains constant, even if $\frac{d}{n} \rightarrow \infty$. This insight underpins the *rate invariance* of IHT: under suitable conditions, the number of iterations required to achieve (near) optimal statistical precision remains constant even as both d and n grow.

Analogous to the Lipschitz smoothness constant, the RSS constant must be estimated in practice. Thus, the natural question in this context, and the starting point to the work in this paper is: **(i)** Do already existing approaches to adaptively tune γ via the estimation of the Lipschitz smoothness constant work in the high-dimensional setting? Our criteria to determine whether a step-size rule works in the high dimension, additional to convergence will be determined by the answers to the following questions: **(ii)** Can they achieve the same or better guarantees than by choosing the optimal fixed step-size? **(iii)** Can they guarantee rate invariance as $\frac{d}{n} \rightarrow \infty$?

1.1 Related works

Over the past three decades, numerous methods have been proposed to solve the problem in (1), including Matching Pursuit Mallat and Zhang (1993), Orthogonal Matching Pursuit Pati et al. (1993), and CoSaMP Needell and Tropp (2009). Iterative Hard Thresholding (IHT) was first introduced in Blumensath and Davies (2009) with many variants proposed since. While initial convergence guarantees for IHT required the Restricted Isometry Property (RIP)-a condition often too stringent in practice-Jain et al. (2014) extended IHT's convergence guarantees to problems fulfilling the restricted strong convexity (RSC) and RSS; establishing linear convergence to near optimal statistical precision. Observe that in the M-estimation context, convergence to arbitrary precision is unnecessary and convergence to the best achievable statistical precision is preferred. However, for the results in Jain et al. (2014)

to hold, if the optimal parameter to recover is s^* -sparse, $s \geq \mathcal{O}(\bar{\kappa}^2 s^*)$ is required, where $\bar{\kappa}$ is the restricted condition number. Khanna and Kyrillidis (2018) proposed an accelerated version of IHT which was extended to the stochastic setting in Zhou et al. (2018). Zhou et al. (2018) establishes that with a sufficiently large mini-batch size, acceleration can be achieved and the faster rate requires only $s \geq \mathcal{O}(\bar{\kappa} s^*)$. Axiotis and Sviridenko (2022) proposed a variant of IHT with an adaptively chosen weighted ℓ_2 penalty for which only $s \geq \mathcal{O}(\bar{\kappa} s^*)$ is required. They further establish that for IHT $s \geq \mathcal{O}(\bar{\kappa} s^*)$ is in fact a necessary condition to achieve near optimal statistical precision. Li et al. (2016) and Shen and Li (2018) introduced variants of IHT that incorporate variance reduction. Shen and Li (2017b) and Yuan et al. (2018) propose Partial Hard Thresholding and Gradient Hard Thresholding pursuit respectively, with a focus on support recovery under high SNR assumptions. Yuan and Li (2021) establishes generalization bounds for solutions found via IHT. Further, Zhang et al. (2025) establishes that IHT, for a range of $\bar{\kappa}$, finds solutions that can be shown to achieve the oracle estimation rate under a high SNR condition.

With no exception, the discussed works establishing convergence under the RSS require knowledge of the RSS constant $\bar{L}$. Knowledge of $\bar{L}$ is crucial for the requirement $s \geq \mathcal{O}(\bar{\kappa}^2 s^*)$ being sufficient for convergence to optimal statistical precision. More generally, convergence can be established whenever $s \geq \mathcal{O}\left(\frac{s^*}{\bar{\mu}^2 \gamma^2}\right)$, with $\gamma \leq \frac{1}{\bar{L}}$. Consequently, overestimating $\bar{L}$ forces a smaller step size γ , which in turn slows down convergence and leads to denser solutions.

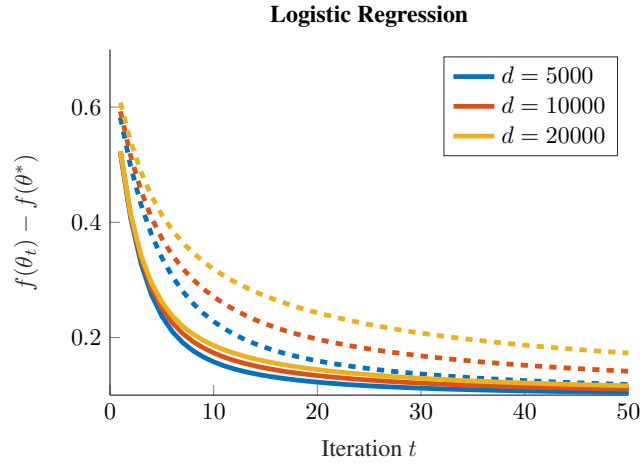


Figure 1: Performance of Polyak's step size (dashed) and Sparse Polyak (solid) on logistic regression problems with increasing d and n . The quantities s , s^* , $\bar{\kappa}$ and $\frac{\log(d)}{n}$ remain constant. With Polyak's step-size the performance degrades as d increases whereas Sparse Polyak exhibits rate invariance, i.e. the number of iterations to achieve (near) optimal statistical precision does not change.

Several recent works have proposed adaptive step-size schemes based on estimating the local Lipschitz constant Malitsky and Mishchenko (2020, 2024); Latafat et al. (2024). With the goal to jointly exploit the function’s local regularity and the algorithm’s trajectory, Mishkin et al. (2024) derive new convergence results for gradient descent as a function of the objective function’s local Lipschitz smoothness and strong convexity parameters. They additionally establish that Polyak’s step-size obtains fast and path-dependent rates. However, their results do not naturally extend to constrained problems, particularly non-convex ones like (1).

Polyak’s original step-size rule, first proposed in Polyak (1969), has gathered renewed attention in the machine learning community Ren et al. (2022); Hazan and Kakade (2019); Loizou et al. (2021); Wang et al. (2023); Zamani and Glineur (2024), but it remains ill-suited for non-convex constrained settings such as (1) unless additional assumptions are imposed. Some efforts have been made to adapt Polyak’s step size to constrained problems. For instance, Cheng and Li (2012) addresses box constraints, while Devanathan and Boyd (2024) consider convex constraints.

To this day, except for Li et al. (2024), and works that employ inexact line-search strategies Xiao and Zhang (2013); Wang et al. (2014), there has been no study of adaptive step-size schemes in the high-dimensional context. The work in Li et al. (2024) handles (1) in the stochastic setting and proposes the use of Polyak’s step-size with no modification. The results in Li et al. (2024) are limited even with no stochasticity, as they imply bounds on the restricted condition number, i.e. $\bar{\kappa} \leq \frac{\eta}{\eta-1}$ where $\eta = (\sqrt{5} + 1)^2/4$. Further, it is worth mentioning that the notion of RSS assumed in Li et al. (2024) is more restrictive than that assumed in the present paper and Jain et al. (2014). Further, as shown in Fig. 1 Polyak’s step-size with no modification presents a performance that degrades as the size of the problem increases even if $\bar{\kappa}$ and the optimal statistical precision (of the order of $\mathcal{O}(\frac{s^* \log d}{n})$ for this particular statistical model) remain constant. This effect is highly undesirable in the high-dimensional setting and as shown in the present work, can be avoided with a suitable modification of Polyak’s step-size.

1.2 Major contributions

Our main contribution is the first adaptive step-size rule that performs well in high-dimensions and preserves the rate invariance property. To develop this scheme, we address question (i) posed in the introduction. We observe that adaptive step-size rules that estimate the Lipschitz smoothness constant do not necessarily work well in high-dimensions. This is because, in many common statistical models, the Lipschitz smoothness constant scales as $\mathcal{O}(d)$ with high probability. Consequently, we answer questions (i) through (iii) in the negative, by demonstrating empirically (c.f. Fig. 1) that estimating the Lipschitz smoothness constant via Polyak’s step-size (dashed line) does not yield rate invariance as d grows even if $\frac{\log(d)}{n}$ is held constant.

To overcome this limitation, we design an adaptive step-size rule that estimates the restricted Lipschitz smoothness constant instead. With this modification, we answer questions (ii) and (iii) in the affirmative. This is captured in Theorem 1 and its Corollaries, which particularize the results of Theorem 1 to relevant statistical models. In Sections 3, 6, and in Appendix D we provide theoretical and empirical evidence, both on synthetic and real data, that our proposed method outperforms Polyak’s step-size for high-dimensional M-estimation tasks and achieves rate invariance. Moreover, we theoretically and empirically show that Sparse Polyak converges to optimal statistical precision at least as fast as IHT with the optimal fixed step-size $\gamma = \mathcal{O}(1/\bar{L})$. We also establish sufficient conditions under which we can guarantee support recovery, and particularize our results for specific statistical learning models in Section 4.

Our guarantees are derived under standard assumptions, identical to those in Jain et al. (2014); Shen and Li (2017b); Yuan et al. (2018). To guarantee convergence to optimal statistical precision we require the knowledge of $f(\theta^*)$ where θ^* denotes the true parameter. While $f(\theta^*)$ is known to be of the order $\mathcal{O}(\frac{\log d}{n})$, its exact value is typically unknown. Consequently, we provide in Appendix B a double loop method that estimates a surrogate to $f(\theta^*)$ and allows convergence to optimal statistical precision as long as lower bound to $f(\theta^*)$ is known. Observe that in our context $f(\theta^*) \geq 0$ making 0 a valid lower bound.

In addition, we prove linear convergence for statistical models that do not satisfy the regularity conditions in Jain et al. (2014); Axiotis and Sviridenko (2022); Yuan and Li (2021) but instead fulfill

the weaker condition in Loh and Wainwright (2015). This extends the results in Jain et al. (2014); Yuan and Li (2021) to cover additional GLMs with an adaptive step-size rule. We provide these additional results in Appendix C.2.

Our proof technique may be of independent interest as it provides a clear pathway to establishing convergence of the IHT algorithm. We believe this is key in extending theoretical guarantees to adaptive step-size schemes to solve (1). We provide a sketch of the proof of our main results in Section 5 and the formal proof in Appendix A.

Finally, although our analysis applies only to Polyak’s step-size, we conjecture that other adaptive step-size rules such as Barzilai and Borwein (1988), Zhou et al. (2025), Malitsky and Mishchenko (2020) may also suffer from similar performance degradation in high dimensions, and would therefore require analogous re-engineering to be effective.

Notation. Throughout this paper, we adopt the following notations. For vectors $x, \in \mathbb{R}^d$, we denote by x_i the i^{th} element. We denote the ℓ_∞ -norm of x as $\|x\|_\infty$, the Euclidean norm as $\|x\|$, the ℓ_1 -norm as $\|x\|_1$, and the ℓ_0 -norm as $\|x\|_0$. Recall $\|x\|_0 = |\{i : x_i \neq 0\}|$. Also, we let $|x|_{\min}$ denote the minimal entry of x in the sense of absolute value. The inner product between two vectors is denoted as $\langle x, y \rangle$. Matrices such as $X \in \mathbb{R}^{d \times d}$, are capitalized. For any matrix Σ , we denote its largest singular value by $\sigma_{\max}(\Sigma)$ and its smallest singular value by $\sigma_{\min}(\Sigma)$. Similarly, if $\Sigma \in \mathbb{R}^{d \times d}$ diagonalizes, we use $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ to denote the largest and smallest eigenvalues of Σ respectively. The Frobenius norm of X is given by $\|X\|_F$, and the nuclear norm by $\|X\|_*$.

2 Setup and background

We make the following assumptions regarding the objective function f .

Assumption 1 (RSC Agarwal et al. (2012)). *The objective function f is (μ, τ) –restricted strongly convex in $\mathbb{R}^d$, i.e.*

$$\frac{\mu}{2} \|\theta_1 - \theta_2\|^2 - \frac{\tau}{2} \|\theta_1 - \theta_2\|_1^2 \leq f(\theta_1) - f(\theta_2) - \langle \nabla f(\theta_2), \theta_1 - \theta_2 \rangle, \forall \theta_1, \theta_2 \in \mathbb{R}^d. \quad (2)$$

Assumption 2 (RSS Agarwal et al. (2012)). *The objective function f is (L, τ) –restricted smooth in $\mathbb{R}^d$, i.e.*

$$f(\theta_1) - f(\theta_2) - \langle \nabla f(\theta_2), \theta_1 - \theta_2 \rangle \leq \frac{L}{2} \|\theta_1 - \theta_2\|^2 + \frac{\tau}{2} \|\theta_1 - \theta_2\|_1^2, \forall \theta_1, \theta_2 \in \mathbb{R}^d.$$

Assumptions 1 and 2 extend the classical notions of strong convexity and L –Lipschitz smoothness. These assumptions reduce to their classical counterparts when τ is sufficiently small and the direction $\theta_1 - \theta_2$ is appropriately sparse. Observe that when $\theta_1 - \theta_2$ is dense and f is convex, Assumption 1 becomes vacuous and the upper bound in Assumption 2 scales linearly with the problem dimension d . This highlights that the RSC and RSS depend not only on the magnitude of the direction $\theta_1 - \theta_2$ but also its structure.

In high-dimensional statistical learning settings where $\frac{d}{n} \rightarrow \infty$, standard strong convexity and smoothness assumptions fail to hold. However, many important problems still satisfy variants of the RSC and RSS, with both μ and L remaining dimension-independent and with τ exhibiting only moderate dependence on d , e.g., Jain et al. (2014); Agarwal et al. (2012). We leverage this in Section 4 to establish fast computational and near optimal statistical guarantees for a variety of high-dimensional statistical learning problems.

We further highlight that our results can be generalized by adopting a weaker RSC condition, where we assume that (2) holds only for pairs θ, θ^* satisfying $\|\theta - \theta^*\| \leq 1$, where θ denotes the ground truth, rather than requiring it to hold globally. This relaxation broadens the applicability of our approach, allowing it to accommodate a wider class of functions. Notably, for some generalized linear models (GLMs), the loss function does not necessarily satisfy (2) without imposing additional constraints. We provide a detailed discussion of these results in Appendix C.2.

3 Main Result

In this section we present our main theoretical result, which establishes convergence guarantees for Sparse Polyak (c.f. Algorithm 1). In this section, we focus on the deterministic setting of f , while

the statistical case will be addressed in Section 4. Note that IHT corresponds to projected gradient descent when applied to (1). The projection onto the ℓ_0 is given by the hard thresholding operator, already discussed in the introduction and formally defined as follows.

Definition 1 (Hard Thresholding Operator). *For any $s > 0$, and $z \in \mathbb{R}^d$, we define $\text{HT}_s(z)$ as the projection of z on $B_0(s) := \{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq s\}$. i.e.,*

$$\text{HT}_s(z) = \underset{\theta \in B_0(s)}{\operatorname{argmin}} \|\theta - z\|_2,$$

where ties are broken lexicographically.

Algorithm 1 Iterative Hard Thresholding (IHT) with Polyak Step-Size

- 1: **Input:** Function f , target function value $\hat{f}$, sparsity parameter s , number of iterations T
 - 2: **Initialize:** $\theta_0 \in \mathbb{R}^d$, with $\|\theta_0\|_0 \leq s$
 - 3: **for** $t = 0$ to $T - 1$ **do**
 - 4: Compute step-size $\gamma_t = \frac{\max\{f(\theta_t) - \hat{f}, 0\}}{5\|\text{HT}_s(\nabla f(\theta_t))\|^2}$
 - 5: Update: $\theta_{t+1} = \text{HT}_s(\theta_t - \gamma_t \nabla f(\theta_t))$
 - 6: **end for**
 - 7: **Output:** θ_T
-

The adaptive step-size rule in Algorithm 1 differs from the classical Polyak rule by replacing $\|\nabla f(\theta_t)\|^2$ with $\|\text{HT}_s(\nabla f(\theta_t))\|^2$. This approach contrasts with the low-dimensional case and with the work in Li et al. (2024). Observe that even if the current iterate θ_t is sparse, sparsity of $\nabla f(\theta_t)$ can not be guaranteed. In fact, the worst-case relationship $\|\nabla f(\theta_t)\| \leq \sqrt{\frac{d}{s}} \|\text{HT}_s(\nabla f(\theta_t))\|$, which holds with equality for a vector in which all coordinates are identical, may hold. As a result, using the full gradient norm can lead to using overly conservative step-sizes, slowing down convergence dramatically as d increases. Unless strong additional conditions are imposed on the problem, convergence may be even be jeopardized (see discussion in Section 5 for more details).

Before providing our main result we introduce some notation. Consider fixed values $s \geq s^* > 0$, and let $f^* \triangleq \min_{\theta: \|\theta\|_0 \leq s^*} f(\theta)$. Assume the chosen target function value satisfies $\hat{f} \geq f^*$. Let $\bar{L} = L + 3\tau s$, $\bar{\mu} = \mu - 3\tau s$, and $\bar{\kappa} = \bar{L}/\bar{\mu}$ denote the restricted condition number.

Theorem 1. *Let $\{\theta_t\}_{t \geq 1}$ denote the sequence of iterates generated by Algorithm 1. Suppose the objective function f satisfies the RSC and RSS in Assumptions 1 and 2, respectively. Let $\hat{\theta}$ be any s^* -sparse vector such that $f(\hat{\theta}) = \hat{f}$, and assume $\bar{\mu} > 0$ and $s \geq (240\bar{\kappa})^2 s^*$. Then, for any iterate θ_t such that $\|\theta_t - \hat{\theta}\|^2 \geq \frac{36\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2}{\bar{\mu}^2}$ we can guarantee*

$$\|\theta_{t+1} - \hat{\theta}\|^2 \leq \left(1 - \frac{1}{80\bar{\kappa}}\right) \|\theta_t - \hat{\theta}\|^2.$$

Moreover, let $t_0 \geq 0$ be the first iteration for which $\|\theta_{t_0} - \hat{\theta}\|^2 < \frac{36\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2}{\bar{\mu}^2}$. Then, for all $t \geq t_0$, $\|\theta_t - \hat{\theta}\|^2 \leq \left(1 + \frac{1}{80\bar{\kappa}}\right) \frac{36\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2}{\bar{\mu}^2}$.

Theorem 1 implies linear convergence at a rate scaling with $\bar{\kappa}^{-1}$ up to precision $\mathcal{O}\left(\frac{\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2}{\bar{\mu}^2}\right)$. This result is near equivalent to that in Theorem 3 in Jain et al. (2014), where the RSS constant is assumed to be known, up to a constant factor. Thus, we successfully answer in the affirmative question (ii) posed in Section 1. Further, observe that, if $\bar{L}$, $\bar{\mu}$ and $\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2$ remain constant as the size of the problem grows, the rate remains unchanged and the achievable precision does so too. This implies that for a variety of statistical models the rate and final precision will remain invariant as the problem size increases, as long as the aforementioned quantities do not change. Thus, we also answer in the affirmative question (iii) posed in Section 1. We particularize the result to specific statistical models and provide further discussion in Section 4 and Appendix C.

Here, $\hat{f}$ is a user-defined target value that reflects the desired level of optimization, which can be set above or equal to the statistical accuracy of the problem. Such a relaxation is natural in learning

problems, as it is often counterproductive to optimize to full precision. The continuity and restricted strong convexity of f imply that an s^* -sparse vector $\hat{\theta}$ such that $f(\hat{\theta}) = \hat{f}$ always exists.

The additional factor $\frac{1}{80\bar{\kappa}}$ in the final precision (c.f. Theorem 1 when $t \geq t_0$) stems from the expansiveness of the hard thresholding operator. The removal of this factor and support recovery guarantees can be achieved under the signal to noise ratio (SNR) condition (c.f.(3)) and are provided in the following Corollary. Such a condition is widely used in hard thresholding and support recovery studies, as seen in (Shen and Li, 2017a, Prop 2) Bouchot et al. (2016); Shen and Li (2017c); Yuan et al. (2016).

Corollary 1. *Under the assumptions stated in Theorem 1, if, further, the SNR condition:*

$$|\hat{\theta}|_{\min} \geq \frac{7\|\text{HT}_s(\nabla f(\hat{\theta}))\|}{\bar{\mu}} \quad (3)$$

holds, for any $t \geq t_0$, where t_0 is defined in Theorem 1, the support of θ_t contains that of $\hat{\theta}$, and the sequence $\|\theta_t - \hat{\theta}\|^2$ is non-increasing and upper bounded by $\frac{36\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2}{\bar{\mu}^2}$.

Algorithm 1 offers an approach to obtain a target accuracy that we assume known in advance to (1), without requiring precise knowledge of L, μ, τ . For scenarios in which we only have access to a lower bound on the problem, Algorithm 2 serves as an alternative; in most learning problems, the bound can be simply set to 0, making the method broadly applicable. This adaptive variant of Algorithm 1 builds on the framework of Hazan and Kakade (2019), which reviews gradient descent with Polyak's step size and its double-loop counterpart. By updating the lower bound adaptively in an outer loop, the method ensures that either the accuracy $\mathcal{O}\left(\frac{\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2}{\bar{\mu}^2}\right)$ is attained or the updated lower bound remains valid.

Algorithm 2 IHT with Adaptive Polyak

```

1: Input: Function  $f$ , a lower bound  $\tilde{f}_1$ , and sparsity parameter  $s$ .
2: Initialize:  $\theta_0 = 0 \in \mathbb{R}^d$ 
3: for  $k=1$  to  $K$  do
4:   for  $t = 0$  to  $T - 1$  do
5:     Compute step-size  $\gamma_t = \frac{f(\theta_t) - \tilde{f}_k}{10\|\text{HT}_s(\nabla f(\theta_t))\|^2}$ 
6:     Update:  $\theta_{t+1} = \text{HT}_s(\theta_t - \gamma_t \nabla f(\theta_t))$ 
7:   end for
8:    $\bar{\theta}_k = \arg\min_{t \leq T} f(\theta_t)$ 
9:    $\tilde{f}_{k+1} = \frac{f(\bar{\theta}_k) + \tilde{f}_k}{2}$ 
10:   $\theta_0 = \bar{\theta}_k$ 
11: end for
12: Output:  $\bar{\theta} = \arg\min_{k \leq K} f(\bar{\theta}_k)$ 

```

Let $s \geq s^* > 0$, and define $f^* := \min_{\|\theta\|_0 \leq s^*} f(\theta)$, attained by some s^* -sparse vector θ^* .

Theorem 2. *Consider the iterates $\{\bar{\theta}_k\}$ generated by Algorithm 2. Assume that the function f fulfills Assumptions 1 and 2. Then for $\varepsilon = (1 + \frac{1}{160\bar{\kappa}}) \frac{36(\bar{L} + \bar{\mu})\|\text{HT}_s(\nabla f(\theta^*))\|^2}{\bar{\mu}^2}$, when $\bar{\mu} > 0, s \geq (480\bar{\kappa})^2 s^*$, Algorithm 2 requires at most $\tilde{T} := \left(1 + \log_2 \frac{2(f(\theta_0) - f(\theta^*))}{\varepsilon}\right) T$ gradient evaluations to achieve $f(\bar{\theta}) - f(\theta^*) \leq \varepsilon$ and $\|\bar{\theta} - \theta^*\|^2 \leq (1 + \frac{1}{160\bar{\kappa}}) \frac{36\|\text{HT}_s(\nabla f(\theta^*))\|^2}{\bar{\mu}^2}$. Here $T = \left\lceil \frac{1}{\log(1/(1-1/160\bar{\kappa}))} \log \left(\frac{\bar{\mu}^2 \|\theta_0 - \theta^*\|^2}{36(1+1/160\bar{\kappa})\|\text{HT}_s(\nabla f(\theta^*))\|^2} \right) \right\rceil$.*

This theorem focuses on the distance of the iterates to θ^* . The quantity T can be interpreted as the number of iterations required to reach the desired accuracy when applying Algorithm 1 with $\hat{f} = f^*$. The additional term $\mathcal{O}\left(\frac{f(\theta_0) - f(\theta^*)}{\varepsilon}\right)$ in the definition of $\tilde{T}$ corresponds to the number of outer iterations needed to obtain a sufficiently tight lower bound for the targeted accuracy. Similar order guarantees are established in Theorems 1 and 2. The proof of Theorem 2 is provided in Appendix B.

4 Statistical Guarantees

The results in Section 3 are deterministic in nature and consequently, do not depend on a data generation model. In, contrast, in this section we use Theorem 1 to provide guarantees for specific statistical models. Corollaries 2 and 3 establish the computational-statistical performance guarantees for sparse logistic regression and low-rank matrix regression respectively. We provide guarantees for additional statistical models, including sparse linear regression, in Appendix C.

4.1 Logistic Regression

We consider a dataset consisting of observations $z_i = (x_i, y_i)$ for $i = 1, \dots, n$, where $x_i \in \mathbb{R}^d$ denote the feature vectors, and $y_i \in \{0, 1\}$ denote the corresponding responses. The feature vectors are organized into the design matrix

$$X \triangleq (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times d}.$$

We assume that the relationship between y_i and x_i follows the model

$$\Pr(y_i = 1 \mid x_i) = \frac{1}{1 + \exp(-x_i^\top \theta^*)}, \quad (4)$$

where θ^* is an s^* -sparse vector representing the underlying ground truth parameter. The objective function is defined as

$$f(\theta) = \frac{1}{n} \sum_{i=1}^n \log(1 + \exp(x_i^\top \theta)) - y_i x_i^\top \theta.$$

We assume that each covariate vector x_i is drawn independently from a multivariate normal distribution $\mathcal{N}(0, \Sigma)$, where Σ is a non-singular covariance matrix. By invoking Corollary 1 in Yuan and Li (2021), we can establish that the objective function $f(\theta)$ satisfies the RSS and RSC conditions, as formalized in the following lemma.

Lemma 1. *Consider the sparse linear logistic regression problem described above. Suppose the covariates x_i are uniformly bounded such that $\|x_i\| \leq 1$ for all $i \in [n]$. Then $f(\theta)$ is $\bar{L}$ -smooth with $\bar{L} = 1$. Moreover, with probability at least $1 - e^{-c_0 n}$, the RSC condition holds with curvature parameter $\mu := \frac{1}{2} \exp(-4R) \sigma_{\min}(\Sigma)$ and tolerance $\tau := c_1 \exp(-4R) \zeta(\Sigma) \frac{\log d}{n}$, where $R := \|\theta^*\|$, $\zeta(\Sigma) = \max_{i=1, \dots, d} \Sigma_{ii}$, and $c_0, c_1 > 0$ are universal constants.*

Corollary 2. *Consider the sparse linear logistic regression problem described above. Under the assumptions of Lemma 1, further suppose that the sample size is sufficiently large so that $\bar{\mu} > 0$. Further, assume the design matrix $X \in \mathbb{R}^{n \times d}$ is normalized such that $\|X_j / \sqrt{n}\| \leq C$ for all $j = 1, \dots, d$. Let $\{\theta_t\}_{t \geq 0}$ be the sequence of iterates produced by Algorithm 1 when applied to the sparse logistic regression problem. Assume the sparsity parameter satisfies $s \geq (240 \bar{\kappa})^2 s^*$, and $\hat{f} = f(\theta^*)$. Then, with probability at least $1 - e^{-c_0 n} - \frac{2}{d}$, the following hold:*

- (i) If $\|\theta_t - \theta^*\|^2 \geq 72C^2 \frac{s \log d}{n \bar{\mu}^2}$, the iterates exhibit contraction toward θ^* , i.e., $\|\theta_{t+1} - \theta^*\|^2 \leq (1 - \frac{1}{80 \bar{\kappa}}) \|\theta_t - \theta^*\|^2$.
- (ii) Let t_0 denote the first iteration at which $\|\theta_{t_0} - \theta^*\|^2 < 72C^2 \frac{s \log d}{n \bar{\mu}^2}$. Then for all $t \geq t_0$, the iterates remain confined in a neighborhood of θ^* : $\|\theta_t - \theta^*\|^2 \leq (1 + \frac{1}{80 \bar{\kappa}}) 72C^2 \frac{s \log d}{n \bar{\mu}^2}$.
- (iii) If θ^* satisfies the SNR condition (3), then the iterates remain confined in a neighborhood of θ^* $\|\theta_t - \theta^*\|^2 \leq 72C^2 \frac{s \log d}{n \bar{\mu}^2}$ for all $t \geq t_0$, and the support of θ^* is exactly recovered and preserved for all subsequent iterations.

The proof of Corollary 2 is provided in Appendix A.3.

Remark 1. *The assumption $\|x_i\| \leq 1, \forall i \in [n]$ is required in Yuan and Li (2021) to provide performance guarantees of HT with a fixed step-size on Logistic Regression. However, we note that this assumption is extremely restrictive in the high-dimensional setting. We provide additional results that do not require $\|x_i\| \leq 1, \forall i \in [n]$ in Appendix C.2. Our results in Appendix C.2 further apply to additional GLMs that do not satisfy the RSC condition (2) globally, and are aligned with the results in Loh and Wainwright (2015) both in terms of sample and asymptotic convergence rates. In these cases, convergence at a rate of $(1 - c_0 / \bar{\kappa})$ can be guaranteed if the algorithm is suitably initialized. However, a stricter requirement on the sparsity level, specifically $s \geq \mathcal{O}(\bar{\kappa}^4) s^*$,*

is necessary otherwise, and a rate of $(1 - 1/c_1\bar{\kappa}^2)$ can be guaranteed. Here c_0 , and c_1 are universal constants.

The result above and the result in Appendix C.2 match the observed behavior of Sparse Polyak in Fig 1. Namely, we observe that Sparse Polyak indeed achieves rate invariance, as $\frac{d}{n} \rightarrow \infty$, the rate and final precision remain constant as long as $\frac{s \log d}{n}$ is left unchanged.

4.2 Matrix Regression

Consider the data generation model

$$y_i = \langle X_i, \Theta^* \rangle + \varepsilon_i, \quad \text{for } i = 1, 2, \dots, n,$$

where $\Theta^* \in \mathbb{R}^{d \times d}$ is a matrix of rank at most s^* . We assume, $X_i \in \mathbb{R}^{d \times d}$, $\text{vec}(X_i) \sim \mathcal{N}(0, \Sigma)$ are i.i.d, and $\Sigma \succ 0$. Further, $\varepsilon_i \sim N(0, \sigma^2)$ are i.i.d. and independent of X_i . Define $f(\Theta) = \frac{1}{2n} \sum_{i=1}^n (y_i - \langle X_i, \Theta \rangle)^2$.

By invoking (Agarwal et al., 2012, Lemma 7), we establish the RSS and RSC properties of $f(\Theta)$, as formalized in the following lemma.

Lemma 2. *Consider the low-rank matrix regression problem described above. Then, with probability at least $1 - e^{-c_0 n}$, $f(\Theta)$ satisfies the RSS and RSC conditions with respect to the Frobenius norm and the nuclear norm. The corresponding parameters are given by:*

$$L = 2\sigma_{\max}(\Sigma), \quad \mu = \frac{1}{2}\sigma_{\min}(\Sigma), \quad \text{and} \quad \tau = c_1\zeta(\Sigma)\frac{d}{n},$$

where $\zeta(\Sigma) := \sup_{\|u\|=1, \|v\|=1} \text{Var}(u^\top X_1 v)$, and $c_0, c_1 > 0$ are universal constants.

To enforce the suitable low-rank structure on the iterates we define

$$\text{PM}_s(W) = \sum_{i=1}^s \sigma_i u_i v_i^\top,$$

where $\sigma_i, i = 1 \dots, s$ are the s largest singular values of W , and u_i, v_i the corresponding singular vectors. We substitute all instances of the HT_s operator by PM_s in Algorithm 1 and 2 (c.f. Appendix B).

Corollary 3. *Consider the low rank matrix regression problem described above. Let $\{\Theta_t\}_{t \geq 0}$ be the sequence of iterates generated by Algorithm 1 when applied to a low rank matrix regression problem. Suppose that $\hat{f} = f(\Theta^*)$, n is sufficiently large such that $\bar{\mu} > 0$, and $s \geq (240\bar{\kappa})^2 s^*$. Then, with probability at least $1 - e^{-c_0 n} - 2e^{-4d}$, the following holds: (i) If $\|\Theta_t - \Theta^*\|_F^2 \geq \frac{7200\sigma^2\zeta(\Sigma)sd}{n\bar{\mu}^2}$, then the iterates contract relative to Θ^* as $\|\Theta_{t+1} - \Theta^*\|_F^2 \leq (1 - \frac{1}{80\bar{\kappa}}) \|\Theta_t - \Theta^*\|_F^2$. (ii) Let t_0 be the first iteration for which $\|\Theta_{t_0} - \Theta^*\|_F^2 < \frac{7200\sigma^2\zeta(\Sigma)sd}{n\bar{\mu}^2}$. Then, for all $t \geq t_0$, the iterates remain in a stable neighborhood around Θ^* , with $\|\Theta_t - \Theta^*\|_F^2 \leq (1 + \frac{1}{80\bar{\kappa}}) \frac{7200\sigma^2\zeta(\Sigma)sd}{n\bar{\mu}^2}$.*

The proof of Corollary 3 can be found in Appendix A.4.

Remark 2. *The result in Corollary 1 also extends to Algorithm 2. If θ^* satisfies the SNR condition (3), the iterates of Algorithm 2 recover the support after $\tilde{T}$ gradient descent steps.*

5 Sketch of the Proof for Theorem 1 and Corollary 1

We provide a sketch of the proof of the main results of the paper. We refer the reader to the appendix for the complete proof. The proof of Theorem 1 follows the outline: (i) study the behavior of $\|\theta_{t+1} - \hat{\theta}\|^2$ given γ_t and $\|\theta_t - \hat{\theta}\|$, (ii) establish that under the assumption that γ_t is sufficiently large the expansive effect of the Hard Thresholding operator can be offset by the contractive effect of the gradient update, (iii) show that γ_t is sufficiently large until we reach optimal statistical precision. Both to finalize the proof of Theorem 1 and to establish Corollary 1: (iv) the iterates remain confined within a neighborhood of θ^* , and, given that the SNR condition holds, the support can be identified, providing further benefits to the algorithm's performance. We elaborate on points (i) through (iv).

(i) To understand the dynamics of $\|\theta_t - \hat{\theta}\|^2$ we analyze the combined effect of gradient descent under the RSC and RSS and the Hard Thresholding operator.

For this we suitably apply Lemma 1 from Jain et al. (2014), exploit the RSC, and the properties of θ_{t+1} in relation to HT_s to yield:

$$\|\theta_{t+1} - \theta^*\|^2 \leq \left(1 + \sqrt{\frac{s^*}{s}}\right)^2 ((1 - \bar{\mu}\gamma_t) \|\theta_t - \theta^*\|^2 \quad (5a)$$

$$-2\gamma_t (f(\theta_t) - f(\theta^*)) + 10\gamma_t^2 \|\text{HT}_s(\nabla f(\theta_t))\|^2). \quad (5b)$$

Setting $\gamma_t = \frac{f(\theta_t) - f(\theta^*)}{5\|\text{HT}_s(\nabla f(\theta_t))\|^2}$ makes (5b) nonpositive and yields a choice that is invariant with d . The use of $\|\text{HT}_s(\nabla f(\theta_t))\|^2$ in γ_t and the RSS are critical to avoiding a step-size that scales with the Lipschitz smoothness constant (which scales as $\mathcal{O}(d)$ in the high-dimensional setting). Setting $\gamma_t = 1/(40\bar{L})$ in (5) recovers the result in Jain et al. (2014), which establishes that IHT requires at most $\mathcal{O}(\bar{\kappa}^{-1} \log(1/\varepsilon))$ to be within ε -accuracy of near optimal statistical precision.

(ii) To achieve the optimal linear rate under our choice of step size, we require that

$$\left(1 + \sqrt{\frac{s^*}{s}}\right)^2 (1 - \bar{\mu}\gamma_t) \leq 1 - c_0 \frac{\bar{\mu}}{\bar{L}}. \quad (6)$$

Sufficient conditions for (6) are $s \geq c_1 \bar{\kappa}^2 s^*$ and $\gamma_t \geq \frac{c_2}{\bar{L}}$ for some universal constants c_0, c_1 and c_2 . However, if γ_t were to scale with the Lipschitz smoothness constant, i.e. with d^{-1} we would require $s \geq d^2 s^*$ to establish linear convergence. Observe that this requirement can not be fulfilled as $s \leq d$.

(iii) We exploit the RSS and RSC to show that $\gamma_t \geq \frac{1}{40\bar{L}}$ when $\|\theta_t - \hat{\theta}\|^2 \geq \frac{36\|\nabla f(\hat{\theta})\|_s^2}{\bar{\mu}^2}$. If, further, $s \geq (240\bar{\kappa})^2 s^*$, (6) holds with $c_0 = 1/160$, yielding $\|\theta_{t+1} - \hat{\theta}\|^2 \leq (1 - c_0 \bar{\kappa}^{-1}) \|\theta_t - \hat{\theta}\|^2$.

The condition $\|\theta_t - \hat{\theta}\|^2 \geq \frac{36\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2}{\bar{\mu}^2}$ stems from (1) being a constrained problem. If the Polyak step-size had been left unaltered, additional regularity conditions are required to establish convergence in the constrained case. As established in Theorem 3 in Polyak (1969), one such condition is $\frac{f(\theta_t) - f(\hat{\theta})}{\|\theta_t - \hat{\theta}\|} \geq c$ for some $c > 0$ and any $\theta_t \in \mathbb{R}^d$, which does not uniformly in high-dimensional M-estimation. For some GLMs, we can show the condition holds locally and exploit this fact to provide more general results in Appendix C.2.

(iv) From (i)-(iii) it follows that there exists t_0 at which $\|\theta_{t_0} - \hat{\theta}\|^2 \leq \frac{36\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2}{\bar{\mu}^2}$. Since the potential expansion is at most $(1 + \frac{1}{80\bar{\kappa}})$ Theorem 1 follows. If $\hat{\theta}$ satisfies (3), we show that the inequality $\|\theta_{t_0} - \hat{\theta}\|^2 \leq \frac{36\|\text{HT}_s(\nabla f(\hat{\theta}))\|^2}{\bar{\mu}^2}$ guarantees that $\hat{\mathcal{S}} \subset \mathcal{S}_{t_0}$. From here, we establish that, this results in two possible scenarios: (a) $\gamma_t < \frac{1}{40\bar{L}}$, implying $\hat{\mathcal{S}} \subset \mathcal{S}_{t_0+1}$, ensuring that $\|[\tilde{\theta}_{t_0+1} - \hat{\theta}]_{\hat{\mathcal{S}}_{t_0+1}}\|^2 = \|\tilde{\theta}_{t_0+1} - \hat{\theta}\|^2$, and eliminating the expansion term in (5); or, (b) $\gamma_t \geq \frac{1}{40\bar{L}}$, and (6) holds. Based on (a) and (b), Corollary 1 is established by induction.

6 Numerical experiments

We first consider sparse linear regression and sparse logistic regression on synthetic data. This is done to illustrate the algorithm's performance as the size of the problem grows while the problem conditioning is kept the same. In all scenarios that rely on synthetic data, we set $d \in \{5000, 10000, 20000\}$, and $s^* = 300$. The design matrix $X \in \mathbb{R}^{d \times n}$ is generated to reflect a time-series structure with a correlation parameter $\omega = 0.5$. We set the sample size n according to $n = \lceil \alpha s \log d \rceil$, where $\alpha > 0$ is a constant. In this section, we use $\alpha = 5$. For each column index $j \in \{1, \dots, n\}$, we generate a sequence of i.i.d. standard normal variables $\varepsilon_1, \dots, \varepsilon_{d-1}$, and construct $x_{1,j} = \varepsilon_1 / \sqrt{1 - \omega^2}$. The subsequent entries are generated recursively as $x_{t+1,j} = \omega x_{t,j} + \varepsilon_t$ for $t \in \{1, \dots, d-1\}$, where $\varepsilon_t \sim \mathcal{N}(0, 1)$. The true parameter θ^* is created by sampling each entry from $\mathcal{N}(0, 1)$, and assigning nonzero values to s^* randomly chosen entries.

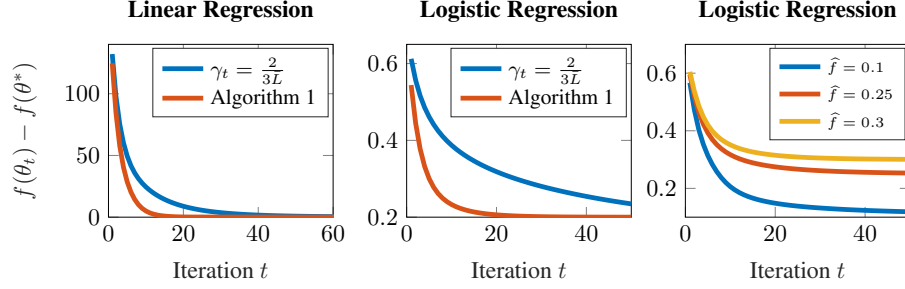


Figure 2: **Left and center:** IHT with $\frac{2}{3L}$ (blue) vs. Algorithm 1 (red) on linear and logistic regression respectively. **Right:** Choice of $\hat{f}$ on Algorithm 1. In all scenarios $\alpha = 5$, $d = 5000$ and $s = 700$.

In the case of linear regression, each sample i (where $i \in \{1, \dots, n\}$) is generated according to the model:

$$y_i = x_i^T \theta^* + w_i, \quad w_i \sim \mathcal{N}(0, 0.25).$$

For logistic regression, the relationship between y_i and x_i follows the model (4).

(i) Comparison to fixed step-size: In Figure 6 we present a comparison between IHT with a fixed step size (blue) and the adaptive step size used in Algorithm 1 (red) when solving linear regression (left panel) and logistic regression (center panel) respectively. When working with a constant step-size, we set the step size to $\frac{2}{3L}$ following Jain et al. (2014) for both linear and logistic regression. $\bar{L}$ for linear regression can be upper bounded as $\lambda_{\max}(\Sigma)(3 + \frac{2(2s+s^*)}{s\alpha})$ (Loh and Wainwright, 2015, Appendix D.1), whereas $\bar{L}$ for logistic regression is one fourth of that of linear regression. In both settings, $\lambda_{\max}(\Sigma) \leq \frac{2}{(1-\omega)^2(1+\omega)}$ Agarwal et al. (2012). Although both step size strategies share the same theoretical guarantees, we observe that the adaptive step size speeds-up convergence. This advantage arises because an adaptive step size can adapt to the local curvature and therefore be significantly more aggressive, allowing for faster progress towards the solution. As expected, this effect is more pronounced when solving logistic regression where the functions' curvature will depend on the point.

(ii) Rate invariance: In Figure 1 (c.f. Section 1), we compare the classical Polyak step size with the step size employed in Algorithm 1 across different problem dimensions d , while maintaining $\alpha = 5$. The solid line represents the performance of Algorithm 1, whereas the dashed line corresponds to the classical Polyak step size. This demonstrates that when the condition number of Σ remains unchanged, the complexity of the method remains almost identical under our chosen step size. In contrast, Polyak's step size leads to an increased number of iterations, even if the achievable statistical precision and $(\lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma))$ remain the same.

(iii) Choice of $\hat{f}$: Finally, Figure 6 (right) highlights the impact of the choice of $\hat{f}$. The results confirm that $\hat{f}$ determines the best achievable accuracy. Additionally, from the formulation of γ_t , we observe that $\hat{f}$ directly influences the step size magnitude, thus is impacting the convergence rate.

Experiments on real world data are provided in Appendix D. All experiments were conducted on a laptop equipped with 16 GB of RAM and a 12th Gen Intel Core i5-12500H 3.10 GHz CPU.

References

- Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452–2482, 2012.
- Kyriakos Axiotis and Maxim Sviridenko. Iterative hard thresholding with adaptive regularization: Sparser solutions without sacrificing runtime. In *International Conference on Machine Learning*, pages 1175–1197. PMLR, 2022.
- Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.
- Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.
- Jean-Luc Bouchot, Simon Foucart, and Pawel Hitczenko. Hard thresholding pursuit algorithms: Number of iterations. *Applied and Computational Harmonic Analysis*, 41(2):412–435, 2016. doi: <https://doi.org/10.1016/j.acha.2016.03.002>.
- David Chapman and Ajay Jain. Musk (Version 2) data set. UCI Machine Learning Repository, 1994. DOI: <https://doi.org/10.24432/C51608>.
- Wanyou Cheng and Donghui Li. An active set modified Polak–Ribière–Polyak method for large-scale nonlinear bound constrained optimization. *Journal of Optimization Theory and Applications*, 155: 1084–1094, 2012.
- Nikhil Devanathan and Stephen Boyd. Polyak minorant method for convex optimization. *Journal of Optimization Theory and Applications*, pages 1–20, 2024.
- Elad Hazan and Sham Kakade. Revisiting the Polyak step size. *arXiv preprint arXiv:1905.00313*, 2019.
- Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional M-estimation. *Advances in neural information processing systems*, 27:685–693, 2014.
- Rajiv Khanna and Anastasios Kyrillidis. IHT dies hard: Provable accelerated iterative hard thresholding. In *International Conference on Artificial Intelligence and Statistics*, pages 188–198. PMLR, 2018.
- Puya Latafat, Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient. *Mathematical Programming*, pages 1–39, 2024.
- Changhao Li, Zhixin Ma, Dazhi Sun, Guoming Zhang, and Jinming Wen. Stochastic IHT with stochastic Polyak step-size for sparse signal recovery. *IEEE Signal Processing Letters*, 31:2035–2039, 2024. doi: 10.1109/LSP.2024.3426353.
- Xingguo Li, Tuo Zhao, Raman Arora, Han Liu, and Jarvis Haupt. Stochastic variance reduced optimization for nonconvex sparse learning. In *International Conference on Machine Learning*, pages 917–925. PMLR, 2016.
- Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.
- Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak step-size for SGD: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.
- Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In *Proceedings of the 37th International Conference on Machine Learning*, pages 6702–6712, 2020.

- Yura Malitsky and Konstantin Mishchenko. Adaptive proximal gradient method for convex optimization. *Advances in Neural Information Processing Systems*, 37:100670–100697, 2024.
- Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.
- Aaron Mishkin, Ahmed Khaled, Yuanhao Wang, Aaron Defazio, and Robert Gower. Directional smoothness and gradient methods: Convergence and adaptivity. *Advances in Neural Information Processing Systems*, 37:14810–14848, 2024.
- Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.
- Deanna Needell and Joel A Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3):301–321, 2009.
- Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.
- Mehdi Neshat, Bradley Alexander, Nataliia Sergiienko, and Markus Wagner. Large-scale Wave Energy Farm data set. UCI Machine Learning Repository, 2020. DOI: <https://doi.org/10.24432/C5GG7Q>.
- Yagyensh Chandra Pati, Ramin Rezaifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on signals, systems and computers*, pages 40–44. IEEE, 1993.
- B.T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969. doi: [https://doi.org/10.1016/0041-5553\(69\)90061-5](https://doi.org/10.1016/0041-5553(69)90061-5).
- Tongzheng Ren, Fuheng Cui, Alexia Atsidakou, Sujay Sanghavi, and Nhat Ho. Towards statistical and computational complexities of Polyak step size gradient descent. In *International Conference on Artificial Intelligence and Statistics*, pages 3930–3961. PMLR, 2022.
- Jie Shen and Ping Li. Partial hard thresholding: Towards a principled analysis of support recovery. In *Advances in Neural Information Processing Systems*, volume 30, 2017a.
- Jie Shen and Ping Li. Partial hard thresholding: Towards a principled analysis of support recovery. *Advances in Neural Information Processing Systems*, 30, 2017b.
- Jie Shen and Ping Li. On the iteration complexity of support recovery via hard thresholding pursuit. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3115–3124, 2017c.
- Jie Shen and Ping Li. A tight bound of hard thresholding. *Journal of Machine Learning Research*, 18(208):1–42, 2018.
- Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.
- Xiaoyu Wang, Mikael Johansson, and Tong Zhang. Generalized Polyak step size for first order optimization with momentum. In *International Conference on Machine Learning*, pages 35836–35863. PMLR, 2023.
- Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.
- Lin Xiao and Tong Zhang. A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization*, 23(2):1062–1091, 2013.
- Xiao-Tong Yuan, Ping Li, and Tong Zhang. Exact recovery of hard thresholding pursuit. In *Advances in Neural Information Processing Systems*, volume 29, pages 3558–3566, 2016.

- Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43, 2018.
- Xiaotong Yuan and Ping Li. Stability and risk bounds of iterative hard thresholding. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1702–1710, 2021.
- Moslem Zamani and François Glineur. Exact convergence rate of the subgradient method by using Polyak step size. *arXiv preprint arXiv:2407.15195*, 2024.
- Yanhang Zhang, Zhifan Li, Shixiang Liu, Xueqin Wang, and Jianxin Yin. Rethinking hard thresholding pursuit: Full adaptation and sharp estimation. *arXiv preprint arXiv:2501.02554*, 2025.
- Danqing Zhou, Shiqian Ma, and Junfeng Yang. AdaBB: Adaptive Barzilai-Borwein method for convex optimization. *Mathematics of Operations Research*, 2025.
- Pan Zhou, Xiaotong Yuan, and Jiashi Feng. Efficient stochastic gradient hard thresholding. *Advances in Neural Information Processing Systems*, 31:1988–1997, 2018.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Yes. The main contributions, including the novel theoretical analysis and assumptions, are explicitly stated in the abstract and introduction, and are rigorously supported in the main sections.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: The paper discusses the limitations of the proposed method, particularly its reliance on prior knowledge of the optimal objective value. This limitation is clearly stated in the main text. To address this issue, a double-loop variant of the algorithm is proposed in the appendix, which can be applied when the required information is unavailable.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: For each theoretical result, the paper explicitly states all required assumptions. A proof sketch of the main theorem is included in the main text, while complete and rigorous proofs of all results are provided in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The paper provides sufficient details to reproduce the main experimental results, including the data generation process, parameter choices, and implementation details. These are described in the numerical experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code is not provided, as the implementations involve standard optimization routines with no algorithmic novelty. All data used in the experiments are either synthetically generated using clearly described procedures or drawn from publicly available sources. The paper includes sufficient details to allow reproduction of the results without requiring access to the code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all experimental settings. Full details of the synthetic data and standard optimization benchmarks are provided in the numerical experiments section of the main body, while the settings for the real-data experiments are presented in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: This paper focuses on deterministic optimization problems, where the goal is to evaluate algorithmic performance rather than statistical model accuracy. As such, the use of error bars or statistical significance tests is not applicable or necessary in this context.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: This paper does not involve large-scale computation; the experiments are limited to evaluating optimization performance on moderate-size synthetic problems and standard benchmarks. All relevant information is provided in the Numerical Experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Yes, the research conducted in the paper conforms to the NeurIPS Code of Ethics in every respect.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: The focus of the paper is purely on theoretical advancements in optimization methods. The primary objective is to improve algorithmic performance and theoretical understanding, which does not have direct societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses open-source datasets, and proper credit is given to the creators or original owners of these datasets. The license and terms of use for these datasets are respected, as specified by the respective sources.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

Appendix

A Main Theorems	21
A.1 Proof of Theorem 1	24
A.2 Proof of Corollary 1	24
A.3 Proof of Corollary 2	25
A.4 Proof of Corollary 3	26
B Adaptive Lower Bound	26
B.1 Proof of Theorem 2	26
C Other Statistical Guarantees	27
C.1 Sparse Linear Regression	27
C.2 Generalized Linear Models	29
D Experiments on real data	32
D.1 Linear Regression	32
D.2 Logistic Regression	32

A Main Theorems

In this section we provide the formal proof to the statements included in the main body of the paper. To formally establish Theorem 1 we build on (Jain et al., 2014, Lemma 1) and Lemmas 4-6. (Jain et al., 2014, Lemma 1) allows us to control the expansive properties of the Hard Thresholding operator, whereas Lemma 4 allows us to establish (5) (c.f. Section 5). Further, in Lemma 5 we establish consequences of the RSS (Assumption 2) that are instrumental in establishing a lower bound on the step-size γ_t . We lower bound γ_t in Lemma 6 and establish conditions under which this lower bound holds. We combine these results in the proof of Theorem 1 in Appendix A.1, followed by a formal proof of the corollaries included in the main body of the paper in the remaining sections of Appendix A.

For simplicity, we let $\widehat{S}_t = S_t \cup \widehat{S}$. Also, we let $g_t = \nabla f(\theta_t)$, and $\widehat{g} = \nabla f(\widehat{\theta})$. When discussing specific statistical models we denote by θ^* the ground truth, $g^* = \nabla f(\theta^*)$ and $S^* = \text{supp}(\theta^*)$. For any index set S and vector $\theta \in \mathbb{R}^d$, we define $[\theta]_S$ as the vector that retains the entries indexed by S , while setting all other entries to zero.

We include the following fundamental lemma, which plays a key role in our analysis. It corresponds to Lemma 1 in Jain et al. (2014), and is presented here for completeness.

Lemma 3. *For any index set I , any $z \in \mathbb{R}^I$, let $\theta = \text{HT}_s(z)$. Then for any $\widehat{\theta} \in \mathbb{R}^I$ such that $\|\widehat{\theta}\|_0 \leq s^*$, we have*

$$\|\theta - z\|^2 \leq \frac{|I| - s}{|I| - s^*} \|\widehat{\theta} - z\|^2.$$

Lemma 4. *Let $\widehat{\theta}$ be any s^* -sparse vector, and θ_t be any s -sparse vector. Assume that the function f fulfills Assumption 1 with $\bar{\mu} = \mu - 3\tau s > 0$. Let $\theta_{t+1} := \text{HT}_s(\theta_t - \gamma_t g_t)$, and $\widetilde{\theta}_{t+1} := \theta_t - \gamma_t g_t$. For any $\gamma_t \leq \frac{f(\theta_t) - f(\widehat{\theta})}{5\|\text{HT}_s(\widehat{\theta})\|^2}$ we have*

$$\begin{aligned} \|\theta_{t+1} - \widehat{\theta}\|^2 &\leq \left(1 + \sqrt{\frac{s^*}{s}}\right)^2 \left\| [\widetilde{\theta}_{t+1}]_{\widehat{S}_{t+1}} - \widehat{\theta} \right\|^2 \\ &\leq \left(1 + \sqrt{\frac{s^*}{s}}\right)^2 (1 - \bar{\mu}\gamma_t) \|\theta_t - \widehat{\theta}\|^2. \end{aligned}$$

Proof. For the first inequality,

$$\|\theta_{t+1} - \widehat{\theta}\| \stackrel{(i)}{\leq} \|\theta_{t+1} - [\widetilde{\theta}_{t+1}]_{\widehat{S}_{t+1}}\| + \|[\widetilde{\theta}_{t+1}]_{\widehat{S}_{t+1}} - \widehat{\theta}\| \stackrel{(ii)}{\leq} \left(1 + \sqrt{\frac{s^*}{s}}\right) \|[\widetilde{\theta}_{t+1}]_{\widehat{S}_{t+1}} - \widehat{\theta}\|,$$

where in (i) we use the triangle inequality and in (ii) we used Lemma 3.

For the second inequality, we consider the expansion

$$\begin{aligned}\|[\tilde{\theta}_{t+1}]_{\hat{\mathcal{S}}_{t+1}} - \hat{\theta}\|^2 &= \|[\tilde{\theta}_{t+1} - \hat{\theta}]_{\hat{\mathcal{S}}_{t+1}}\|^2 \\ &= \|[\theta_t - \hat{\theta}]_{\hat{\mathcal{S}}_{t+1}}\|^2 - 2\gamma_t \langle \theta_t - \hat{\theta}, [g_t]_{\hat{\mathcal{S}}_{t+1}} \rangle + \gamma_t^2 \| [g_t]_{\hat{\mathcal{S}}_{t+1}} \|^2,\end{aligned}$$

where

$$-2\gamma_t \langle \theta_t - \hat{\theta}, [g_t]_{\hat{\mathcal{S}}_{t+1}} \rangle = -2\gamma_t \langle \theta_t - \hat{\theta}, g_t \rangle + 2\gamma_t \langle \theta_t - \hat{\theta}, [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} \rangle,$$

and consequently

$$\left\| [\tilde{\theta}_{t+1} - \hat{\theta}]_{\hat{\mathcal{S}}_{t+1}} \right\|^2 \leq \|\theta_t - \hat{\theta}\|^2 - 2\gamma_t \langle \theta_t - \hat{\theta}, g_t - [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} \rangle + \gamma_t^2 \left\| [g_t]_{\hat{\mathcal{S}}_{t+1}} \right\|^2$$

Using the RSC yields

$$\begin{aligned}\|[\tilde{\theta}_{t+1} - \hat{\theta}]_{\hat{\mathcal{S}}_{t+1}}\|^2 &\leq (1 - \bar{\mu}\gamma_t) \|\theta_t - \hat{\theta}\|^2 - 2\gamma_t (f(\theta_t) - f(\hat{\theta})) + \gamma_t^2 \|\text{HT}_{s+s^*}(g_t)\|^2 \\ &\quad + 2\gamma_t \langle \theta_t - \hat{\theta}, [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} \rangle.\end{aligned}\tag{7}$$

To obtain (5) we must upper bound the inner product in (7), for which we have:

$$\begin{aligned}\langle [\theta_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}}, \gamma_t [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} \rangle &= \langle [\theta_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}}, \gamma_t [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} \rangle \\ &= \langle [\theta_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} - \gamma_t [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} + \gamma_t [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}}, \gamma_t [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} \rangle \\ &\leq \|[\theta_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} - \gamma_t [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}}\| \|\gamma_t [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}}\| + \|\gamma_t [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}}\|^2.\end{aligned}$$

Then,

$$\|[\theta_t - \gamma_t g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}}\| \stackrel{(i)}{\leq} \|[\theta_t - \gamma_t g_t]_{\mathcal{S}_t \setminus \mathcal{S}_{t+1}}\| \stackrel{(ii)}{\leq} \|[\theta_t - \gamma_t g_t]_{\mathcal{S}_{t+1} \setminus \mathcal{S}_t}\| = \|[\gamma_t g_t]_{\mathcal{S}_{t+1} \setminus \mathcal{S}_t}\|,$$

where in (i) we use that $\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1} \subseteq \mathcal{S}_t \setminus \mathcal{S}_{t+1}$, and in (ii) we exploit that $|\mathcal{S}_t \setminus \mathcal{S}_{t+1}| = |\mathcal{S}_{t+1} \setminus \mathcal{S}_t|$ and that $\mathcal{S}_{t+1}$ contains the indexes of the s largest elements of $\theta_t - \gamma_t g_t$. Thus, we obtain the overall upper bound

$$2\gamma_t \langle \theta_t - \hat{\theta}, [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} \rangle \leq 2\gamma_t^2 \| [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} \| \| [g_t]_{\mathcal{S}_{t+1} \setminus \mathcal{S}_t} \| + 2\gamma_t^2 \| [g_t]_{\mathcal{S}_t \setminus \hat{\mathcal{S}}_{t+1}} \|^2,$$

which together with (7) yields

$$\|[\tilde{\theta}_{t+1} - \hat{\theta}]_{\hat{\mathcal{S}}_t}\|^2 \leq (1 - \bar{\mu}\gamma_t) \|\theta_t - \hat{\theta}\|^2 - 2\gamma_t (f(\theta_t) - f(\hat{\theta})) + 5\gamma_t^2 \|\text{HT}_{2s}(g_t)\|^2.\tag{8}$$

Given the upperbound on γ_t , the two right most terms together are negative, and we thus the proof is complete. $\square$

Observe that in the proof of Lemma 4 we use the iterate $\tilde{\theta}_{t+1}$ to treat the effect of the hard thresholding operator and gradient descent separately. We then restrict to the support $\hat{\mathcal{S}}_{t+1}$ to avoid the scaling of any bound with the ambient dimension d .

Lemma 5. (RSS-gradient bound) Assume that f fulfills Assumptions 1 where $\bar{\mu} = \mu - 3\tau s > 0$, and Assumption 2. Then, for any pair x, y of s -sparse vectors there holds

$$\frac{1}{2(L + 3\tau s)} \|\text{HT}_s(\nabla f(x) - \nabla f(y))\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle.$$

Proof. We define

$$\phi(t) = f(t) - \langle \nabla f(x), t - x \rangle.$$

From its formulation, we know $\phi(t)$ inherits the RSS and RSC property of f .

As a result, for any $2s$ -sparse vector z , we have

$$\begin{aligned}\phi(x) &\leq \phi(z) - \langle 0, z - x \rangle - \frac{\bar{\alpha}}{2} \|x - z\|^2 \\ &= \phi(z) - \frac{\bar{\alpha}}{2} \|x - z\|^2 \\ &\leq \phi(z) \\ &\leq \phi(y) + \langle \nabla \phi(y), z - y \rangle + \frac{L + 3\tau s}{2} \|z - y\|^2.\end{aligned}$$

Set $z = y - \frac{1}{L+3\tau s} \text{HT}_s(\nabla\phi(y))$, the inequality above indicates

$$\phi(x) \leq \phi(y) - \frac{1}{2(L+3\tau s)} \|\text{HT}_s(\nabla\phi(y))\|^2,$$

which is equivalent to

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2(L+3\tau s)} \|\text{HT}_s(\nabla f(y) - \nabla f(x))\|^2 \leq f(y).$$

□

Lemma 6. Consider the iterates $\{\theta_t\}_{t \geq 1}$ generated by Algorithm 1 to solve (1). Assume that f fulfills Assumptions 1 where $\bar{\mu} = \mu - 3\tau s > 0$, and 2. Further, denote by $\hat{\theta}$ an arbitrary s^* -sparse vector for which $f(\hat{\theta})$ is known and desirable. Then, the step-size

$$\gamma_t = \frac{f(\theta_t) - f(\hat{\theta})}{5\|\text{HT}_s(g_t)\|^2} \geq \frac{1}{40\bar{L}}$$

for each $t \geq 0$ for which

$$\begin{aligned} \|\theta_t - \hat{\theta}\|^2 &\geq \frac{18\|\text{HT}_{s+s^*}(\hat{g})\|^2}{\bar{\mu}^2} \\ f(\theta_t) &> f(\hat{\theta}). \end{aligned}$$

Proof. Assume that $f(\theta_t) - f(\hat{\theta}) > 0$, then

$$\gamma_t \geq \frac{f(\theta_t) - f(\hat{\theta})}{10\|\text{HT}_s(g_t - \hat{g})\|^2 + 10\|\text{HT}_s(\hat{g})\|^2}.$$

Given that f fulfills Assumption 2 we may invoke Lemma 5 yielding the bound

$$\gamma_t \geq \frac{f(\theta_t) - f(\hat{\theta})}{20\bar{L}(f(\theta_t) - f(\hat{\theta}) - \langle \hat{g}, \theta_t - \hat{\theta} \rangle) + 10\|\text{HT}_s(\hat{g})\|^2}.$$

If

$$10 \left(2\bar{L}\langle \hat{g}, \hat{\theta} - \theta_t \rangle + \|\text{HT}_s(\hat{g})\|^2 \right) \leq 20\bar{L}(f(\theta_t) - f(\hat{\theta})) \quad (9)$$

we can guarantee that

$$\gamma_t \geq \frac{1}{40\bar{L}}.$$

Rearranging (9) we have that the condition can be equivalently written as

$$\|\text{HT}_s(\hat{g})\|^2 \leq 2\bar{L} \left(f(\theta_t) - f(\hat{\theta}) + \langle \hat{g}, \theta_t - \hat{\theta} \rangle \right).$$

Invoking the RSC, a sufficient condition for the above is

$$\|\text{HT}_s(\hat{g})\|^2 \leq 2\bar{L} \left(\frac{\bar{\mu}}{2} \|\theta_t - \hat{\theta}\|^2 + 2\langle \hat{g}, \theta_t - \hat{\theta} \rangle \right),$$

which can be guaranteed as long as

$$\begin{aligned} \|\text{HT}_s(\hat{g})\|^2 &\leq 2\bar{L} \left(\frac{\bar{\mu}}{2} \|\theta_t - \hat{\theta}\|^2 - 2\|\text{HT}_{s+s^*}(\hat{g})\| \|\theta_t - \hat{\theta}\| \right) \\ &= 2\bar{L} \left(\frac{\bar{\mu}}{2} \|\theta_t - \hat{\theta}\| - 2\|\text{HT}_{s+s^*}(\hat{g})\| \right) \|\theta_t - \hat{\theta}\|. \end{aligned}$$

To guarantee that the above holds it is sufficient to request that $\|\theta_t - \hat{\theta}\| \geq \frac{106}{25\bar{\mu}} \|\text{HT}_{s+s^*}(\hat{g})\|$, and thus the result follows.

□

A.1 Proof of Theorem 1

As a consequence of Lemma 6 we distinguish three cases for any t : **(i)** $\|\theta_t - \hat{\theta}\|^2 \geq \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2}$ and $f(\theta_t) - \hat{f} > 0$, **(ii)** $\|\theta_t - \hat{\theta}\|^2 < \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2}$ and $f(\theta_t) - \hat{f} > 0$, or **(iii)** $f(\theta_t) - \hat{f} \leq 0$.

In case **(iii)** no progress is made, i.e. $\theta_{t+1} = \theta_t$ and by the RSC there holds

$$\frac{\bar{\mu}}{2} \|\theta_t - \hat{\theta}\|^2 \leq \|\theta_t - \hat{\theta}\| \|\text{HT}_{s+s^*}(\hat{g})\|$$

and thus

$$\|\theta_{t+1} - \hat{\theta}\|^2 = \|\theta_t - \hat{\theta}\|^2 < \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2}.$$

Further, from the above it follows that if $\|\theta_t - \hat{\theta}\|^2 \geq \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2}$ we can guarantee that $\gamma_t > 0$.

For case **(i)** we begin by invoking Lemma 4, which guarantees that

$$\|\theta_{t+1} - \hat{\theta}\|^2 \leq \left(1 + 3\sqrt{\frac{s^*}{s}}\right)^2 (1 - \bar{\mu}\gamma_t) \|\theta_t - \hat{\theta}\|^2.$$

Using Lemma 6 we can guarantee a lower bound on the step-size $\gamma_t \geq \frac{1}{40L}$. Further, under our assumption on s , namely, $s \geq (240\bar{\kappa})^2 s^*$, we can bound the contraction factor:

$$\left(1 + \sqrt{\frac{s^*}{s}}\right)^2 (1 - \bar{\mu}\gamma_t) \leq \left(1 + \sqrt{\frac{s^*}{s}}\right) (1 - \bar{\mu}\gamma_t) \leq \left(1 + \frac{1}{80\bar{\kappa}}\right) \left(1 - \frac{1}{40\bar{\kappa}}\right) \leq 1 - \frac{1}{80\bar{\kappa}},$$

and therefore,

$$\|\theta_{t+1} - \hat{\theta}\|^2 \leq \left(1 - \frac{1}{80\bar{\kappa}}\right) \|\theta_t - \hat{\theta}\|^2.$$

Thus, the first part of the theorem's statement follows, i.e. when $\|\theta_t - \hat{\theta}\|$ is sufficiently large, we can guarantee that θ_{t+1} will approach $\hat{\theta}$. We are now left with establishing the veracity of the second statement. For this, let t_0 be the time defined in the theorem's statement. Then, we are under case **(ii)** or case **(iii)**. If we are under case **(iii)** there is nothing left to prove. If we are in case **(ii)**, there are two further cases: **(a)** the iterates remain confined within a ball of radius $6\|\text{HT}_s(\hat{g})\|^2/\bar{\mu}^2$, i.e.

$$\forall t \geq t_0, \|\theta_t - \hat{\theta}\|^2 < \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2},$$

and the theorem's second statement is therefore true, or **(b)** there exists a time $t_1 > t_0$ at which for the first time

$$\frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2} \leq \|\theta_{t_1} - \hat{\theta}\|^2.$$

By Lemma 4 and by definition of t_1 we have

$$\|\theta_{t_1} - \hat{\theta}\|^2 \leq \left(1 + \frac{1}{80\bar{\kappa}}\right) (1 - \bar{\mu}\gamma_{t_1}) \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2} < \left(1 + \frac{1}{80\bar{\kappa}}\right) \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2}.$$

This implies, we find ourselves again in case **(i)**. Observe that going forward, no iterate can escape the above ball and therefore the second statement of the theorem holds.

A.2 Proof of Corollary 1

If (3) holds for $\hat{\theta}$, for any θ_t fulfilling $\|\theta_t - \hat{\theta}\|^2 \leq \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2}$ there holds

$$\min_{i \in \mathcal{S}} |[\theta_t]_i| \geq \frac{7\|\text{HT}_s(\hat{g})\|}{\bar{\mu}} - \frac{6\|\text{HT}_s(\hat{g})\|}{\bar{\mu}} > 0. \quad (10)$$

This can be established by contradiction, i.e. if the condition above is violated, either $\|\theta_t - \hat{\theta}\|^2 > \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2}$ or (3) can not hold. Consequently, $\hat{\mathcal{S}} \subset \mathcal{S}^t$. Now we consider the iterate θ_{t+1} . Note that $\hat{\mathcal{S}} \subset \mathcal{S}^{t+1}$, if

$$\gamma_t \|g_t\|_\infty < \frac{\|\text{HT}_s(\hat{g})\|}{2\bar{\mu}}. \quad (11)$$

To see this note that for any $i \notin \mathcal{S}_t$ we have

$$|[\theta_t]_i - \gamma_t [g_t]_i| = \gamma_t |[g_t]_i| < \frac{\|\text{HT}_s(\hat{g})\|}{2\bar{\mu}},$$

and for any $i \in \hat{\mathcal{S}}$, given that $\|\theta_t - \hat{\theta}\| \leq \frac{6\|\text{HT}_s(\hat{g})\|}{\bar{\mu}}$, we have

$$|[\theta_t]_i - \gamma_t [g_t]_i| > \frac{7\|\text{HT}_s(\hat{g})\|}{\bar{\mu}} - \frac{6\|\text{HT}_s(\hat{g})\|}{2\bar{\mu}} - \frac{\|\text{HT}_s(\hat{g})\|}{2\bar{\mu}} = \frac{\|\text{HT}_s(\hat{g})\|}{2\bar{\mu}}.$$

Because the Hard Thresholding operator selects the s largest components of θ_{t+1} , in the selection of the elements that should go into $\mathcal{S}^{t+1}$ the operator will not deselect elements from $\hat{\mathcal{S}}$ in benefit of any outside of $\mathcal{S}^t$. Thus, $\hat{\mathcal{S}} \subset \mathcal{S}^{t+1}$.

We now find conditions on γ_t under which (11) can be guaranteed. Observe that

$$\begin{aligned} \gamma_t \|g_t\|_\infty &\leq \gamma_t (\|\hat{g}\|_\infty + \|g_t - \hat{g}\|_\infty) \\ &\stackrel{(i)}{\leq} \gamma_t \left(\|\hat{g}\|_\infty + \bar{L} \|\theta_t - \hat{\theta}\| \right) \stackrel{(ii)}{\leq} \gamma_t \left(\|\hat{g}\|_\infty + \frac{6\bar{L}}{\bar{\mu}} \|\text{HT}_s(\hat{g})\| \right), \end{aligned}$$

where (i) follows from Lemma 5, and (ii) follows from the assumption that $\|\theta_t - \hat{\theta}\|^2 \leq \frac{36\|\text{HT}_s(\hat{g})\|^2}{\bar{\mu}^2}$. A sufficient condition for (11) to hold is thus given by

$$\gamma_t < \frac{1}{2\bar{\mu} + 12\bar{L}}.$$

As a result, when $\gamma_t < \frac{1}{2\bar{\mu} + 12\bar{L}}$, we have $\hat{\mathcal{S}} \subset \mathcal{S}^{t+1}$, and thus $\|\theta_{t+1} - \hat{\theta}\|^2 \leq \|\theta_t - \hat{\theta}\|^2$ by $\|\theta_{t+1} - \hat{\theta}\|^2 = \|[\theta_{t+1} - \hat{\theta}]_{\hat{\mathcal{S}}^{t+1}}\|^2 \leq (1 - \bar{\mu}\gamma_t) \|\theta_t - \hat{\theta}\|^2$. Otherwise, $\gamma_t \geq \frac{1}{2\bar{\mu} + 12\bar{L}} > \frac{1}{40\bar{L}}$, we still have $\|\theta_{t+1} - \hat{\theta}\|^2 \leq \|\theta_t - \hat{\theta}\|^2$ by Lemma 4. Thus, the proof is completed by induction.

A.3 Proof of Corollary 2

By invoking Lemma 1, we can guarantee that f fulfills the RSC and RSM with probability at least $1 - e^{-c_0 n}$. The function f is convex by construction, and assuming that Algorithm 1 is provided suitable parameters $\hat{f}$ and s as stipulated by the Corollary, we may invoke Theorem 1. Thus, with probability at least $1 - e^{-c_0 n}$, the iterates satisfy a contractive relation for $\|\theta_t - \theta^*\|^2$ until the point where $\|\theta_t - \theta^*\|^2 < \frac{36\|\text{HT}_s(g^*)\|^2}{\bar{\mu}^2}$.

To complete the proof, it remains to establish that, with probability at least $1 - \frac{2}{d}$, it holds that

$$\frac{36\|\text{HT}_s(g^*)\|^2}{\bar{\mu}^2} \leq 72C^2 \frac{s \log d}{n\bar{\mu}^2}. \quad (12)$$

The proof of this claim essentially follows the arguments in (Wainwright, 2019, Example 7.14); for completeness, we include the full details here.

Define $\sigma(x) = \frac{1}{1 + \exp(-x)}$. The gradient of f evaluated at θ^* can be expressed as

$$g^* = \frac{1}{n} \sum_{i=1}^n (\sigma(x_i^\top \theta^*) - y_i) x_i.$$

Recall the relation between x_i and y_i is governed by the model

$$\Pr(y_i = 1 \mid x_i) = \frac{1}{1 + \exp(-x_i^\top \theta^*)}.$$

Under this model, it follows that each term $\sigma(x_i^\top \theta^*) - y_i$ is a zero-mean sub-Gaussian random variable with sub-Gaussian parameter $\sigma^2 = \frac{1}{4}$.

Thus, $\|g^*\|_\infty$ is the maximum of d independent zero-mean sub-Gaussian random variables, each with variance proxy at most $\sigma^2 = \frac{C^2}{4n}$. By standard sub-Gaussian maximal inequalities, we have

$$\Pr\left(\|g^*\|_\infty \geq C\sqrt{\frac{\log d}{2n}} + \frac{C\delta}{2}\right) \leq 2e^{-\frac{n\delta^2}{2}}$$

for any $\delta > 0$. Setting $\delta = \sqrt{\frac{2\log d}{n}}$ completes the proof of the bound in (12).

Applying the union bound allows us to claim that the guarantees provided by Theorem 1 hold with probability at least $1 - c_0 e^{-n} - \frac{2}{d}$. Guarantees on support recovery follow then by direct application of Corollary 1.

A.4 Proof to Corollary 3

By (Jain et al., 2014, Lemma 2), we can verify the result of Theorem 1 translates to the matrix case, where the vector ℓ_2 -norm is replaced by the Frobenius norm, the ℓ_1 -norm is replaced by the nuclear norm, and the HT_s operator is substituted by PM_s .

By applying Theorem 1 and Lemma 2, we conclude that, with probability at least $1 - e^{-c_0 n}$, the iterates satisfy a contractive relation for $\|\Theta_t - \Theta^*\|_F^2$ until the point where $\|\Theta_t - \Theta^*\|_F^2 < \frac{36\|\text{PM}_s(g^*)\|_F^2}{\bar{\mu}^2}$.

To complete the proof, it remains to establish that, with probability at least $1 - \frac{2}{d}$,

$$\frac{36\|\text{PM}_s(g^*)\|_F^2}{\bar{\mu}^2} \leq \frac{7200\sigma^2\zeta(\Sigma)sd}{n\bar{\mu}^2}. \quad (13)$$

Note that $g^* = \frac{1}{n}\varepsilon_i X_i$, by (Wainwright, 2019, Corollary 10.10), we have

$$\Pr\left(\|g^*\|_2 \geq \frac{\lambda_n}{2}\right) \leq 2e^{-2n\delta^2},$$

where $\lambda_n = 10\sigma\sqrt{\zeta(\Sigma)}\left(\sqrt{\frac{2d}{n}} + \delta\right)$.

By setting $\delta = \sqrt{\frac{2d}{n}}$, we have

$$\|g^*\|_2 \leq 10\sigma\sqrt{\zeta(\Sigma)}\sqrt{\frac{2d}{n}},$$

with probability at least $1 - 2e^{-4d}$. We thus apply the union bound to complete the proof of claim (13).

B Adaptive Lower Bound

B.1 Proof of Theorem 2

Let $a_t := \frac{f(\theta_t) - f(\theta^*)}{5\|\text{HT}_s(g_t)\|}$. Suppose the step size γ_t used in Algorithm 2 satisfies $\gamma_t = ba_t$ for some scalar $b \in [\frac{1}{2}, 1]$. By invoking Lemmas 4 and 6, as long as $\|\theta_t - \theta^*\|^2 \geq \frac{36\|\text{HT}_s(g^*)\|^2}{\bar{\mu}^2}$, we have:

$$\|\theta_{t+1} - \theta^*\|^2 \leq \left(1 + \frac{1}{160\bar{\kappa}}\right) \left(1 - \frac{b}{40\bar{\kappa}}\right) \|\theta_t - \theta^*\|^2 \leq \left(1 - \frac{1}{160\bar{\kappa}}\right) \|\theta_t - \theta^*\|^2.$$

This establishes that the iterates exhibit contractive behavior until they enter a small neighborhood of the optimum.

We now consider two possible cases depending on whether the lower bound surrogate $\tilde{f}_k$ is valid and how the step size compares to a_t during epoch k .

Case (i): Suppose $\tilde{f}_k$ is a valid lower bound for $f(\theta^*)$, and that $\gamma_t \leq a_t$ holds for all iterations $t = 0, \dots, T(\alpha)$ within epoch k . Then, the contractive relation applies repeatedly, and we obtain

$$\|\theta_{T(\alpha)} - \theta^*\|^2 \leq \left(1 - \frac{\bar{\mu}}{160\bar{L}}\right)^{T(\alpha)} \|\theta_0 - \theta^*\|^2 \leq (1 + \alpha) \left(1 + \frac{1}{160\bar{\kappa}}\right) \frac{36\|\text{HT}_s(g^*)\|^2}{\bar{\mu}^2}.$$

By the restricted smoothness and strong convexity assumptions, this implies that the function suboptimality satisfies $f(\theta_{T(\alpha)}) - f^* \leq \varepsilon(\alpha)$, thus completing the proof for this case.

Case (ii): Alternatively, suppose that $\tilde{f}_k$ is a valid lower bound, but $\gamma_t > a_t$ for some t in epoch k . This condition implies that

$$f(\theta_t) - \tilde{f}_k > 2(f(\theta_t) - f(\theta^*)),$$

which in turn yields

$$\tilde{f}_{k+1} = \frac{f(\bar{\theta}_k) + \tilde{f}_k}{2} \leq \frac{f(\theta_t) + \tilde{f}_k}{2} < f(\theta^*).$$

Hence, $\tilde{f}_{k+1}$ is also a valid lower bound. By induction, we conclude that if Case I never occurs, then all $\tilde{f}_k$, for $k = 1, \dots, K$, remain valid lower bounds.

Moreover, under this scenario, the sequence $f^* - \tilde{f}_k$ decreases geometrically. In particular,

$$f^* - \tilde{f}_{k+1} \leq f^* - \frac{f^* + \tilde{f}_k}{2} = \frac{f^* - \tilde{f}_k}{2}.$$

The geometric decrease of $f^* - \tilde{f}_k$ in case (ii) ensures that if case (i) never occurs, then there exists some k_0 such that $f(\theta^*) - \tilde{f}_{k_0} < \varepsilon(\alpha)$. In that case, $\bar{\theta}_{k_0}$ is either an output corresponding to case (i) (which completes the proof) or an output under case (ii), i.e., it satisfies

$$f(\bar{\theta}_{k_0}) - f(\theta^*) < f(\theta^*) - \tilde{f}_{k_0} < \varepsilon(\alpha).$$

C Other Statistical Guarantees

In this section we provide guarantees for additional statistical models not provided in the body of the paper. The results in Appendix C.1 hold for sparse linear regression. The results in Appendix C.2 apply to some GLMs and require the analysis of the behavior of Sparse Polyak under different regularity conditions. These are not entirely captured in Theorem 1.

C.1 Sparse Linear Regression

In this section, we assume the dataset consists of data points $z_i = (x_i, y_i)$ for $i = 1, \dots, n$, where $x_i \in \mathbb{R}^d$ denote the feature vectors, and $y_i \in \mathbb{R}$ denote the responses. The feature vectors are aggregated into the design matrix

$$\mathbb{R}^{n \times d} \ni X \triangleq \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}$$

and the responses are aggregated in $\mathbb{R}^n \ni y \triangleq (y_1, \dots, y_n)^\top$. Let θ^* denote the ground truth of the statistical model, with $\|\theta^*\|_0 \leq s^*$, and f^* denote the corresponding objective value. Specifically, we assume that the responses y_i and feature vectors x_i are related by $y_i = x_i^\top \theta^* + \varepsilon_i$, where x_i are drawn from a $N(0, \Sigma)$ distribution, Σ is non singular, $\varepsilon_i \sim N(0, \sigma^2)$, and x_i and ε_i are i.i.d and independent of one another. Additionally, the objective function $f(\theta) = \frac{1}{2n} \|X\theta - y\|^2$.

From Agarwal et al. (2012)[Lemma 6], it follows that the RSS and RSC conditions hold with probability at least $1 - e^{-c_0 n}$ with coefficients $L = 2\sigma_{\max}(\Sigma)$, $\mu = \frac{1}{2}\sigma_{\min}(\Sigma)$, and $\tau = c_1 \zeta(\Sigma) \frac{\log d}{n}$, where $\zeta(\Sigma) = \max_{i=1, \dots, d} \Sigma_{ii}$. Here c_0 and c_1 are universal constants.

Corollary 4. *Consider the sparse linear regression problem described above. Let $\{\theta_t\}_{t \geq 0}$ be the sequence of iterates generated by Algorithm 1 or Algorithm 2 when employed to solve a sparse linear*

regression problem. Suppose that $\hat{f} = f^*$ for Algorithm 1. Assume we have sufficient samples for $\bar{\mu} > 0$, with $s \geq (240\bar{\kappa})^2 s^*$ for Algorithm 1, and $s \geq (480\bar{\kappa})^2 s^*$ for Algorithm 2. Further, assume that each column of X is C -normalized, i.e., $\| \frac{X_j}{\sqrt{n}} \| \leq C$ for $j = 1, \dots, d$. Here X_j denotes the j -th column of X . Then, with probability at least $1 - e^{-c_0 n} - \frac{2}{d}$, for any $\alpha \geq \frac{1}{80}$ after $T(\alpha)$ iterations, $\min_{t \leq T(\alpha)} \|\theta_t - \theta^*\|^2$ is upper bounded by

$$\varepsilon(\alpha) = (1 + \alpha) \frac{288C^2 \sigma^2 s \log d}{n\bar{\mu}^2}.$$

The required number of iterations $T(\alpha)$ fulfills:

$$\begin{aligned} T(\alpha) &\leq \left\lceil \frac{1}{\log(1/(1 - 1/80\bar{\kappa}))} \log \left(\frac{\|\theta_0 - \theta^*\|^2}{\varepsilon(\alpha)} \right) \right\rceil, \text{ and} \\ T(\alpha) &\leq \left(1 + \log_2 \frac{4(f(\theta_0) - f(\theta^*))}{\bar{\mu}\varepsilon(\alpha)} \right) \left\lceil \frac{1}{\log(1/(1 - 1/160\bar{\kappa}))} \log \left(\frac{\|\theta_0 - \theta^*\|^2}{\varepsilon(\alpha)} \right) \right\rceil, \end{aligned}$$

for Algorithm 1 and Algorithm 2 respectively.

Moreover, if θ^* satisfies the SNR condition (3), we can ensure that after $T(0)$ iterations, the error is upper bounded by $\varepsilon(0)$, and the support of θ^* has been identified, i.e. $\mathcal{S}^* \subset \mathcal{S}^t \forall t \geq T(0)$.

Corollary 4 establishes the convergence properties of Algorithm 1 and Algorithm 2. The error term is of order $\mathcal{O}\left(\frac{\bar{\kappa}^2 s^* \log d}{n\bar{\mu}^2}\right)$, which is of the same order as that in (Jain et al., 2014, Theorem 3), where a fixed step size is considered under the assumption that $\bar{L}$ is known.

Proof. For Algorithm 1, the proof follows the same steps as the proof of Corollary 2. The only difference is the upper bound for $\frac{36\|\text{HT}_s(g^*)\|^2}{\bar{\mu}^2}$, which we provide next.

When the columns of X are C -normalized, by (Wainwright, 2019, Example 7.14), with probability $1 - \frac{2}{d}$,

$$\|g^*\|_\infty^2 = \|X^T \varepsilon\|_\infty^2 \leq 8C^2 \sigma^2 \frac{\log d}{n}.$$

Using the union bound together with (Agarwal et al., 2012, Lemma 6), and the assumptions stated in the Corollary, yield the required assumptions for Theorem 1 to hold. Thus, this completes the proof for Algorithm 1.

As we can see from the proof of Theorem 1 and Theorem 2, the accuracy level $\varepsilon(0)$ is determined by the point at which a lower bound on the step size can be established. According to the formulation of γ_t in Algorithm 2, it is guaranteed to be at least half the step size used in Algorithm 1. This observation implies that the accuracy level $\varepsilon(0)$ can also be achieved by Algorithm 2 when the conditions of Case (i) in the proof of Theorem 2 are satisfied.

To complete the proof for Algorithm 2, we only need to show that $f(\theta_t) - f^* \leq \frac{18\|\text{HT}_s(g^*)\|^2}{\bar{\mu}}$ implies $\|\theta_t - \theta^*\|^2 \leq \varepsilon(0)$.

By RSC,

$$f(\theta_t) - f^* \geq \bar{\mu} \|\theta_t - \theta^*\|^2 - \|\text{HT}_s(g^*)\| \|\theta_t - \theta^*\|.$$

When $f(\theta_t) - f^* \leq \frac{18\|\text{HT}_s(g^*)\|^2}{\bar{\mu}}$, it implies

$$\frac{\|\text{HT}_s(g^*)\|_s}{\bar{\mu}} \left(\frac{3\|\text{HT}_s(g^*)\|}{\bar{\mu}} + \|\theta_t - \theta^*\| \right) \geq \|\theta_t - \theta^*\|^2.$$

A necessary condition for the inequality above is

$$\|\theta_t - \theta^*\|^2 \leq \frac{36\|\text{HT}_s(g^*)\|_s^2}{\bar{\mu}^2} = \varepsilon(0).$$

□

C.2 Generalized Linear Models

In this section, we consider a dataset consisting of observations $z_i = (x_i, y_i)$ for $i = 1, \dots, n$, where $x_i \in \mathbb{R}^d$ denote the feature vectors, and $y_i \in \mathbb{R}$ denotes the corresponding responses. The feature vectors are organized into the design matrix

$$X \triangleq \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d},$$

and the responses are collected in the vector $y \triangleq (y_1, \dots, y_n)^\top \in \mathbb{R}^n$.

For notational convenience, we let θ^* denote the true underlying parameter of the statistical model and f^* the corresponding objective function value. We assume the relationship between x_i and y_i is characterized by the conditional distribution

$$\Pr(y_i \mid x_i, \theta^*, \sigma) = \exp \left\{ \frac{y_i x_i^\top \theta^* - \psi(x_i^\top \theta^*)}{c(\sigma)} \right\},$$

where $\sigma > 0$ is a scale parameter, and ψ is the cumulant function. Given this data generation model, we define the objective function

$$f(\theta) = \frac{1}{n} \sum_{i=1}^n (\psi(x_i^\top \theta) - y_i x_i^\top \theta).$$

We assume that ψ is infinitely differentiable with $\psi''(t) > 0$ and uniformly bounded for all $t \in \mathbb{R}$. These assumptions are satisfied in a variety of settings, including logistic regression and multinomial regression Loh and Wainwright (2015). We assume the feature vectors x_i are i.i.d. and drawn from a multivariate normal distribution $N(0, \Sigma)$, where Σ is non-singular. Under the setting described above, the RSC condition does not hold everywhere. However, in the described setting it can be shown that the following milder RSC condition holds Loh and Wainwright (2015)

$$f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq \begin{cases} \frac{\mu}{2} \|y - x\|_2^2 - \frac{\tau}{2} \|y - x\|_1^2, & \text{if } \|y - x\|_2 \leq 1, \quad (14a) \\ \|y - x\|_2 \left(\frac{\mu}{2} - \frac{\tau}{2} \frac{\|y - x\|_1^2}{\|y - x\|_2^2} \right), & \text{if } \|y - x\|_2 > 1. \quad (14b) \end{cases}$$

We now present a lemma grouping the results that we need to proceed: the RSS condition, the condition (14), and the order of achievable statistical accuracy for GLMs in our setting. This lemma aggregates results from (Loh and Wainwright, 2015, Proof of Corollary 2, Appendix D.1) and (Negahban et al., 2012, Proposition 2). Notably, while (Negahban et al., 2012, Proposition 2) is originally stated centered only around the ground truth $x = \theta^*$ (c.f. (14)) its proof extends to any given x . This implies that while our results are currently stated for $x = \theta^*$, to achieve optimal statistical precision, we can instead state equivalent results to those in Theorem 1 where $\hat{\theta}$ is such that $f(\hat{\theta}) = \hat{f}$.

Lemma 7. *For the statistical models described above, with probability at least $1 - c_1 d^{-1} - c_2 e^{-n}$ Assumption 2 and (14) hold, and*

$$\|\nabla f(\theta^*)\|_\infty \leq c_0 \sqrt{\frac{\log d}{n}},$$

where $c_0, c_1, c_2 > 0$ are universal constants. The constants μ , and L in (14) and Assumption 2, respectively, depend on ψ , and Σ . Further $\tau = c_3 \frac{\log d}{n}$ where $c_3 > 0$ is a universal constant.

In the following, we keep the definition $\bar{L} = L + 3\tau s$, and $\bar{\mu} = \mu - 3\tau s$, where μ is that of (14). Observe that Corollary 5 is stated for Algorithm 1 for simplicity but an analogous statement for Algorithm 2 follows.

Corollary 5. Let $\{\theta_t\}_{t \geq 0}$ denote the iterates generated by Algorithm 1 when applied to the generalized linear models described above. Set the step size rule according to

$$\gamma_t = \frac{\max\{f(\theta_t) - \hat{f}, 0\}}{5\|\text{HT}_{2s}(g_t)\|^2}.$$

Define $R := \|\theta^*\|^2$, and $R_0 := 4R + 1$. Assume we set $\hat{f} = f^*$, $\theta_0 = 0$, and suppose $s \geq (480R_0\bar{\kappa}^2)^2 s^*$. Further, assume the sample size is large enough to ensure $\bar{\mu} > 5c_0\sqrt{\frac{2s \log d}{n}}$.

Then, with probability at least $1 - \frac{c_1}{d} - c_2e^{-n}$, we guarantee that for all $t \geq T$ where

$$T \leq \mathcal{O}\left(\bar{\kappa} \log\left(\frac{n}{s \log(d)}\right)\right) + \mathcal{O}(\bar{\kappa}^2 \log(R))$$

there holds

$$\|\theta_t - \theta^*\|^2 \leq \left(1 + \frac{1}{160R_0\bar{\kappa}^2}\right) \frac{36c_0s \log d}{n\bar{\mu}^2}.$$

Moreover, if θ^* satisfies the SNR condition (3), we can guarantee that for all $t \geq T$, $\|\theta_t - \theta^*\|^2 \leq \frac{36c_0s \log d}{n}$ and that the support of θ^* has been recovered, i.e. $\mathcal{S}^* \subset \mathcal{S}_t$.

Proof. To establish this result we leverage that the condition (14) combined with Sparse Polyak allow us to establish convergence despite the lack of RSC. More specifically, we will establish that Algorithm 1 exhibits, under the conditions described in the Corollary, at most three modes of convergence:

$$\|\theta_{t+1} - \theta^*\|^2 \leq \begin{cases} \left(1 - \frac{1}{160R_0\bar{\kappa}^2}\right) \|\theta_t - \theta^*\|^2, & \text{if } \|\theta_t - \theta^*\| \geq 1, \\ \left(1 - \frac{1}{160\bar{\kappa}}\right) \|\theta_t - \theta^*\|^2, & \text{if } \|\theta_t - \theta^*\| < 1, \text{ and } \|\theta_t - \theta^*\|^2 \geq \frac{36c_0^2s \log d}{n\bar{\mu}^2}, \\ \left(1 + \frac{1}{160R_0\bar{\kappa}^2}\right) \frac{36c_0^2s \log d}{n\bar{\mu}^2}, & \text{otherwise.} \end{cases} \quad (15)$$

Observe that under our current assumptions $\frac{36c_0^2s \log d}{n\bar{\mu}^2} < \frac{18}{25}$ and therefore, the list above is exhaustive and the conditions on the second case are compatible.

We begin assuming that $R \geq 1$. We will establish that if the modes of convergence provided above hold for $R \geq 1$, when $R < 1$ we only observe the two last cases.

(i) We start off by exploiting (14). Assuming that $\|\theta_t - \theta^*\| \geq 1$ we may follow the strategy in Lemma 4 and exploit that f is convex (but not RSC) to obtain

$$\|\theta_{t+1} - \theta^*\|^2 \leq \left(1 + \sqrt{\frac{s^*}{s}}\right)^2 \left(\|\theta_t - \theta^*\|^2 - 2\gamma_t(f(\theta_t) - f(\theta^*)) + 5\gamma_t^2\|\text{HT}_{2s}(g_t)\|^2\right). \quad (16)$$

With the choice of step size

$$\gamma_t = \frac{\max\{f(\theta_t) - f(\theta^*), 0\}}{5\|\text{HT}_{2s}(g_t)\|^2},$$

(16) simplifies to:

$$\|\theta_{t+1} - \theta^*\|^2 \leq \left(1 + \sqrt{\frac{s^*}{s}}\right)^2 \left(\|\theta_t - \theta^*\|^2 - \frac{(f(\theta_t) - f(\theta^*))^2}{5\|\text{HT}_{2s}(g_t)\|^2}\right). \quad (17)$$

From Lemma 7 and under the assumption that $\bar{\mu} > 5c_0\sqrt{\frac{2s \log d}{n}}$, it follows that with probability at least $1 - c_1d^{-1} - c_2e^{-n}$ there holds

$$f(\theta_t) - f(\theta^*) \geq \frac{\bar{\mu}}{2}\|\theta_t - \theta^*\| - \|\text{HT}_{2s}(g^*)\|\|\theta_t - \theta^*\| \geq \frac{\bar{\mu}}{4}\|\theta_t - \theta^*\|,$$

thus by applying the above bound onto (17) we have

$$\|\theta_{t+1} - \theta^*\|^2 \leq \left(1 + \sqrt{\frac{s^*}{s}}\right)^2 \left(\|\theta_t - \theta^*\|^2 - \frac{\bar{\mu}^2}{80\|\text{HT}_{2s}(g_t)\|^2} \|\theta_t - \theta^*\|^2\right). \quad (18)$$

Further, we may upper bound the norm of the gradient as

$$\begin{aligned} \|\text{HT}_{2s}(g_t)\|^2 &\leq 2(\|\text{HT}_{2s}(g^*)\|_{2s}^2 + \|\text{HT}_{2s}(g_t - g^*)\|^2) \\ &\stackrel{(i)}{\leq} 2(\|\text{HT}_{2s}(g^*)\|_{2s}^2 + 2\bar{L}^2\|\theta_t - \theta^*\|^2) \end{aligned} \quad (19)$$

where in (i) we invoke Lemma 5. Combining (18) with (19) yields

$$\|\theta_{t+1} - \theta^*\|^2 \leq \left(1 + \sqrt{\frac{s^*}{s}}\right)^2 \left(1 - \frac{\bar{\mu}^2}{160(\|\text{HT}_{2s}(g^*)\|^2 + 2\bar{L}^2\|\theta_t - \theta^*\|^2)}\right) \|\theta_t - \theta^*\|^2. \quad (20)$$

To guarantee that the above implies our first regime of convergence we need to establish that $\|\theta_t - \theta^*\|^2 \leq R$ for all $t \geq 0$. We return to this point after exploring the second and third regime which will be useful in establishing the first.

(ii) On the other hand, assume instead that $\|\theta_t - \theta^*\|^2 \leq 1$ and $\|\theta_t - \theta^*\|^2 \geq \frac{36c_0^2 s \log d}{n\bar{\mu}^2}$ then, from (14) the RSC holds and therefore we can invoke the result of Theorem 1 in which

$$\|\theta_{t+1} - \theta^*\|^2 \leq \left(1 - \frac{1}{160\bar{\kappa}}\right) \|\theta_t - \theta^*\|^2. \quad (21)$$

(iii) If, instead $\|\theta_t - \theta^*\|^2 < \frac{36c_0^2 s \log d}{n\bar{\mu}^2}$ we have from Theorem 1 that

$$\|\theta_{t+1} - \theta^*\|^2 \leq \left(1 + \frac{1}{160\bar{\kappa}}\right) \frac{36c_0^2 s \log d}{n\bar{\mu}^2} < 1. \quad (22)$$

Observe that from the three cases we consider ((20)- (22)), (21) and (22) already correspond to one of our stated modes of convergence, and thus we are to establish the first.

Clearly, when $R < 1$, θ_0 is in a region in which the RSC holds, and therefore, we will only observe the behavior in (21) and (22). Thus, we are only to prove the first regime of convergence for $R \geq 1$. To establish that the behavior in the first regime holds, we need to establish that $\|\theta_t - \theta^*\|^2 \leq R$ holds for all $t \geq 0$. We proceed to establish this and consequently the behavior in the first regime by induction. Note that by our initial condition $\|\theta_0 - \theta^*\|^2 = \|\theta^*\|^2 = R$ and thus the condition holds for $t = 0$. Suppose $\|\theta_t - \theta^*\|^2 \leq R$ for some t . If $\|\theta_t - \theta^*\|^2 < 1$ then either (21) and (22) hold and the proof by induction is complete. If instead, $\|\theta_t - \theta^*\|^2 \geq 1$, (20) holds. Then, by induction hypothesis and under the Corollary's assumption on the sample size there holds

$$\|\text{HT}_{2s}(g^*)\|^2 + 2\bar{L}^2\|\theta_t - \theta^*\|^2 \leq \left(\frac{1}{16} + 2R\right) \bar{L}^2 \leq \frac{R_0 \bar{L}^2}{2},$$

thus, combining with (20) and using that by assumption $s \geq (480R_0\bar{\kappa}^2)^2$ we obtain

$$\|\theta_{t+1} - \theta^*\|^2 \leq \left(1 - \frac{1}{160R_0\bar{\kappa}^2}\right) \|\theta_t - \theta^*\|^2 < R.$$

We have thus established the veracity of (15). Observe that (15) together with $\theta_0 = 0$ and $\|\theta^*\|^2 = R$ imply that there exists $t_0 \geq 0$ fulfilling

$$t_0 \leq \lceil -\log(R)/\log(1 - 1/160R_0\bar{\kappa}^2) \rceil$$

such that for all $t \geq t_0$

$$\|\theta_t - \theta^*\|^2 < 1.$$

Further, this implies that in at most

$$\left\lceil \log \left(\frac{36c_0^2 s (1 + 1/(160\bar{\kappa})) \log(d)}{n\bar{\mu}^2} \right) / \log(1 - 1/160\bar{\kappa}) \right\rceil$$

additional iterations optimal statistical precision is reached. Finally, if the SNR condition holds, the term $(1 + \frac{1}{160\bar{\kappa}})$ in the third regime is replaced by 1 as a consequence of Corollary 1. $\square$

Corollary 5 recovers the result in (Loh and Wainwright, 2015, Theorem 3) with the following similarities and differences. To achieve optimal statistical precision, as $\alpha = s \frac{\log d}{n} \rightarrow 0$ both Sparse Polyak and (Loh and Wainwright, 2015, Theorem 3) require $\mathcal{O}(\bar{\kappa} \log(\alpha^{-1}))$ iterations. However, we require additional iterations $\mathcal{O}(\bar{\kappa}^2 \log(R))$ when $R \geq 1$. We observe however, that our result holds under more general conditions, as we do not make the assumption that $\|\theta^*\| \leq 1$ which is necessary in (Loh and Wainwright, 2015, Theorem 3), where this condition can be relaxed at the expense of requiring the RSC to hold within a larger radius.

We conclude this Appendix by highlighting the fact that for both sparse linear regression and sparse GLMs we verify the rate invariance of Sparse Polyak theoretically. When the problem size increases much faster than the sample size $\frac{d}{n} \rightarrow \infty$ but $\frac{s^* \log d}{n}$ and Σ remain constant, IHT with Sparse Polyak will reach a ε neighborhood of the optimal statistical precision within at most $\mathcal{O}(\bar{\kappa}^{-1} \log(1/\varepsilon))$ for linear regression and at most $\mathcal{O}(\bar{\kappa}^{-1} \log(1/\varepsilon)) + \mathcal{O}(\bar{\kappa}^{-1} \log(R))$ when $R > 1$ in the case of GLMs. In both cases, this number does not change with increasing d and n . Observe that these results allow us to answer in the affirmative questions (ii) and (iii) posed in Section 1.

D Experiments on real data

Linear Regression We consider a linear regression task using the Large-scale Wave Energy Farm dataset from the UCI Machine Learning Repository Neshat et al. (2020), which is publicly available under the CC BY 4.0 license. The terms of use are described at <https://archive.ics.uci.edu/#terms>. The goal is to predict the total power output of the wave farm based on a sparse linear model. We randomly select 120 samples from the dataset, each containing 149 features. In our experiment, we set the sparsity level to $s = 20$. For the IHT method with a fixed step size, we choose the step size as 8×10^{-12} . This value is determined via a grid search over the range $[10^{-13}, 9 \times 10^{-12}]$, as step sizes outside this interval result in poor convergence or divergence. The results are presented in Figure 3 (left).

Logistic Regression We evaluate sparse logistic regression using the Molecule Musk dataset Chapman and Jain (1994) from the UCI Machine Learning Repository, which is publicly available under the CC BY 4.0 license. The terms of use are described at <https://archive.ics.uci.edu/#terms>. The task is to classify molecules as musks or non-musks. We randomly select 120 samples from the dataset, each with 166 features. In our experiment, we set the sparsity level to $s = 20$. For the IHT method with a fixed step size, we select a step size of 1.9×10^{-5} , chosen via a grid search over the interval $[3 \times 10^{-6}, 4 \times 10^{-5}]$. Step sizes outside this range lead to poor convergence or divergence. The results are shown in Figure 3 (right).

We observe that in Fig 3 (right) Sparse Polyak performs better than both classic Polyak and IHT with the fixed step-size, even if the step-size is optimized by grid search. This is expected, an adaptive step-size can adapt to the curvature at any point in the algorithm’s trajectory, whereas a fixed step-size cannot. We observe that Sparse Polyak performs better than Classic Polyak. On the other hand, in Fig 3 (left) we observe that the best fixed step-size performs better than an adaptive step-size. We conjecture that this is due to the factor $1/5$ in Sparse Polyak which we presume is an artifact of our analysis and is not in fact necessary. Despite this fact, sparse Polyak consistently outperforms the classical Polyak rule in the high-dimensional setting.

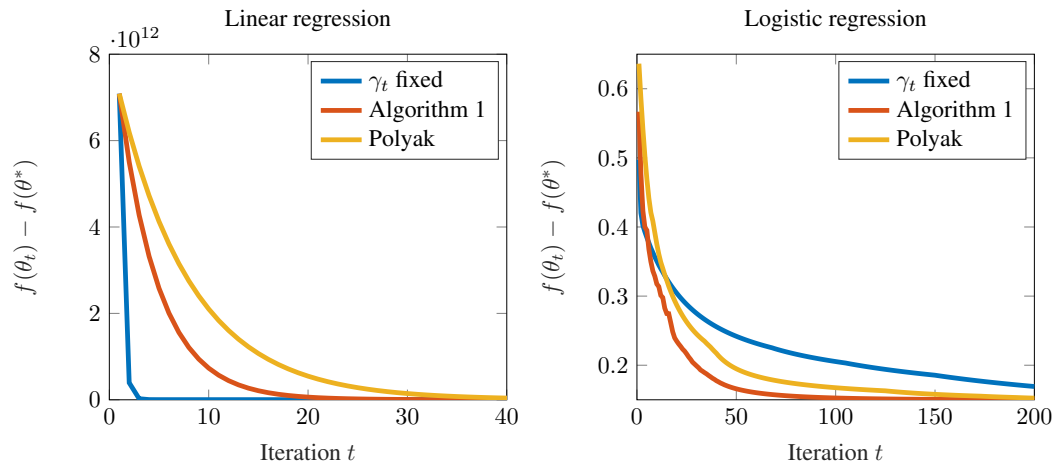


Figure 3: Performance comparison of IHT with optimal constant step size, Sparse Polyak and classical Polyak when performing: **(left)** linear regression on the Wave Energy Farm data set, and **(right)** logistic regression on the Molecule Musk data set.