# Between hard and soft thresholding: optimal iterative thresholding algorithms

Haoyang Liu and Rina Foygel Barber

June 11, 2018

**Abstract**

Iterative thresholding algorithms seek to optimize a differentiable objective function over a sparsity or rank constraint by alternating between gradient steps that reduce the objective, and thresholding steps that enforce the constraint. This work examines the choice of the thresholding operator, and asks whether it is possible to achieve stronger guarantees than what is possible with hard thresholding. We develop the notion of relative concavity of a thresholding operator, a quantity that characterizes the convergence performance of any thresholding operator on the target optimization problem. Surprisingly, we find that commonly used thresholding operators, such as hard thresholding and soft thresholding, are suboptimal in terms of convergence guarantees. Instead, a general class of thresholding operators, lying between hard thresholding and soft thresholding, is shown to be optimal with the strongest possible convergence guarantee among all thresholding operators. Examples of this general class includes $\ell_q$ thresholding with appropriate choices of $q$, and a newly defined *reciprocal thresholding* operator. We also investigate the implications of the improved optimization guarantee in the statistical setting of sparse linear regression, and show that this new class of thresholding operators attain the optimal rate for computationally efficient estimators, matching the Lasso.

## 1 Introduction

We consider the general problem of sparse optimization, where we seek to optimize a likelihood function or loss function subject to a sparsity constraint,

$$\min_{x\in\mathbb{R}^d, \|x\|_0\le s} \mathsf{f}(x).$$

Here $\mathsf{f}:\mathbb{R}^d\to\mathbb{R}$ is the target function that we would like to minimize, while the constraint $\|x\|_0\le s$ requires that the solution vector $x$ has at most $s$ many nonzero entries. Similarly, we may work with a matrix parameter $X\in\mathbb{R}^{n\times m}$ and search for a low-rank solution,

$$\min_{X\in\mathbb{R}^{n\times m}, \mathrm{rank}(X)\le s} \mathsf{f}(X).$$

Optimization problems over a sparsity constraint or a rank constraint are ubiquitous in high-dimensional statistics and machine learning. Sparsity of a vector parameter $x$ represents the idea that we can model the data using a small fraction of the available features, which, for instance, may correspond to covariates in a regression model or to basis expansion terms in a nonparametric function estimation problem. Similarly, a rank constraint on a matrix parameter $X$ might correspond to an underlying factor model with a small number of factors. We will focus on problems where $\mathsf{f}$ is a differentiable function, as is often the case for many likelihood models and other loss functions.

In this work, we will study the iterative thresholding approach, where gradient steps that lower the value of the target function $\mathsf{f}$ are alternated with thresholding steps to enforce the sparsity constraint—for instance, *hard thresholding* sets all but the largest $s$ entries to zero, while *soft thresholding* shrinks all values towards zero equally until the sparsity constraint is satisfied. (The same ideas apply to a rank constraint, by thresholding or shrinking singular values instead of vector entries. For simplicity, we will primarily discuss the sparse minimization problem, and will return to the low-rank problem later on.)

For sparse minimization of a differentiable target function $\mathsf{f}(x)$, many existing algorithms can be broadly described as iterating steps of the following form:

$$\begin{cases} \text{Gradient step: } x'_t = x_{t-1} - \eta_t \cdot \nabla \mathsf{f}(x_{t-1}) \text{ for some step size } \eta_t, \\ \text{Sparsity step: } x_t = \text{some sparse (or nearly sparse) approximation to } x'_t. \end{cases} \tag{1}$$

Our aim in this work is to characterize the type of thresholding operators that are likely to be most successful at converging to a good solution, i.e. to a value of $\mathsf{f}(x)$ that is as low as possible. Is an iterative thresholding algorithm most likely to succeed if we use hard thresholding, soft thresholding, or yet another form of thresholding to enforce the sparsity constraint?

In this work, we find that a success of a thresholding operator, relative to a broad class of target functions $\mathsf{f}$ that we may want to minimize, is fully characterized by a simple measure that we call the *relative concavity*. The relative concavity studies the behavior of the sparse thresholding map $x'_t \mapsto x_t$ in the iterative algorithm (1), viewed as an approximate projection onto the space of $s$-sparse vectors. Using relative concavity as a tool to evaluate and compare different thresholding operators, we find that commonly used thresholding operators, for example hard thresholding and soft thresholding, are indeed suboptimal. Instead, we characterize a general class of thresholding operators, lying between hard thresholding and soft thresholding, that we show to be optimal. This class includes $\ell_q$ norm thresholding, where $q \in (0, 1)$ is chosen adaptively relative to the particular problem; furthermore, choosing $q = 2/3$ is "universal" in the sense that it is nearly optimal across all sparse thresholding problems. We also develop the *reciprocal thresholding* operator, which enjoys the same optimality guarantees as $\ell_q$ thresholding, but with a closed-form equation for the iterative thresholding step. These simple and efficient iterative thresholding methods are then applied to the statistical setting of sparse linear regression problem:

$$y = X\theta_0 + z, \tag{2}$$

and are shown to match the Lasso in terms of the resulting guarantee on estimating the true mean vector $X\theta_0$.

# 2 Background: sparse minimization

Before defining relative concavity and the reciprocal thresholding operator, we first review some of the recent literature on hard thresholding and related methods, and define the convexity and smoothness properties of the objective function $\mathsf{f}$ that we will assume throughout this work.

## 2.1 Restricted strong convexity and restricted smoothness

In many problems in high-dimensional statistics, we aim to optimize loss functions that may be very poorly conditioned in general, but nonetheless exhibit convergence properties of a well-conditioned function when working only with sparse or approximately sparse vectors. This behavior is captured in the notions of restricted strong convexity and restricted smoothness (see e.g. Negahban et al. [2009], Loh and Wainwright [2013] for background).

A differentiable function $\mathsf{f} : \mathbb{R}^d \to \mathbb{R}$ satisfies *restricted strong convexity* with parameter $\alpha$ at sparsity level $s$, abbreviated as $(\alpha, s)$-RSC, if

$$\mathsf{f}(y) \geq \mathsf{f}(x) + \langle \nabla \mathsf{f}(x), y - x \rangle + \frac{\alpha}{2}\|x - y\|_2^2 \text{ for all } x, y \in \mathbb{R}^d \text{ with } \|x\|_0 \leq s, \|y\|_0 \leq s.$$

Similarly, $\mathsf{f}$ satisfies *restricted smoothness* with parameter $\beta$ at sparsity level $s$, abbreviated as $(\beta, s)$-RSM, if

$$\mathsf{f}(y) \leq \mathsf{f}(x) + \langle \nabla \mathsf{f}(x), y - x \rangle + \frac{\beta}{2}\|x - y\|_2^2 \text{ for all } x, y \in \mathbb{R}^d \text{ with } \|x\|_0 \leq s, \|y\|_0 \leq s.$$

Our results will focus on $\kappa = \beta/\alpha$, the *condition number* of the function $\mathsf{f}$ (at the given sparsity level $s$).

## 2.2 Iterative hard thresholding

Recent work by Jain et al. [2014] studies the *iterative hard thresholding algorithm*, which alternates between taking a gradient step, $x - \eta \nabla \mathsf{f}(x)$, and projecting onto the sparsity constraint. Specifically, given a target sparsity level $s$ and an initial point $x_0 \in \mathbb{R}^d$, the iterative step of the algorithm is defined by

$$x_t = \Psi_s^{\mathrm{HT}}\big(x_{t-1} - \eta \nabla \mathsf{f}(x_{t-1})\big), \tag{3}$$

where $\Psi_s^{\mathrm{HT}}$ is the "hard thresholding" operator, which truncates any vector $z \in \mathbb{R}^d$ to its $s$ largest entries,

$$\big(\Psi_s^{\mathrm{HT}}(z)\big)_i = \begin{cases} z_i, & i \in S, \\ 0, & i \notin S, \end{cases}$$

where $S \subset \{1, \ldots, d\}$ indexes the $s$ largest-magnitude entries of $z$.[1]

**Restricted optimality for iterative hard thresholding** It is well known that, due to the nonconvexity of the sparsity constraint $\|x\|_0 \leq s$, the iterative hard thresholding algorithm cannot be guaranteed to find the global minimum, $\min_{\|x\|_0 \leq s} \mathsf{f}(x)$—at least, not without strong assumptions. In other words, it may be the case that $\lim_{t\to\infty} \mathsf{f}(x_t)$ is strictly larger than $\min_{\|x\|_0 \leq s} \mathsf{f}(x)$. However, Jain et al. [2014]'s analysis of the iterative hard thresholding algorithm (3) proves that IHT achieves a weaker optimization guarantee, converging to a loss value that is at least as small as the best value attained under a more restricted constraint $\|x\|_0 \leq s'$ where $s' < s$. More precisely, Jain et al. [2014, Theorem 1] prove that, for an objective function $\mathsf{f}$ satisfying $(\alpha, s)$-RSC and $(\beta, s)$-RSM,

$$\mathsf{f}(x_t) \leq \min_{\|y\|_0 \leq s/(32\kappa^2)} \left\{ \mathsf{f}(y) + \left(1 - \frac{1}{12\kappa}\right)^t \cdot (\mathsf{f}(x_0) - \mathsf{f}(y)) \right\}, \tag{4}$$

where $\kappa = \beta/\alpha$, and where the step size is taken to be $\eta \propto 1/\beta$. In other words, their result proves linear convergence to the bound

$$\lim_{t\to\infty} \mathsf{f}(x_t) \leq \min_{\|y\|_0 \leq s/(32\kappa^2)} \mathsf{f}(y),$$

[1] To be fully precise, in the case of a tie between different entries of $z$, we may need to choose which entries to keep and which to set to zero. This choice will not matter from the point of view of our theoretical analysis, and from this point on, we will assume that we have fixed some map $z \mapsto S$, mapping each vector $z \in \mathbb{R}^d$ to a set $S \subset \{1, \ldots, d\}$ corresponding to the indices of the $s$ largest entries, so that $|S| = s$ and $\min_{i \in S} |z_i| \geq \max_{j \notin S} |z_j|$, for every $z$. For instance, in the case of a tie between $z_i$ and $z_j$ for the position of the $s$th largest-magnitude entry, we might follow the rule that we choose to keep entry $i$ if $i < j$ and to keep entry $j$ otherwise. Since the exact choice of the rule for breaking ties is not relevant for our results here, we will implicitly assume it to be fixed for the remainder of this paper.

meaning that while IHT may not find the global minimum of $\mathsf{f}(x)$ relative to the $s$-sparsity constraint, it is nonetheless guaranteed to perform at least as well as the best $s/(32\kappa^2)$-sparse solution. An analogous result is proved for the low-rank setting, thresholding singular values instead of vector entries.

In this work, we will refer to this type of result as a *restricted optimality* guarantee, where the output of an $s$-sparse optimization algorithm is guaranteed to perform well relative to a more restrictive $s'$-sparsity constraint, for some $s' < s$. In particular, we will be interested in the sparsity ratio $s'/s$—the ratio between the sparsity level $s$ used in the algorithm, versus the level $s'$ appearing in the guarantee. Ideally, we would like this ratio to be as close to 1 as possible, for the strongest possible guarantee.

## 2.3 Related literature

**Iterative thresholding** There exists a vast literature on the properties of iterative thresholding algorithms, especially iterative hard thresholding, regarding the optimization properties and statistical guarantees of these algorithms. Recent results in this area include the work of Blumensath and Davies [2009], Jain et al. [2014], Chen and Wainwright [2015], Bhatia et al. [2015], Jain et al. [2016], Cai et al. [2016], Kyrillidis and Cevher [2014].

Accelerated forms of the iterative hard thresholding algorithm are studied in Kyrillidis and Cevher [2011], Blumensath [2012], Khanna and Kyrillidis [2017]. In particular, Khanna and Kyrillidis [2017] finds substantial theoretical and empirical improvement over the original non-accelerated version of the algorithm. Nguyen et al. [2017] studies iterative hard thresholding in the context of stochastic gradient descent, where at each step $t$ we only have access to a noisy vector that approximates the true current gradient, $\nabla\mathsf{f}(x_t)$. The works mentioned here also consider thresholding algorithms for the low-rank setting, truncating singular values instead of vector entries. More broadly, Nguyen et al. [2017]'s work considers approximate thresholding procedures and more general definitions of sparsity.

To the best of our knowledge, the question of optimality among thresholding operators has not been addressed before, and it is the goal of this work to provide a framework to identify the convergence behavior of *all* thresholding operators and to find the optimal ones.

**Penalized and constrained optimization methods** The sparse optimization problem can alternately be approximated by a penalized minimization problem,

$$\min_{x\in\mathbb{R}^d} \left\{\mathsf{f}(x) + \lambda\mathsf{R}(x)\right\},$$

or a constrained optimization problem,

$$\min_{x\in\mathbb{R}^d} \left\{\mathsf{f}(x) \ : \ \mathsf{R}(x) \le c\right\},$$

where $\mathsf{R}(x)$ is a sparsity-promoting regularizer, and $\lambda$ and $c$ are tuning parameters controlling the penalization or constraint. Of course, choosing $\mathsf{R}(x) = \|x\|_0$ would reduce to the original target optimization problem, but these minimizations are generally only feasible to solve if $\mathsf{R}(x)$ is some relaxation of the sparsity constraint/penalty. For example, the Lasso [Tibshirani, 1996] uses a convex regularizer, $\mathsf{R}(x) = \|x\|_1$, which enjoys many strong guarantees of accurate estimation of the true sparse signal $x$ and of its support. More recently, many nonconvex penalties have been proposed that reduce the shrinkage bias of the Lasso, at the cost of a more challenging optimization problem, such as the SCAD [Fan and Li, 2001] and MCP [Zhang, 2010] penalties. The $\ell_q$ norm, for $q \in (0, 1)$, has also been extensively studied as a compromise between the convex but biased $\ell_1$ norm (as in the Lasso), and the theoretically optimal but computationally infeasible $\ell_0$ norm (i.e. the sparsity constraint, $\|x\|_0 \le s$). Results for the $\ell_q$ norm include work by Chartrand [2007],

Foucart and Lai [2009], Kabashima et al. [2009], Lai and Wang [2011]. Zheng et al. [2015]'s recent work studies the $\ell_q$ norm using the framework of approximate message passing to characterize its superior performance relative to the convex $\ell_1$ norm. While the resulting optimization problem is nonconvex for these alternatives to the $\ell_1$ norm, Loh and Wainwright [2013] show that restricted strong convexity in the objective function $\mathsf{f}$ is sufficient to outweigh bounded concavity in the penalty, to ensure successful optimization within a small error tolerance.

The penalized or constrained formulations of the sparse minimization problem may initially appear to be fundamentally different from the iterative thresholding approach. However, these penalized or constrained problems are often optimized with proximal gradient descent or projected gradient descent algorithms—specifically, for a penalty, the proximal gradient descent algorithm iterates the steps

$$\begin{cases} \text{Gradient step: } x_t' = x_{t-1} - \eta_t \cdot \nabla \mathsf{f}(x_{t-1}) \text{ for some step size } \eta_t, \\ \text{Proximal step: } x_t = \arg\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2}\|x - x_t'\|_2^2 + \eta_t \lambda \mathsf{R}(x) \right\}, \end{cases}$$

while for a constraint, projected gradient descent iterates the steps

$$\begin{cases} \text{Gradient step: } x_t' = x_{t-1} - \eta_t \cdot \nabla \mathsf{f}(x_{t-1}) \text{ for some step size } \eta_t, \\ \text{Projection step: } x_t = \arg\min_{x \in \mathbb{R}^d} \left\{ \frac{1}{2}\|x - x_t'\|_2^2 \ : \ \mathsf{R}(x) \le c \right\}. \end{cases}$$

Since $\mathsf{R}(x)$ is a sparsity-promoting regularizer, each iteration $x_t$ will therefore be sparse or approximately sparse. In this way, the penalized loss or constrained loss formulations of the sparse minimization problem can be viewed as analogous to the family of iterated thresholding algorithms, where the thresholding step is replaced by penalizing or constraining a regularizer $\mathsf{R}(x)$ that is a relaxation of the sparsity constraint.

# 3 Convergence of iterative thresholding

In this section, we examine the performance of gradient descent with iterative thresholding, for various choices of the thresholding operator $\Psi_s$. Specifically, after initializing at any point $x_0 \in \mathbb{R}^d$, the algorithm proceeds by alternating between taking a gradient descent step, and applying a thresholding operator:

$$x_t = \Psi_s\big(x_{t-1} - \eta_t \nabla \mathsf{f}(x_{t-1})\big), \tag{5}$$

where $\Psi_s : \mathbb{R}^d \to \{x \in \mathbb{R}^d : \|x\|_0 \le s\}$ is some thresholding operator that enforces $s$-sparsity at each step.

**Step size choice** Throughout the paper, we will primarily study this generalized iterative thresholding algorithm under the choice of a universal fixed step size $\eta = 1/\beta$, where $\beta$ is the restricted smoothness parameter for the function $\mathsf{f}$. When $\beta$ is unknown, we will also consider the following adaptive choice of step size based on exact line search:

$$\begin{cases} \text{Define } \widetilde{x}_t(\eta) = \Psi_s\big(x_{t-1} - \eta \nabla \mathsf{f}(x_{t-1})\big), \\ \text{Choose } \eta_t = \max\left\{ \eta \ge 0 : \mathsf{f}(\widetilde{x}_t(\eta)) \le \mathsf{f}(x_{t-1}) + \langle \widetilde{x}_t(\eta) - x_{t-1}, \nabla \mathsf{f}(x_{t-1}) \rangle + \frac{1}{2\eta}\|\widetilde{x}_t(\eta) - x_{t-1}\|_2^2 \right\}, \\ \text{Set } x_t = \widetilde{x}_t(\eta_t). \end{cases} \tag{6}$$

Note that, since $x_{t-1}$ and $\widetilde{x}_t(\eta)$ are both $s$-sparse, the curvature condition

$$\mathsf{f}(\widetilde{x}_t(\eta)) \le \mathsf{f}(x_{t-1}) + \langle \widetilde{x}_t(\eta) - x_{t-1}, \nabla \mathsf{f}(x_{t-1}) \rangle + \frac{1}{2\eta}\|\widetilde{x}_t(\eta) - x_{t-1}\|_2^2 \tag{7}$$

is necessarily satisfied for any $\eta \leq \frac{1}{\beta}$ due to the restricted smoothness property. Therefore we will always have $\eta_t \geq \frac{1}{\beta}$. Intuitively, the rule not only helps us get rid of the need to know $\beta$, but also allows the algorithm to take larger step size for more progress when possible. In practice, we would consider using a backtracking line search, that is, starting from a large step size and iteratively shrinking it until condition (7) is satisfied. In this way, condition (7) is similar to the classical Armijo rule for backtracking line search. For simplicity of our theoretical result we do not treat inexact linesearch in the following.

**Restricted optimality** Given an iterative algorithm that keeps the sparsity of the iterations at $s$, as discussed in Section 2.2, we cannot hope to achieve *global optimality* (i.e. a guarantee that $\mathsf{f}(x_t)$ is nearly as good as the best $s$-sparse solution, $\min_{\|x\|_0 \leq s} \mathsf{f}(x)$), but we can instead prove guarantees of *restricted optimality*, that is $\lim_{t\to\infty} \mathsf{f}(x_t) \leq \min_{\|x\|_0 \leq s'} \mathsf{f}(x)$, for some tighter sparsity constraint $s' \leq s$. We will assess a thresholding operator $\Psi_s$ based on its ability to guarantee restricted optimality relative to a sparsity level $s'$ that is as close to $s$ as possible, i.e. a sparsity ratio $\rho = s'/s$ that is as close to 1 as possible.

## 3.1 Relative concavity of a thresholding operator

Let $s \in \{1, \ldots, d\}$ be any fixed sparsity level and let $\rho \in [0, 1]$. We define the *relative concavity* of an $s$-sparse thresholding operator $\Psi_s$ relative to sparsity proportion $\rho$ as

$$\gamma_{s,\rho}(\Psi_s) = \sup \left\{ \frac{\langle y - \Psi_s(z), z - \Psi_s(z) \rangle}{\|y - \Psi_s(z)\|_2^2} \ : \ y, z \in \mathbb{R}^d, \ \|y\|_0 \leq \rho s, \ y \neq \Psi_s(z) \right\}.$$

Note that $\frac{\langle y-\Psi_s(z), z-\Psi_s(z)\rangle}{\|y-\Psi_s(z)\|_2^2}$ is the coefficient of projection when projecting $z-\Psi_s(z)$ onto $y-\Psi_s(z)$, and measures how much these two vectors align. To understand the term "relative concavity" in the name, we note that if $\Psi_s$ were a projection operator to some convex constraint set $\mathcal{C}$, then we would have $\langle y - \Psi_s(z), z - \Psi_s(z) \rangle \leq 0$ for any $y \in \mathcal{C}$, by the properties of convex projections. For sparse estimation, the constraint $\|x\|_0 \leq s$ is not convex; any positive values of $\langle y-\Psi_s(z), z-\Psi_s(z)\rangle$ with $\|y\|_0 \leq s$ measure the extent to which the thresholding operator $\Psi_s$ behaves *differently* from a convex projection. By taking a more restrictive constraint on $y$, namely $\|y\|_0 \leq \rho s$ rather than $\|y\|_0 \leq s$, we reduce this measure of concavity; the relative concavity of $\Psi_s$ will be smaller for lower values of $\rho$.

This notion of relative concavity is closely related to the *local concavity coefficients* developed in Barber and Ha [2017] for the purpose of studying projected gradient descent with an arbitrary nonconvex constraint. We will compare the two later on, after presenting our main theorems.

## 3.2 Relative concavity and iterative thresholding

We now examine how the relative concavity of $\Psi_s$ relates to the convergence behavior of iterative thresholding with a fixed step size. The main message, casted informally, is this:

> Given sparsity levels $s$ and $s' = \rho s$, and an $s$-sparse thresholding operator $\Psi_s$, the condition $\gamma_{s,\rho}(\Psi_s) \leq \frac{1}{2\kappa}$ is both necessary and sufficient for restricted optimality to hold relative to sparsity level $s'$.

**Stationary points** Before giving our formal results, we start with a warm-up—supposing that $x$ is a stationary point of the iterative thresholding algorithm with step size $\eta = \frac{1}{\beta}$, what guarantees can we give about $\mathsf{f}(x)$? If $\mathsf{f}$ satisfies $(\alpha, s)$-RSC, then we know that

$$\mathsf{f}(y) \geq \mathsf{f}(x) + \langle y - x, \nabla \mathsf{f}(x) \rangle + \frac{\alpha}{2}\|x - y\|_2^2$$

for any $s$-sparse $y$. Furthermore, writing $z = x - \eta \nabla \mathsf{f}(x)$, we know that $\Psi_s(z) = x$ since $x$ is a stationary point. Therefore,

$$\langle y - x, \nabla \mathsf{f}(x) \rangle = -\beta \langle y - x, z - x \rangle \geq -\beta \gamma_{s,\rho}(\Psi_s) \|x - y\|_2^2 \geq -\frac{\alpha}{2} \|x - y\|_2^2, \tag{8}$$

as long as $y$ is $\rho s$-sparse and the relative concavity satisfies $\gamma_{s,\rho}(\Psi_s) \leq \frac{1}{2\kappa}$. In other words, this condition on relative concavity is sufficient to ensure that

$$\mathsf{f}(x) \leq \min_{\|y\|_0 \leq \rho s} \mathsf{f}(y) \text{ for any stationary point } x.$$

Conversely, if $\gamma_{s,\rho}(\Psi_s) > \frac{1}{2\kappa}$, Theorem 2 below will construct a stationary point $x$ that *fails* to satisfy $\mathsf{f}(x) \leq \min_{\|y\|_0 \leq \rho s} \mathsf{f}(y)$.

**Convergence results** Next we turn to results for the iterated thresholding algorithm initialized at an arbitrary $s$-sparse point $x_0$ (for example, initialized at zero). Our first theorem accounts for the sufficiency of the condition.

**Theorem 1.** *Consider any objective function $\mathsf{f} : \mathbb{R}^d \to \mathbb{R}$, any sparsity levels $s \geq s'$, and any $s$-sparse thresholding operator $\Psi_s$. Assume the objective function $\mathsf{f}$ satisfies $(\alpha, s)$-RSC and $(\beta, s)$-RSM. Let $\rho = s'/s$ and $\kappa = \beta/\alpha$, and assume that*

$$\gamma_{s,\rho}(\Psi_s) < \frac{1}{2\kappa}.$$

*Then, for any $s$-sparse $x_0 \in \mathbb{R}^d$ and any $s'$-sparse $y \in \mathbb{R}^d$, the iterated thresholding algorithm (5) initialized at $x_0$ and run with fixed step size $\eta = 1/\beta$ satisfies*

$$\min_{t=1,\ldots,T} \mathsf{f}(x_t) \leq \mathsf{f}(y) + \left( \frac{1 - 1/\kappa}{1 - 2\gamma_{s,\rho}(\Psi_s)} \right)^T \cdot \frac{\beta}{2} \|x_0 - y\|_2^2$$

*for each $T \geq 1$. The same result holds for the iterative thresholding algorithm with adaptive step size (6).*

In other words, the condition $\gamma_{s,\rho}(\Psi_s) < \frac{1}{2\kappa}$ guarantees restricted optimality on the class of $\kappa$-conditioned objective functions at sparsity proportion $\rho$. Next, we examine the necessity of the bound on $\gamma_{s,\rho}(\Psi_s)$. The following result proves that, if $\gamma_{s,\rho}(\Psi_s) > \frac{1}{2\kappa}$, then there exists an objective function $\mathsf{f}(x)$ on which the restricted optimality guarantee fails, when we run iterative thresholding with fixed step size $\eta = \frac{1}{\beta}$.

**Theorem 2.** *Consider any sparsity levels $s \geq s'$, any $s$-sparse thresholding operator $\Psi_s$, and any constants $\beta \geq \alpha > 0$. Let $\rho = s'/s$ and $\kappa = \beta/\alpha$, and assume that*

$$\gamma_{s,\rho}(\Psi_s) > \frac{1}{2\kappa}.$$

*Then there exists an objective function $\mathsf{f}(x)$ that satisfies $(\alpha, s)$-RSC and $(\beta, s)$-RSM, and an $s$-sparse $x_0 \in \mathbb{R}^d$ and $s'$-sparse $y \in \mathbb{R}^d$, such that the iterated thresholding algorithm (5) run with step size $\eta = 1/\beta$ and initialization point $x_0$ satisfies*

$$\lim_{t \to \infty} \mathsf{f}(x_t) > \mathsf{f}(y).$$

This result is proved by constructing an objective function $\mathsf{f}$ and an $s$-sparse point $x_0$, such that $\mathsf{f}(x_0) > \mathsf{f}(y)$, but $x_0$ is a stationary point of the iterated thresholding algorithm, i.e. by initializing at $x_0$, we obtain $x_t = x_0$ for all $t \geq 1$. This proves that the iterated thresholding algorithm does not satisfy restricted optimality (at the given sparsity levels), since it is trapped at an $s$-sparse point $x_0$ whose objective value is strictly worse than that of the $s'$-sparse point $y$.

**Local vs global guarantees** We have seen that the condition $\gamma_{s,\rho}(\Psi_s) < \frac{1}{2\kappa}$ on the relative concavity, is sufficient to ensure a restricted optimality result, without any initialization conditions—that is, this is a global result, rather than a result that holds only in some neighborhood of the optimal solution. We can compare this framework to the local concavity coefficients of Barber and Ha [2017], where the convergence guarantee is of a local type.

In the present work, to achieve our convergence result via relative concavity, we require that, for any $z \in \mathbb{R}^d$ and for $x = \Psi_s(z)$, we have $\langle y - x, z - x\rangle \leq \gamma_{s,\rho}(\Psi_s)\|x - y\|_2^2$ for all $\rho s$-sparse $y$. For a stationary point $x$ of the iterated thresholding algorithm with step size $\eta = \frac{1}{\beta}$, we would have $z = x - \eta\nabla \mathsf{f}(x)$, and so the requirement above can be rewritten as

$$\langle y - x, -\nabla \mathsf{f}(x)\rangle \leq \beta \cdot \gamma_{s,\rho}(\Psi_s) \cdot \|x - y\|_2^2 \text{ for all } y \text{ with } \|y\|_0 \leq \rho s, \tag{9}$$

and the term $\beta\gamma_{s,\rho}(\Psi_s)$ on the right-hand side is bounded as $\beta\gamma_{s,\rho}(\Psi_s) < \beta \cdot \frac{1}{2\kappa} = \frac{\alpha}{2}$ according to the conditions of Theorem 1.

In contrast, Barber and Ha [2017]'s local concavity coefficient framework requires that, for any $z \in \mathbb{R}^d$ and any $y \in \mathcal{C}$, $\langle y - x, z - x\rangle \leq \gamma_x(\mathcal{C}) \cdot \|z - x\| \cdot \|y - x\|_2^2$ where $x = P_{\mathcal{C}}(z)$ is the projection of $z$ to the constraint set $\mathcal{C}$. At a stationary point $x$ of projected gradient descent, we have $z = x - \eta\nabla \mathsf{f}(x)$, and so equivalently,

$$\langle y - x, -\nabla \mathsf{f}(x)\rangle \leq \gamma_x(\mathcal{C}) \cdot \|\nabla \mathsf{f}(x)\| \cdot \|y - x\|_2^2 \text{ for all } y \in \mathcal{C} \ . \tag{10}$$

Barber and Ha [2017]'s main results prove convergence to the global minimum over $\mathcal{C}$, as long as the algorithm is initialized in a neighborhood within which the condition $\gamma_x(\mathcal{C})\|\nabla \mathsf{f}(x)\| < \frac{\alpha}{2}$ holds uniformly.[2]

Comparing the relative concavity framework (9) with Barber and Ha [2017]'s local concavity coefficient framework (10), we see that in both settings, $\langle y - x, -\nabla \mathsf{f}(x)\rangle$ is required to strictly less than $\frac{\alpha}{2}\|y - x\|_2^2$. The difference is that:

- Barber and Ha [2017]'s work requires this bound to hold for all $y$ in the constraint set $\mathcal{C}$, but only for $x$ in some neighborhood the global optimum. If the algorithm is initialized in this neighborhood, then global optimality is guaranteed.
- Our present work requires this bound to hold only for a more restricted set of $y$'s, i.e. with the restricted sparsity level $\|y\|_0 \leq s' = \rho s$, but for all $x$ in the constraint set of $s$-sparse vectors. Regardless of where the algorithm is initialized, we obtain a restricted optimality guarantee.

Overall, by requiring the concavity bound to hold only for a more restricted set of $y$'s, our new result is able to avoid initialization conditions, at the cost of obtaining restricted optimality rather than global optimality as the final guarantee.

# 4 Upper and lower bounds on relative concavity

We have now seen that the relative concavity $\gamma_{s,\rho}(\Psi_s)$ fully characterizes the performance of the thresholding operator $\Psi_s$ in the gradient descent algorithm, with a convergence guarantee in Theorem 1 and a matching lower bound in Theorem 2 (assuming a fixed step size). In this next section, we turn to the question of investigating the relative concavity in greater detail, in order to determine which thresholding operators are most likely to lead to successful optimization. Along the way, we will focus on the following questions:

[2]The norm $\|\cdot\|$ measuring the magnitude of the gradient $\nabla \mathsf{f}(x)$ is not necessarily the $\ell_2$ norm—it is typically chosen to be smaller than the $\ell_2$ norm, for instance, the $\ell_\infty$ norm in the case of sparse estimation—but this is not relevant to the comparison here.

- What is the relative concavity of commonly used thresholding operators, for example, hard thresholding and soft thresholding?
- What is the best (i.e. lowest) possible relative concavity $\gamma_{s,\rho}(\Psi_s)$ among all thresholding operators $\Psi_s$, and which thresholding operators are optimal?

Throughout this section, for providing upper and lower bounds on $\gamma_{s,\rho}(\Psi_s)$, we will assume without comment that $s, s' \in \{1, \dots, d\}$ are two sparsity levels satisfying $1 \le s' \le s \le d$ and $s + s' \le d$, and we will define $\rho = s'/s$ as usual.

## 4.1 Relative concavity of hard and soft thresholding

First, we consider hard thresholding, $\Psi_s = \Psi_s^{\mathrm{HT}}$. The following result computes the relative concavity for the hard thresholding operator:

**Lemma 1.** *The relative concavity of hard thresholding is given by*

$$\gamma_{s,\rho}(\Psi_s^{\mathrm{HT}}) = \frac{\sqrt{\rho}}{2}$$

*for every sparsity proportion* $\rho \in (0, 1]$.

In particular, with Lemma 1, the condition $\gamma_{s,\rho}(\Psi_s^{\mathrm{HT}}) < \frac{1}{2\kappa}$ becomes $\rho < \frac{1}{\kappa^2}$. In light of Theorems 1 and 2, we see that for iterative hard thresholding algorithm, $\rho < \frac{1}{\kappa^2}$ is necessary and sufficient to guarantee restricted optimality with sparsity level $s$ and $s'$, tightening the condition obtained in Jain et al. [2014] where they prove restricted optimality with the sparsity proportion $\rho = \frac{1}{32\kappa^2}$.

We might wonder whether the highly discontinuous nature of the hard thresholding function might not be ideal—by smoothing out the discontinuity, could we attain better performance? However, we find that any continuous thresholding operator is necessarily worse than hard thresholding:

**Lemma 2.** *For any continuous map* $\Psi_s : \mathbb{R}^d \to \{x \in \mathbb{R}^d : \|x\|_0 \le s\}$*, its relative concavity satisfies*

$$\gamma_{s,\rho}(\Psi_s) \ge 1$$

*for every sparsity proportion* $\rho \in (0, 1]$.

In particular, since $\kappa \ge 1$, the condition $\gamma_{s,\rho}(\Psi_s) < \frac{1}{2\kappa}$ never holds if $\Psi_s$ is continuous. Comparing to Theorem 2, we see that no continuous operator can guarantee restricted optimality at any sparsity ratio $\rho$, even in the ideal setting where $\mathsf{f}$ is well-conditioned.

## 4.2 Optimal value of relative concavity

In this section we turn to the question of optimality: what is the optimal value of relative concavity among all thresholding operators at a given sparsity proportion $\rho$? We will establish that

$$\inf_{\Psi_s : \mathbb{R}^d \to \{x \in \mathbb{R}^d : \|x\|_0 \le s\}} \gamma_{s,\rho}(\Psi_s) = \frac{\rho}{1+\rho}.$$

That is, the lowest relative concavity among all thresholding operators at a given sparsity proportion $\rho$ is exactly $\frac{\rho}{1+\rho}$. Since this is much smaller than $\frac{\sqrt{\rho}}{2}$ when $\rho$ is small, we see that hard thresholding is suboptimal.

We start with the following lower bound for all thresholding operators:

**Lemma 3.** *For any map $\Psi_s : \mathbb{R}^d \to \{x \in \mathbb{R}^d : \|x\|_0 \le s\}$ and any sparsity proportion $\rho \in (0, 1]$, the relative concavity is lower-bounded as*

$$\gamma_{s,\rho}(\Psi_s) \ge \frac{\rho}{1+\rho}.$$

To show that this lower bound is indeed tight, we will consider $\ell_q$ thresholding and establish upper bound for its relative concavity that matches this lower bound with proper choices of $q$. $\ell_q$ thresholding encourage sparsity without exerting too much shrinkage by constraining the $\ell_q$ norm of the vector after thresholding for some $q \in (0, 1)$. To be precise, let

$$P_{\ell_q}(z; t) = \arg\min \{\|x - z\|_2 : \|x\|_q \le t\}$$

denote projection to the $\ell_q$ ball, where $\|x\|_q = (\sum_i |x_i|^q)^{1/q}$ is the $\ell_q$ "norm" (in fact a nonconvex function since $q < 1$). Then define

$$\Psi_s^{\ell_q}(z) = P_{\ell_q}(z; t(z)), \text{ where } t(z) = \sup \left\{t : \|P_{\ell_q}(z; t)\|_0 \le s\right\}.$$

In words, $\Psi_s^{\ell_q}(z)$ projects $z$ to an $\ell_q$ ball whose radius is chosen to be as large as possible while still ensuring $s$-sparsity.[3] The following result computes the relative concavity for $\ell_q$ thresholding:

**Lemma 4.** *The relative concavity of $\ell_q$ thresholding $\Psi_s^{\ell_q}$ is equal to*

$$\gamma_{s,\rho}(\Psi_s^{\ell_q}) = \frac{\frac{\rho}{\min\{1, (\frac{2-q}{q})^2(1-\rho)\}}}{\frac{4q(1-q)}{(2-q)^2}(1 + \sqrt{1 + (\frac{2-q}{q})^2 \frac{\rho}{\min\{1, (\frac{2-q}{q})^2(1-\rho)\}}})}$$

*for every sparsity proportion $\rho \in (0, 1)$. In particular, if we choose*

$$q = \frac{2(1-\rho)}{3-\rho},$$

*then the resulting thresholding operator attains the lowest possible relative concavity,*

$$\gamma_{s,\rho}(\Psi_s^{\ell_q}) = \frac{\rho}{1+\rho}, \text{ for } q = \frac{2(1-\rho)}{3-\rho}.$$

*In addition, the universal choice $q = 2/3$ yields relative concavity equal to,*

$$\gamma_{s,\rho}(\Psi_s^{\ell_{2/3}}) = \frac{\frac{\rho}{\min\{1, 4(1-\rho)\}}}{\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4\rho}{\min\{1, 4(1-\rho)\}}}} \le \frac{\rho}{\min\{1, 4(1-\rho)\}}, \text{ for all } \rho \in (0, 1).$$

Now we provide some explanation for this result. If we are allowed to choose $q$ depend on $\rho$, then the choice $q = \frac{2(1-\rho)}{3-\rho}$ would lead to a relative concavity of $\frac{\rho}{1+\rho}$, which exactly matches the lower bound in Lemma 3. Of course this specific choice of $q$ is chosen for a specific sparsity proportion $\rho$ and might not work well for other values of the sparsity proportion. To avoid this drawback or the need to tune the parameter $q$, one can have the universal choice $q = 2/3$. Due to the expression for $\gamma_{s,\rho}(\Psi_s^{\ell_{2/3}})$, we see that $\gamma_{s,\rho}(\Psi_s^{\ell_{2/3}}) \approx \rho$ when $\rho$ is small, thus nearly matching the lower bound $\frac{\rho}{1+\rho}$.

In particular, with the optimal value of relative concavity $\gamma_{s,\rho} = \frac{\rho}{1+\rho}$, the condition $\gamma_{s,\rho} < \frac{1}{2\kappa}$ becomes $\rho < \frac{1}{2\kappa-1}$. In light of Theorem 1 and Theorem 2, we see that $\rho < \frac{1}{2\kappa-1}$ is both necessary and sufficient for restricted optimality to hold with sparsity proportion $\rho$. Compare this with the condition $\rho < \frac{1}{\kappa^2}$ required by hard thresholding, we see that the dependence on condition number is greatly improved!

[3]Note that $P_{\ell_q}(z; t)$ may be non-unique. To be fully precise, we define $\Psi_s^{\ell_q}(z)$ by first fixing some map $z \mapsto S$, the possibly non-unique support of its largest $s$ entries, and then defining $t(z)$ and choosing the possibly non-unique projection $P_{\ell_q}(z; t(z))$ in such a way that the nonzero entries in the projection are exactly on this support.

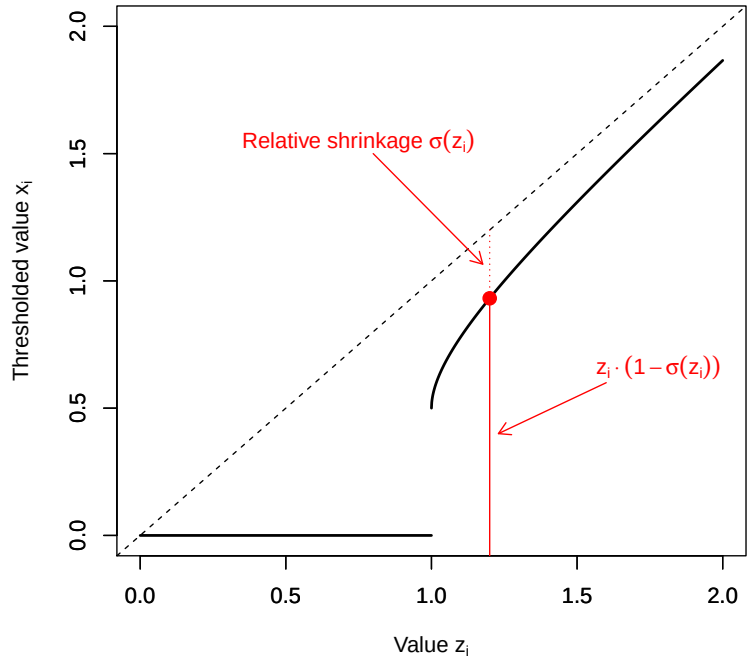


Figure 1: An illustration of the definition of the relative shrinkage function $\sigma$ (for simplicity we use thresholding level $\tau = 1$ in this illustration). Here we use the relative shrinkage function $\sigma(t) = \frac{t-\sqrt{t^2-1}}{2}$, corresponding to the reciprocal thresholding operator defined in Section 4.4.

### 4.3 A general class of thresholding operators

Now that we have seen that $\ell_q$ thresholding operators enjoy good properties in terms of relative concavity, we can ask whether there are other thresholding operators of such optimal and near-optimal properties. In this section we address this problem by showing $\ell_q$ thresholding can be characterized as a special case of a larger class of thresholding operators, which all enjoy the same optimality properties in the sense of their relative concavity. Consider any nonincreasing function

$$\sigma : [1, \infty) \to [0, 1],$$

which we call the "shrinkage function", which will determine the amount of shrinkage on each entry of a vector $z$ at the thresholding step. Defining the support $S$ and thresholding level $\tau = \max_{i \notin S} |z_i|$ as before, we then define the thresholding operator $\Psi_{s;\sigma}$ as

$$\big(\Psi_{s;\sigma}(z)\big)_i = \begin{cases} z_i - \tau\sigma\big(|z_i|/\tau\big), & i \in S, \\ 0, & i \notin S. \end{cases}$$

In other words, for entry $i \in S$, $\sigma\big(|z_i|/\tau\big)$ determines the *relative* amount of shrinkage on this entry. The intuitive meaning of $\sigma$ is illustrated in Figure 1. (If $\tau = 0$, i.e. $z$ is already $s$-sparse, then we would simply take $\Psi_{s;\sigma}(z) = z$; we will ignore this case from this point on.)

Note that since $\sigma$ is nondecreasing, the maximum shrinkage occurs when $|z_i| = \tau$ exactly; the amount of shrinkage in this setting is governed by $\sigma(1)$.

We can now examine the relationship of the choice of $\sigma$ to the relative concavity:

**Lemma 5.** *For any nonincreasing shrinkage function $\sigma : [1, \infty) \to [0, 1]$ such that $0 < \sigma(1) < 1$ and*

$$t \mapsto \sigma(t)(t - \sigma(t)) \text{ is nondecreasing over } t \geq 1, \tag{11}$$

*the thresholding operator $\Psi_{s;\sigma}$ has relative concavity*

$$\gamma_{s,\rho}(\Psi_{s;\sigma}) = \frac{\frac{\rho}{\min\{1,(1-\rho)/\sigma(1)^2\}}}{2\sigma(1)\big(1-\sigma(1)\big)\left(1+\sqrt{1+\frac{\rho/\sigma(1)^2}{\min\{1,(1-\rho)/\sigma(1)^2\}}}\right)}.$$

*In particular, the resulting operator attains the lowest possible relative concavity,*

$$\gamma_{s,\rho}(\Psi_{s;\sigma}) = \frac{\rho}{1+\rho},$$

*if and only if $\sigma(1) = \frac{1-\rho}{2}$. If instead we take a universal shrinkage level $\sigma(1) = \frac{1}{2}$, then the relative concavity is given by*

$$\gamma_{s,\rho}(\Psi_{s;\sigma}) = \frac{\frac{\rho}{\min\{1,4(1-\rho)\}}}{\frac{1}{2}+\frac{1}{2}\sqrt{1+\frac{4\rho}{\min\{1,4(1-\rho)\}}}} \leq \frac{\rho}{\min\{1,4(1-\rho)\}}.$$

Examining the definition of this general family of thresholding operators, we can see that $\ell_q$ thresholding corresponds to setting

$$\sigma(t) = t - \left(\text{the larger-magnitude root } x \text{ of the equation } t = x + \frac{q(2-2q)^{1-q}/(2-q)^{2-q}}{x^{1-q}}\right),$$

for which we have $\sigma(1) = \frac{q}{2-q}$ and which satisfies (11). We also have that $\sigma(1) = \frac{1}{2}$ corresponds to the "universal" choices $q = 2/3$, and $\sigma(1) = \frac{1-\rho}{2}$ (the optimal value) corresponds to the $\rho$-specific choices $q = \frac{2(1-\rho)}{3-\rho}$. As a consequence, the previous result for $\ell_q$ thresholding, Lemma 4, is simply special case of this more general lemma. On the other hand, the hard thresholding operator $\Psi_s^{\mathrm{HT}}$ can be obtained by setting $\sigma(t) = 0$ for all $t \in [1, \infty)$, but this does not satisfy the assumption $\sigma(1) > 0$ required in the lemma. However, if we informally consider fixing $\rho > 0$ and taking a limit $\sigma(1) \to 0$ in the upper bound in the lemma, we see

$$\lim_{\sigma(1)\to 0} \frac{\frac{\rho}{\min\{1,\frac{1-\rho}{\sigma(1)^2}\}}}{2\sigma(1)(1-\sigma(1))\left(1+\sqrt{1+\frac{\rho}{\sigma(1)^2\min\{1,\frac{1-\rho}{\sigma(1)^2}\}}}\right)} = \frac{\sqrt{\rho}}{2},$$

obtaining the relative concavity of hard thresholding calculated earlier.

## 4.4 Reciprocal thresholding and minimal shrinkage

Practically, for two thresholding operators with the same restricted optimality guarantees, i.e. with the exact same value of relative concavity, we may favor the one that exerts smaller amount of shrinkage. Thus it makes sense to ask among the general class of thresholding operators defined in Section 4.3, which operators exert the minimal amount of shrinkage? Consider all operators of

the form $\Psi_{s;\sigma}$, with some fixed value of $\sigma(1) \in (0, 1/2]$. For any $\sigma$ satisfying the assumption (11), for all $t \geq 1$ we have

$$\sigma(t)(t - \sigma(t)) \geq \sigma(1)(1 - \sigma(1)).$$

For convenience, we reparametrize this equation by setting $c = 1 - 2\sigma(1) \in [0, 1)$, and so we are considering all nonincreasing functions $\sigma : [1, \infty) \to [0, 1]$ that satisfy $\sigma(1) = \frac{1-c}{2}$ and

$$\sigma(t)(t - \sigma(t)) \geq \sigma(1)(1 - \sigma(1)) = \frac{1 - c^2}{4}.$$

Thus, we must have

$$\sigma(t) \geq \frac{t - \sqrt{t^2 - (1 - c^2)}}{2} \tag{12}$$

for all $t \geq 1$.

This motivates a new family of thresholding operators, *reciprocal thresholding with parameter* $c$, which is designed to make the inequality (12) an equality. To be specific, we define reciprocal thresholding with parameter $c$ to be

$$\Psi_s^{\mathrm{RT},c} = \Psi_{s;\sigma} \text{ with shrinkage function } \sigma(t) = \frac{t - \sqrt{t^2 - (1 - c^2)}}{2}.$$

To apply this operator to some vector $z \in \mathbb{R}^d$, we first let $S \subset \{1, \ldots, d\}$ be the indices of the largest $s$ entries of $z$ (with our usual caveat about needing to establish some rule for breaking ties) and let $\tau = \max_{i \notin S} |z_i|$ be the magnitude of the $(s+1)$-st largest entry of $z$. Then $\Psi_s^{\mathrm{RT},c}(z)$ operates entry-wise as follows:

$$\left(\Psi_s^{\mathrm{RT},c}(z)\right)_i = \begin{cases} \mathrm{sign}(z_i) \cdot \left(\frac{1}{2}|z_i| + \frac{1}{2}\sqrt{|z_i|^2 - \tau^2(1 - c^2)}\right), & \text{if } i \in S, \\ 0, & \text{if } i \notin S. \end{cases} \tag{13}$$

Here the thresholded value $\left(\Psi_s^{\mathrm{RT},c}(z)\right)_i$ is equal to the larger-magnitude root $t$ of the equation

$$z_i = t + \frac{\tau^2 \cdot \frac{1-c^2}{4}}{t}, \tag{14}$$

hence the name "reciprocal" thresholding.

As before, to avoid the need for selecting $c$ adaptively, we might want to consider some fixed choices. At one extreme, taking $c = 1$ yields $\Psi_s^{\mathrm{RT},1} = \Psi_s^{\mathrm{HT}}$, the hard thresholding operator. At the other extreme, taking $c = 0$ defines the "universal" *reciprocal thresholding* operator:

$$\Psi_s^{\mathrm{RT}} = \Psi_s^{\mathrm{RT},0}.$$

For any $z \in \mathbb{R}^d$, $\Psi_s^{\mathrm{RT}}$ operate entry-wise as:

$$\left(\Psi_s^{\mathrm{RT}}(z)\right)_i = \begin{cases} \mathrm{sign}(z_i) \cdot \left(\frac{1}{2}|z_i| + \frac{1}{2}\sqrt{|z_i|^2 - \tau^2}\right), & \text{if } i \in S, \\ 0, & \text{if } i \notin S. \end{cases} \tag{15}$$

The following lemma calculates the relative concavity of $\Psi_s^{\mathrm{RT},c}$ and $\Psi_s^{\mathrm{RT}}$ as a direct consequence of Lemma 5.

**Lemma 6.** *For any sparsity proportion $\rho \in (0, 1]$, the thresholding operator $\Psi_s^{\mathrm{RT},c}$ with parameter $c = \rho$ has relative concavity equal to*

$$\gamma_{s,\rho}(\Psi_s^{\mathrm{RT},\rho}) = \frac{\rho}{1 + \rho}.$$

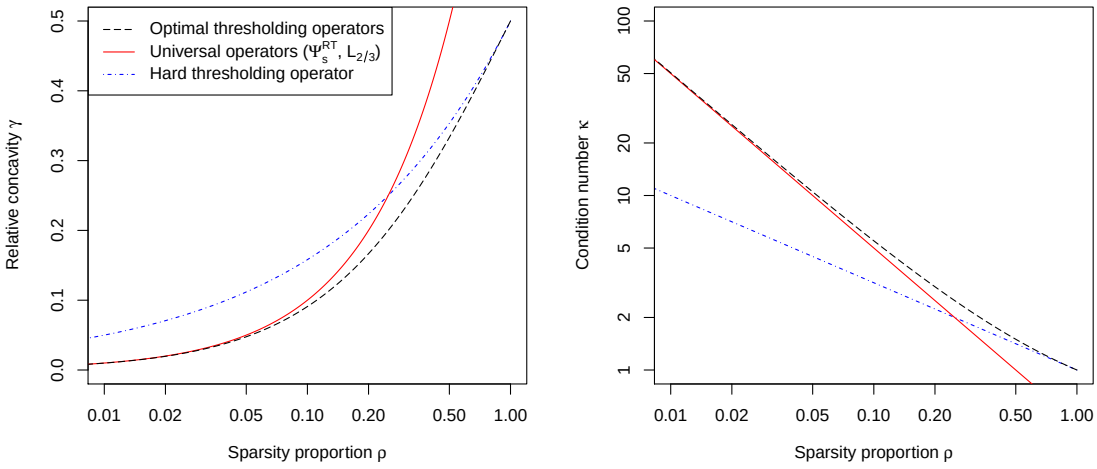


Figure 2: A comparison of three values of relative concavity: the optimal relative concavity (attained by, for instance, $\Psi_s^{\mathrm{RT},c}$ with $c = \rho$, or by $\Psi_s^{\ell_q}$ with $q = \frac{2(1-\rho)}{3-\rho}$); the relative concavity obtained by the "universal" operators, including $\Psi_s^{\mathrm{RT}}$ and by $\ell_{2/3}$ thresholding; and the relative concavity of hard thresholding. The left plot shows the relative concavity as a function of the sparsity proportion $\rho$. The right plot shows the largest possible condition number $\kappa$ of the objective function $\mathsf{f}$ for which a restricted optimality guarantee can be attained (Theorems 1 and 2 show that $\gamma_{s,\rho}(\Psi_s) \leq \frac{1}{2\kappa}$ is necessary and sufficient for a restricted optimality guarantee).

*The reciprocal thresholding operator* $\Psi_s^{\mathrm{RT}}$ *has relative concavity equal to*

$$\gamma_{s,\rho}(\Psi_s^{\mathrm{RT}}) = \frac{\frac{\rho}{\min\{1,4(1-\rho)\}}}{\frac{1}{2} + \frac{1}{2}\sqrt{1 + \frac{4\rho}{\min\{1,4(1-\rho)\}}}} \leq \frac{\rho}{\min\{1, 4(1-\rho)\}}$$

*for every sparsity proportion* $\rho \in (0, 1)$.

Thus, $\Psi_s^{\mathrm{RT},c}$ with $c = \rho$ is exactly optimal among all thresholding operators relative to the sparsity proportion $\rho$ (as is $\Psi_s^{\ell_q}$ with $q = \frac{2(1-\rho)}{3-\rho}$), while $\Psi_s^{\mathrm{RT}}$ is near optimal when $\rho$ is small (as is $\Psi_s^{\ell_{2/3}}$).

## 4.5 An illustrative comparison

Through the development in this section, we see that there are three important benchmarks for relative concavity: the bound $\sqrt{\rho}/2$ attained by hard thresholding $\Psi_s^{\mathrm{HT}}$, the bound $\frac{\frac{\rho}{\min\{1,4(1-\rho)\}}}{\frac{1}{2}+\frac{1}{2}\sqrt{1+\frac{4\rho}{\min\{1,4(1-\rho)\}}}}$ attained by reciprocal thresholding $\Psi_s^{\mathrm{RT}}$ and $\ell_{2/3}$ thresholding $\Psi_s^{\ell_{2/3}}$, and the optimal value $\frac{\rho}{1+\rho}$. In this section we provide a comparison between these three values.

The left-hand plot of Figure 2 displays the three values of relative concavity as functions of the sparsity proportion $\rho$. We see that at small values $\rho \approx 0$, the relative concavity of reciprocal thresholding and $\ell_{2/3}$ thresholding is nearly identical to the optimal bound $\frac{\rho}{1+\rho}$, and is substantially better than the relative concavity for hard thresholding, given by $\sqrt{\rho}/2$. At larger values of $\rho$, the relative concavity for hard thresholding is instead lower.

To view this comparison in another light, given any fixed thresholding operator $\Psi_s$ with certain relative concavity, and given an objective function $\mathsf{f}$ with condition number $\kappa$, for what sparsity ratio $\rho = s'/s$ is the iterative thresholding algorithm guaranteed to achieve restricted optimality? Using the condition $\gamma_{s,\rho}(\Psi_s) \leq \frac{1}{2\kappa}$, for each relative concavity $\gamma_{s,\rho}$ we can solve for the largest possible $\kappa$ for which restricted optimality is assured, as a function of $\rho$. This is illustrated in the right-hand plot of Figure 2, where we see that the reciprocal thresholding operator $\Psi_s^{\mathrm{RT}}$ and the $\ell_{2/3}$ thresholding operator achieve a nearly-optimal sparsity ratio $\rho$ when the condition number $\kappa$ is large and $\rho$ is correspondingly close to zero, while hard thresholding $\Psi_s^{\mathrm{HT}}$ is closer to optimal for $\kappa$ and $\rho$ close to 1. Thus, we can conclude that reciprocal thresholding and $\ell_{2/3}$ thresholding offer stronger theoretical guarantees when $\kappa > 2$, while hard thresholding may be better for very well-conditioned problems where $1 \leq \kappa < 2$. (Empirically, we have observed that it is often the case that the three perform nearly identically in "generic" problems, and only show substantial differences in problems constructed to mimic our lower bound result, Theorem 2, for example, in linear regression problems where a small subset of the features are generated to have covariance structure similar to the construction in Theorem 2.)

# 5 Iterative thresholding for low-rank matrices

We next extend our analysis of iterative thresholding methods to the setting of a low-rank constraint. In fact, our results carry over fully into this setting. Given a rank constraint, $\mathrm{rank}(X) \leq s$, the hard thresholding operator is defined as

$$\widetilde{\Psi}_s^{\mathrm{HT}} : X \mapsto U \cdot \mathrm{diag}(\Psi_s^{\mathrm{HT}}(d)) \cdot V^\top,$$

where $X = U \cdot \mathrm{diag}(d) \cdot V^\top$ is the singular value decomposition of $X$.[4] That is, hard thresholding is performed on the singular values of the matrix $X$, rather than on its entries. Of course, we can extend this to any thresholding operator—given any $\Psi_s : \mathbb{R}^{\min\{n,m\}} \to \{x \in \mathbb{R}^{\min\{n,m\}} : \|x\|_0 \leq s\}$, we can "lift" this thresholding operator to the matrix setting by defining

$$\widetilde{\Psi}_s : X \mapsto U \cdot \mathrm{diag}(\Psi_s(d)) \cdot V^\top. \tag{16}$$

Of course, its possible to construct a rank-$s$ thresholding operator $\widetilde{\Psi}_s$ that is not of the form given in (16), for example, if $\widetilde{\Psi}_s$ does not preserve the left and right singular vectors of $Z$.

We next extend our convergence results, Theorems 1 and 2, to the low-rank setting. In order to do so, we need to define the matrix version of relative concavity—this definition is analogous to the vector case, with rank constraints in place of sparsity constraints:

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) = \sup\left\{ \frac{\langle Y - \widetilde{\Psi}_s(Z), Z - \widetilde{\Psi}_s(Z)\rangle}{\|Y - \widetilde{\Psi}_s(Z)\|_{\mathrm{F}}^2} \;\; : Y, Z \in \mathbb{R}^{n\times m}, \mathrm{rank}(Y) \leq \rho s, Y \neq \widetilde{\Psi}_s(Z) \right\}.$$

As for the vector case, relative concavity is necessary and sufficient for guaranteeing restricted optimality—in fact, the proofs of these are completely identical to the vector case. For completeness, we state the results here, for the matrix version of the iterated thresholding algorithm:

$$X_t = \widetilde{\Psi}_s\big(X_{t-1} - \eta_t \nabla \mathsf{f}(X_{t-1})\big), \tag{17}$$

with either fixed step size $\eta_t = 1/\beta$ or adaptive step size defined as in (6).

[4] In the case of repeated singular values, the singular value decomposition will not be unique, and we assume that we have some mechanism for specifying a specific singular value decomposition. This is analogous to the sparse vector problem, where if the $s$th largest entry in $z$ is not unique, we need to assume some mechanism for breaking ties and choosing the support of the thresholded vector.

**Theorem 3.** *Consider any objective function* $\mathsf{f} : \mathbb{R}^{n\times m} \to \mathbb{R}$*, any ranks* $s \geq s'$*, and any rank-*$s$ *thresholding operator* $\widetilde{\Psi}_s$*. Assume the objective function* $\mathsf{f}$ *satisfies* $(\alpha, s)$*-RSC and* $(\beta, s)$*-RSM relative to the rank constraint.*[5] *Let* $\rho = s'/s$ *and* $\kappa = \beta/\alpha$*, and assume that* $\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) < \frac{1}{2\kappa}$*. Then, for any* $X_0, Y \in \mathbb{R}^{n\times m}$ *with* $\operatorname{rank}(X_0) \leq s$ *and* $\operatorname{rank}(Y) \leq s'$*, the iterated thresholding algorithm* (17) *run with step size* $\eta = 1/\beta$ *and initialization point* $X_0$ *satisfies*

$$\min_{t=1,\dots,T} \mathsf{f}(X_t) \leq \mathsf{f}(Y) + \left(\frac{1 - 1/\kappa}{1 - 2\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s)}\right)^T \cdot \frac{\beta}{2}\|X_0 - Y\|_{\mathrm{F}}^2$$

*for each* $T \geq 1$*.*

**Theorem 4.** *Consider any ranks* $s \geq s'$*, any rank-*$s$ *thresholding operator* $\widetilde{\Psi}_s$*, and any constants* $\beta \geq \alpha > 0$*. Let* $\rho = s'/s$ *and* $\kappa = \beta/\alpha$*, and assume that* $\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) > \frac{1}{2\kappa}$*. Then there exists an objective function* $\mathsf{f}(X)$ *that satisfies* $(\alpha, s)$*-RSC and* $(\beta, s)$*-RSM relative to the rank constraint, and matrices* $X_0, Y \in \mathbb{R}^{n\times m}$ *with* $\operatorname{rank}(X_0) \leq s$ *and* $\operatorname{rank}(Y) \leq s'$*, such that the iterated thresholding algorithm* (17) *run with step size* $\eta = 1/\beta$ *and initialization point* $X_0$ *satisfies*

$$\lim_{t\to\infty} \mathsf{f}(X_t) > \mathsf{f}(Y).$$

In other words, just as for the sparse optimization problem, the relationship between relative concavity and condition number gives a necessary and sufficient condition for guaranteed convergence. We note that these results apply to *any* rank-$s$ thresholding operator $\widetilde{\Psi}_s$, whether or not it can be constructed by "lifting" a $s$-sparse thresholding operator as in (16).

Next, how can we calculate relative concavity of a thresholding operator in the matrix setting? For simplicity, from this point on we assume that we are working with ranks $s \geq s' \geq 1$ with $s + s' \leq \min\{n, m\}$. For this question, we will again see that results from the sparse setting transfer to the low-rank setting. First, we have the same lower bound uniformly over all operators:

**Lemma 7.** *For any map* $\widetilde{\Psi}_s : \mathbb{R}^{n\times m} \to \{X \in \mathbb{R}^{n\times m} : \operatorname{rank}(X) \leq s\}$ *and any sparsity proportion* $\rho \in (0, 1]$*, the relative concavity is lower-bounded as*

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) \geq \frac{\rho}{1 + \rho}.$$

Furthermore, if we restrict our attention to "lifted" thresholding operators of the form (16), the relative concavity of $\Psi_s$ is inherited by the lifted operator $\widetilde{\Psi}_s$—as long as we restrict ourselves to $s$-sparse thresholding operators $\Psi_s$ that satisfy a natural sign condition:

$$\text{For any } z \in \mathbb{R}^d \text{ and any } a \in \{\pm 1\}^d,\ \Psi_s\big(\operatorname{diag}(a) \cdot z\big) = \operatorname{diag}(a) \cdot \Psi_s(z). \tag{18}$$

This effectively means that $\Psi_s(z)$ preserves the signs of $z$, but the signs of $z$ do not affect the amount of shrinkage in the thresholded vector $\Psi_s(z)$. For example, this requires that $\Psi_s(-z) = -\Psi_s(z)$. Under this assumption, the relative concavity of $\Psi_s$ carries over into the matrix setting.

**Lemma 8.** *Let* $\Psi_s$ *be a* $s$*-sparse thresholding operator satisfying the sign condition* (18)*, and let* $\widetilde{\Psi}_s$ *be the lifted thresholding operator defined in* (16)*. Then for every sparsity proportion* $\rho \in (0, 1]$*,*

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) = \gamma_{s,\rho}(\Psi_s).$$

[5] In the low-rank setting, the RSC and RSM conditions are defined with rank in place of sparsity—specifically, we are assuming that $\frac{\alpha}{2}\|X - Y\|_{\mathrm{F}}^2 \leq \mathsf{f}(Y) - \mathsf{f}(X) - \langle \nabla \mathsf{f}(X), Y - X\rangle \leq \frac{\beta}{2}\|X - Y\|_{\mathrm{F}}^2$ whenever $\operatorname{rank}(X) \leq s, \operatorname{rank}(Y) \leq s$.

It is obvious that all the thresholding operators we have considered satisfy the sign condition (18). Thus, all the results of relative concavity that we have proved in the sparse setting, carry over directly to the low-rank setting. In particular, as for the sparse setting, the hard thresholding operator has relative concavity

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s^{\mathrm{HT}}) = \frac{\sqrt{\rho}}{2},$$

while any thresholding operator $\widetilde{\Psi}_{s;\sigma}$ constructed with some shrinkage function $\sigma$ satisfying $\sigma(1) = 1/2$ and the conditions of Lemma 5, such as the reciprocal thresholding operator, $\widetilde{\Psi}_s^{\mathrm{RT}}$, or $\ell_q$ thresholding with $q = 2/3$, $\widetilde{\Psi}_s^{\ell_{2/3}}$, satisfy

$$\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_{s;\sigma}) = \widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s^{\ell_{2/3}}) = \widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s^{\mathrm{RT}}) \leq \frac{\rho}{\min\{1, 4(1-\rho)\}}.$$

If the desired rank proportion $\rho = s'/s$ is fixed in advance, then as before, choosing reciprocal thresholding with parameter $c = \rho$, or $\ell_q$ thresholding with $q = \frac{2(1-\rho)}{3-\rho}$, we again obtain the optimal relative concavity of $\frac{\rho}{1+\rho}$. As before, we can conclude that reciprocal thresholding and $\ell_{2/3}$ each offer lower relative concavity than hard thresholding whenever $\rho$ is small—and, correspondingly, are a safer choice for objective functions $\mathsf{f}$ whose condition number is not close to 1.

# 6 Sparse linear regression

Now that we have discussed the deterministic optimization setting in depth, it is natural to ask what is the implication of these guarantee for a statistically random setting. In this section, we apply our developed machinery to the concrete statistical setting of sparse linear regression. We work with the Gaussian linear model

$$y = X\theta_0 + z \tag{19}$$

where $X \in \mathbb{R}^{n\times p}$ is a fixed design matrix, $\theta_0 \in \mathbb{R}^p$ is the true coefficient vector assumed to be fixed and $s_0$-sparse, and $z \sim N(0, \sigma^2\mathbf{I}_n)$ is the noise vector, with fixed unknown noise level $\sigma^2 > 0$. In this section we will mainly be interested in prediction error, i.e. how well we can estimate the true mean vector $X\theta_0$. One way of capturing the conditioning of the design matrix is by the following definition: at some given sparsity level $s$, we define a set of design matrices $\mathcal{X}(\alpha, \beta, s)$ as

$$\mathcal{X}(\alpha,\beta,s) = \left\{ X \in \mathbb{R}^{n\times p} : \text{ the map } \theta \mapsto \theta^\top \left(\frac{X^\top X}{2n}\right)\theta \text{ satisfies } (\alpha, s)\text{-RSC and } (\beta, s)\text{-RSM} \right\}. \tag{20}$$

As usual, we will be interested in the condition number $\kappa = \beta/\alpha$. A similar definition is the *restricted eigenvalue* condition on the design matrix $X$, which constrains $X$ to the following set

$$\mathcal{X}_{\mathrm{RE}}(\kappa, s_0) = \left\{ X \in \mathbb{R}^{n\times p} : \max_{j=1,\ldots,p} \frac{\|X_j\|_2}{\sqrt{n}} \leq 1, \text{ and } \theta^\top \left(\frac{X^\top X}{2n}\right)\theta \geq \frac{1}{2\kappa}\|\theta\|_2^2 \right.$$
$$\left. \text{for all } \theta \in \mathbb{R}^d \text{ with } \|\theta\|_1 \leq 4 \max_{|S|=s_0} \|\theta_S\|_1 \right\}. \tag{21}$$

To gain some intuition for when these conditions may hold, for a design matrix $X$ whose rows are i.i.d. draws from a normal distribution $N(0, \Sigma)$, Raskutti et al. [2010, Theorem 1] show that the population-level eigenvalues of the covariance $\Sigma$ are approximately preserved in the design matrix, at any sparsity level $s \ll \frac{n}{\log(p)}$.

**Computational lower bound** In terms of prediction error, the optimal method, $\ell_0$ constrained least squares method, is not computable. Thus from the lower bound side, it is of interest to ask what is the lowest prediction error achievable in the class of computational feasible estimator. Recently, Zhang et al. [2014] provide a partial answer to this question, restricting to the class of $s_0$ sparse estimator. Their main result (see Theorem 1 in Zhang et al. [2014]) states the following (informally):

> Under the assumption that $NP \nsubseteq P \setminus poly$, for any $\delta \in (0,1)$, under some assumption on $n, d, s_0$ and for any $\kappa$ in a wide range, there exists a design matrix $X \in \mathcal{X}_{\mathrm{RE}}(\kappa, s_0)$ such that for any computational efficient methods, the maximum prediction error (over all $s_0$ sparse $\theta_0$) is lower bounded by (up to some constant) $\kappa \cdot \frac{\sigma^2 s_0^{1-\delta} \log(d)}{n}$.

Thus if we restrict ourselves to all computationally feasible $s_0$ sparse estimator, then the best achievable squared prediction error is of order $\kappa \cdot \frac{\sigma^2 s_0 \log(d)}{n}$.

**Upper bounds for iterative thresholding methods** In this section we establish prediction error bounds for iterative thresholding algorithm. First we provide some intuition on how to connect restricted optimality guarantee with statistical performance. It is well known that the global optimum of $\ell_0$-constrained least squared loss, i.e.

$$\widehat{\theta} \in \arg \min_{\|\theta\|_0 \le s_0} \|y - X\theta\|_2^2,$$

achieves a squared prediction error scaling as $\frac{\sigma^2 s_0 \log(d)}{n}$. For iterative thresholding algorithms, since we only have restricted optimality rather than global optimality, we are forced to work over a constraint at a larger sparsity $s \ge s_0$ to guarantee $\|y - X\widehat{\theta}\|_2^2 \le \min_{\|\theta\|_0 \le s_0} \|y - X\theta\|_2^2$. The statistical price one has to pay for this computational strategy is the inflation in noise level corresponding to the inflation in sparsity—that is, we have error on $s$ many nonzero coefficients, rather than $s_0$ many—so the final upper bound for prediction error would scale as $\frac{\sigma^2 s \log(d)}{n}$ instead of $\frac{\sigma^2 s_0 \log(d)}{n}$, where $s$ is chosen to be the smallest sparsity level that guaratees restricted optimality relative to the lower sparsity level $s' = s_0$. Now recall from Section 4 that, while hard thresholding offers restricted optimality guarantees at sparsity levels $s \sim \kappa^2 s_0$, the optimal and near-optimal thresholding operators (for example reciprocal thresholding and $\ell_{2/3}$ thresholding) improves this scaling to $s \sim \kappa s_0$. This allows us to improve the upper bound for squared prediction error from scaling as $\kappa^2$ to $\kappa$, when we switch our method from iterative hard thresholding, to iterative thresholding with an operator $\Psi_s$ that enjoys a lower relative concavity. Indeed in Jain et al. [2014], it is shown that iterative hard thresholding achieves a prediction error upper bounded by $\kappa^2 \cdot \frac{\sigma^2 s_0 \log(d)}{n}$. In view of our lower bound result Theorem 2, which states that the restircted optimality guarantee is tight, we postulate that the corresponding prediction error bound is also tight for iterative hard thresholding method.

Now we formulate this rigorously. Consider the iterative thresholding algorithm with some thresholding operator $\Psi_s$ applied to the objective function $\mathsf{f}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$, whose iteration takes the form

$$\widehat{\theta}_t = \Psi_s\big(\widehat{\theta}_{t-1} + \eta_t \cdot \frac{1}{n} X^\top (y - X\widehat{\theta}_{t-1})\big). \tag{22}$$

As usual, for the step size we may choose $\eta_t = 1/\beta$ if $\beta$ is known, or we may choose $\eta_t$ adaptively as in (6). We will work with any thresholding operator $\Psi_s$ satisfying

$$\gamma_{s,\rho}(\Psi_s) \le \rho \text{ for all } \rho \in (0, 1/2). \tag{23}$$

From Section 4, we see that on the one hand, this condition rules out hard thresholding and any continuous thresholding operator; on the other hand, it is satisfied by the reciprocal thresholding

operator, $\Psi_s^{\mathrm{RT}}$, by $\ell_q$ thresholding with $q = 2/3$, $\Psi_s^{\ell_{2/3}}$, and by any shrinkage operator $\Psi_{s;\sigma}$ where $\sigma(1) = 1/2$ and $\sigma$ satisfies the conditions of Lemma 5. We now present our result for this setting:

**Theorem 5.** *Suppose that $y = X\theta_0 + N(0, \sigma^2\mathbf{I}_n)$, where $\theta_0$ is $s_0$-sparse, and where $X \in \mathcal{X}(\alpha, \beta, s)$, where $s = C\kappa s_0$ for some $C > 2$. Suppose that $\Psi_s$ is any $s$-sparse thresholding operator satisfying* (23).

*Let $\widehat{\theta}_t$ be the estimate produced at step $t$ of the iterative thresholding algorithm* (22) *initialized at some $s$-sparse $\widehat{\theta}_0 \in \mathbb{R}^d$. Let $\tilde{\theta}_t \in \arg\min_{\theta \in \{\widehat{\theta}_1, \ldots, \widehat{\theta}_t\}} \frac{1}{2n}\|y - X\theta\|_2^2$, that is, $\tilde{\theta}_t$ is the best estimate seen before time $t$, relative to the loss function $\mathsf{f}(\theta) = \frac{1}{2n}\|y - X\theta\|_2^2$.*

*Then, for any $\delta > 0$ and any $t \geq 1$,*

$$\frac{1}{n}\|X(\tilde{\theta}_t - \theta_0)\|_2^2 \leq \kappa \cdot \frac{28C\sigma^2 s_0 \log(d)}{n} + \frac{12\sigma^2 \log(1/\delta)}{n} + \left(\frac{1 - 1/\kappa}{1 - 2/C\kappa}\right)^t \cdot 2\beta\|\widehat{\theta}_0 - \theta_0\|_2^2,$$

*with probability at least $1 - \delta$.*

Since $t$ can be taken to be large (each iteration is very cheap), the dominant term is the first one, so we essentially have

$$\frac{1}{n}\|X(\tilde{\theta}_t - \theta_0)\|_2^2 \lesssim \kappa \cdot \frac{\sigma^2 s_0 \log(d)}{n}.$$

Comparing with the upper bound for iterative hard thresholding, we see that we now attains the ideal $\kappa$, rather than $\kappa^2$, scaling.

**Comparison with Lasso** The Lasso estimate of $\theta_0$, given by the convex optimization problem

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2}\|y - X\theta\|_2^2 + \lambda\|\theta\|_1 \right\},$$

is proved in Bickel et al. [2009] to achieve a squared prediction error bounded as

$$\frac{1}{n}\|X(\widehat{\theta} - \theta_0)\|_2^2 \lesssim \kappa \cdot \frac{\sigma^2 s_0 \log(d)}{n} \tag{24}$$

with a penalty parameter value $\lambda \sim \sigma\sqrt{\frac{\log(d)}{n}}$, under the assumption that $X \in \mathcal{X}_{\mathrm{RE}}(\kappa, s_0)$. Compared with Lasso, due to Theorem 5, iterative thresholding algorithms with proper thresholding operators, for example the simple and efficient reciprocal thresholding, achieve the same squared prediction error bound. Moreover, both Lasso and iterative reciprocal thresholding method are guaranteed to give an estimator that is $\mathcal{O}(\kappa s_0)$ sparse (this sparsity level for Lasso is proved in Bickel et al. [2009, Eqn. (7.9)]), and thus nearly match the computational lower bound with a gap in sparsity. An open question for future work is whether the larger sparsity level, i.e. $\mathcal{O}(\kappa s_0)$ rather than $s_0$, is unavoidable to achieve the squared prediction error $\kappa \cdot \frac{\sigma^2 s_0 \log(d)}{n}$, or whether there may be an $\mathcal{O}(s_0)$-sparse and computationally efficient estimator that achieves this bound.

# 7 Discussion

Relative concavity offers a framework for comparing theoretical properties of thresholding operators. Under this framework, we find a general class of optimal and near-optimal thresholding operators, among which is the new reciprocal thresholding operator, an alternative to hard and soft thresholding with tighter theoretical guarantees that is able to achieve better dependence on condition number for sparse and low-rank optimization problems. Nonetheless, many open

questions remain for these problems. For example, our upper and lower bounds on $\lim_{t\to\infty} \mathsf{f}(x_t)$ are proved relative to a broad class of functions satisfying (restricted) convexity and smoothness properties, with no underlying statistical model. In a statistical framework, we may be able to make additional assumptions, for instance, assuming that $\nabla \mathsf{f}(y)$ is small at some highly sparse $y$ (e.g. if $y$ is the true model parameter vector, while $\mathsf{f}$ is the negative log-likelihood on the observed data)—is the relative concavity still necessary and sufficient for optimization guarantees, or would we observe different behavior of the various thresholding operators in this statistical setting? Furthermore, how does the choice of the thresholding operator interact with modifications of the gradient descent algorithm, such as decreasing step size, choosing the step size via backtracking or another adaptive method, acceleration of the gradient descent step, replacing gradients with stochastic gradients, or using second-order information? We hope to address these directions in future work.


## Acknowledgements

R.F.B. was partially supported by the National Science Foundation via grant DMS-1654076, and by an Alfred P. Sloan fellowship. The authors are grateful to Chao Gao for helpful discussions and feedback on this work.


# References


Rina Foygel Barber and Wooseok Ha. Gradient descent with nonconvex constraints: local concavity determines convergence. *arXiv preprint arXiv:1703.07755*, 2017.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

Peter J Bickel, Yaacov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Thomas Blumensath. Accelerated iterative hard thresholding. *Signal Processing*, 92(3):752–756, 2012.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

T Tony Cai, Xiaodong Li, and Zongming Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *The Annals of Statistics*, 44(5):2221–2251, 2016.

Rick Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10):707–710, 2007.

Yudong Chen and Martin J Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Simon Foucart and Ming-Jun Lai. Sparsest solutions of underdetermined linear systems via $\ell_q$-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3):395–407, 2009.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional M-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

Prateek Jain, Nikhil Rao, and Inderjit S Dhillon. Structured sparse regression via greedy hard thresholding. In *Advances in Neural Information Processing Systems*, pages 1516–1524, 2016.

Yoshiyuki Kabashima, Tadashi Wadayama, and Toshiyuki Tanaka. A typical reconstruction limit for compressed sensing based on $\ell_p$-norm minimization. *Journal of Statistical Mechanics: Theory and Experiment*, 2009(09):L09003, 2009.

Rajiv Khanna and Anastasios Kyrillidis. IHT dies hard: provable accelerated iterative hard thresholding. *arXiv preprint arXiv:1712.09379*, 2017.

Anastasios Kyrillidis and Volkan Cevher. Recipes on hard thresholding methods. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on*, pages 353–356. IEEE, 2011.

Anastasios Kyrillidis and Volkan Cevher. Matrix recipes for hard thresholding methods. *Journal of mathematical imaging and vision*, 48(2):235–265, 2014.

Ming-Jun Lai and Jingyue Wang. An unconstrained \ell_q minimization with 0q\leq1 for sparse solution of underdetermined linear systems. *SIAM Journal on Optimization*, 21(1):82–101, 2011.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

Nam Nguyen, Deanna Needell, and Tina Woolf. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63(11):6869–6895, 2017.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014.

Le Zheng, Arian Maleki, Haolei Weng, Xiaodong Wang, and Teng Long. Does $\ell_p$-minimization outperform $\ell_1$-minimization? *CoRR*, 2015.

# A Proofs

## A.1 Proofs of upper and lower bounds on convergence

In this section, we prove our upper and lower bounds on convergence for the sparse setting, Theorems 1 and 2. The results for the matrix setting, Theorems 3 and 4, are proved identically, so we do not give those proofs here.

*Proof of Theorem 1.* Fix any $t \in \{1, \ldots, T\}$. Since $x_{t-1}$ and $x_t$ are $s$-sparse by definition of the algorithm, and $\mathsf{f}$ satisfies $(\alpha, s)$-RSC and $(\beta, s)$-RSM, we have

$$\mathsf{f}(y) \geq \mathsf{f}(x_{t-1}) + \langle \nabla \mathsf{f}(x_{t-1}), y - x_{t-1} \rangle + \frac{\alpha}{2}\|y - x_{t-1}\|_2^2,$$
$$\mathsf{f}(x_t) \leq \mathsf{f}(x_{t-1}) + \langle \nabla \mathsf{f}(x_{t-1}), x_t - x_{t-1} \rangle + \frac{1}{2\eta_t}\|x_t - x_{t-1}\|_2^2,$$

where $\eta_t = \eta = \frac{1}{\beta}$ for the fixed step size algorithm (5), or $\eta_t$ is the adaptive step size defined in the algorithm (6)—note that in this second case, since $\mathsf{f}$ satisfies $(\beta, s)$-RSM, we see that $\eta_t \geq \frac{1}{\beta}$ since the step size is chosen by backtracking. Combining these two inequalities, we obtain

$$\mathsf{f}(x_t) - \mathsf{f}(y) \leq \langle \nabla \mathsf{f}(x_{t-1}), x_t - y \rangle + \frac{1}{2\eta_t}\|x_t - x_{t-1}\|_2^2 - \frac{\alpha}{2}\|y - x_{t-1}\|_2^2. \tag{25}$$

We can also calculate

$$\begin{aligned}
&\frac{1}{2\eta_t}\|x_t - y\|_2^2 \\
&= \frac{1}{2\eta_t}\|x_{t-1} - y\|_2^2 - \frac{1}{2\eta_t}\|x_t - x_{t-1}\|_2^2 + \frac{1}{\eta_t}\langle x_{t-1} - x_t, y - x_t \rangle \\
&= \frac{1}{2\eta_t}\|x_{t-1} - y\|_2^2 - \frac{1}{2\eta_t}\|x_t - x_{t-1}\|_2^2 + \frac{1}{\eta_t}\langle (x_{t-1} - \eta_t \nabla \mathsf{f}(x_{t-1})) - x_t, y - x_t \rangle - \langle \nabla \mathsf{f}(x_{t-1}), x_t - y \rangle \\
&\leq \frac{1}{2\eta_t}\|x_{t-1} - y\|_2^2 - \frac{1}{2\eta_t}\|x_t - x_{t-1}\|_2^2 + \frac{1}{\eta_t} \cdot \gamma_{s,\rho}(\Psi_s) \cdot \|x_t - y\|_2^2 - \langle \nabla \mathsf{f}(x_{t-1}), x_t - y \rangle,
\end{aligned} \tag{26}$$

where the last step applies the definition of restricted concavity, since $x_t = \Psi_s\left(x_{t-1} - \eta_t \nabla \mathsf{f}(x_{t-1})\right)$ by definition of the algorithm.

Combining steps (25) and (26), then,

$$\mathsf{f}(x_t) - \mathsf{f}(y) \leq \frac{1}{2\eta_t}\left[(1 - \eta_t\alpha)\|x_{t-1} - y\|_2^2 - (1 - 2\gamma_{s,\rho}(\Psi_s))\|x_t - y\|_2^2\right].$$

Since $\eta_t\alpha \geq \frac{1}{\beta} \cdot \alpha = \frac{1}{\kappa}$, this implies

$$\mathsf{f}(x_t) - \mathsf{f}(y) \leq \frac{1}{2\eta_t}\left[\left(1 - \frac{1}{\kappa}\right)\|x_{t-1} - y\|_2^2 - (1 - 2\gamma_{s,\rho}(\Psi_s))\|x_t - y\|_2^2\right].$$

Taking a weighted sum over $t = 1, \ldots, T$, we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} 2\eta_t \left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T-t} \cdot (\mathsf{f}(x_t) - \mathsf{f}(y)) \\
&\leq \sum_{t=1}^{T} \left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T-t} \cdot \left[\left(1-\frac{1}{\kappa}\right)\|x_{t-1}-y\|_2^2 - (1-2\gamma_{s,\rho}(\Psi_s))\|x_t-y\|_2^2\right] \\
&= (1-2\gamma_{s,\rho}(\Psi_s)) \cdot \sum_{t=1}^{T} \left[\left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T-t+1}\|x_{t-1}-y\|_2^2 - \left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T-t}\|x_t-y\|_2^2\right] \\
&= (1-2\gamma_{s,\rho}(\Psi_s)) \cdot \left[\left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T}\|x_0-y\|_2^2 - \|x_T-y\|_2^2\right] \\
&\leq \left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T}\|x_0-y\|_2^2,
\end{aligned}$$

where the next-to-last step simply cancels terms in the telescoping sum. After rescaling, we have

$$\frac{\sum_{t=1}^{T} 2\eta_t \left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T-t} \mathsf{f}(x_t)}{\sum_{t=1}^{T} 2\eta_t \left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T-t}} \leq \mathsf{f}(y) + \frac{\left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T}\|x_0-y\|_2^2}{\sum_{t=1}^{T} 2\eta_t \left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T-t}}.$$

The left-hand side is a weighted average of $\mathsf{f}(x_1), \mathsf{f}(x_2), \ldots, \mathsf{f}(x_T)$, and is therefore lower-bounded by $\min_{t=1,\ldots,T} \mathsf{f}(x_t)$, while the denominator on the right-hand side is lower-bounded as

$$\sum_{t=1}^{T} 2\eta_t \left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T-t} \geq 2\eta_T \geq \frac{2}{\beta}.$$

After simplifying, we therefore have

$$\min_{t=1,\ldots,T} \mathsf{f}(x_t) \leq \mathsf{f}(y) + \frac{\beta}{2} \cdot \left(\frac{1-1/\kappa}{1-2\gamma_{s,\rho}(\Psi_s)}\right)^{T} \cdot \|x_0-y\|_2^2,$$

as desired. □

*Proof of Theorem 2.* By definition of $\gamma_{s,\rho}(\Psi_s)$, for any $\delta > 0$, there exist some $s'$-sparse $y \in \mathbb{R}^d$ and some $z \in \mathbb{R}^d$ such that $x = \Psi_s(z) \neq y$ and

$$\langle y - x, z - x\rangle \geq \gamma_{s,\rho}(\Psi_s) \cdot \|y-x\|_2^2 \cdot (1-\delta).$$

Let $U \in \mathbb{R}^{d\times d}$ be any orthogonal matrix with its first column equal to $\frac{y-x}{\|y-x\|_2}$. We now define an objective function as

$$\mathsf{f}(w) = -\beta\langle z-x, w-x\rangle + \frac{1}{2}(w-x)^\top U D U^\top (w-x) \text{ where } D = \begin{pmatrix} \alpha & 0 & \ldots & 0 \\ 0 & a_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & a_d \end{pmatrix},$$

for some $a_2, \ldots, a_d \in [\alpha, \beta]$. Clearly, $\mathsf{f}$ satisfies $(\alpha, s)$-RSC and $(\beta, s)$-RSM. Next, we can check

that $\mathsf{f}(x) = 0$, while

$$\begin{aligned}\mathsf{f}(y) &= -\beta\langle z-x, y-x\rangle + \frac{1}{2}(y-x)^\top UDU^\top (y-x) \\ &= -\beta\langle z-x, y-x\rangle + \frac{\alpha}{2}\|y-x\|_2^2 \\ &\le -\beta\gamma_{s,\rho}(\Psi_s)\cdot\|y-x\|_2^2\cdot(1-\delta) + \frac{\alpha}{2}\|y-x\|_2^2 \\ &= -\beta\|y-x\|_2^2\cdot\left(\gamma_{s,\rho}(\Psi_s)\cdot(1-\delta) - \frac{1}{2\kappa}\right),\end{aligned}$$

where the first step uses the definition of $U$, while the inequality follows from the definition of $x, y, z$. Since $\gamma_{s,\rho}(\Psi_s) > \frac{1}{2\kappa}$ by assumption, and $\delta$ can be chosen to be arbitrarily small, we therefore have $\mathsf{f}(y) < 0 = \mathsf{f}(x)$.

Finally, computing $\nabla\mathsf{f}(w) = -\beta(z-w) + UDU^\top(w-x)$, suppose that we run the iterated thresholding algorithm (5) with step size $\eta = \frac{1}{\beta}$, initialized at the point $x_0 = x$. Since we have $\nabla\mathsf{f}(x) = -\beta(z-x)$, the first update step is given by

$$x_1 = \Psi_s\left(x - \frac{1}{\beta}\nabla\mathsf{f}(x)\right) = \Psi_s(z) = x.$$

This proves that $x$ is a stationary point of the algorithm—in other words, if the algorithm is initalized at $x_0 = x$, then $x_t = x$ for all $t \ge 1$. Therefore, $\lim_{t\to\infty}\mathsf{f}(x_t) = \mathsf{f}(x) > \mathsf{f}(y)$, as desired. □

## A.2 Proofs for calculating relative concavity

In this section we give the proofs for all lemmas from Sections 4 and 5, calculating upper and lower bounds on relative concavity in the vector and matrix setting.

*Proof of Lemma 1.* Fix any $z \in \mathbb{R}^d$ and any $s'$-sparse $y \in \mathbb{R}^d$. Let $x = \Psi_s^{\mathrm{HT}}(z)$. Let $S = \mathrm{Support}(x)$ and $S' = \mathrm{Support}(y)$. We can write

$$\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} = \frac{\langle y_{S'\setminus S}, z_{S'\setminus S}\rangle}{\|y_{S'\setminus S}\|_2^2 + \|(y-z)_S\|_2^2}, \tag{27}$$

since $x_S = z_S$ by definition of hard thresholding. Next, let $\tau = \max_{i\notin S}|z_i|$, i.e. the $(s+1)$-st largest magnitude entry of $z$. Then $|z_i| \ge \tau$ for all $i \in S$ by definition of the method, and so $|(y-z)_i| \ge \tau$ for all $i \in S\setminus S'$. Therefore,

$$\|(y-z)_S\|_2^2 \ge \tau^2\cdot(s-\ell),$$

where $\ell = |S\cap S'|$. We also have

$$\langle y_{S'\setminus S}, z_{S'\setminus S}\rangle \le \|y_{S'\setminus S}\|_2\cdot\tau\sqrt{s'-\ell},$$

since $\|z_{S'\setminus S}\|_2 \le \tau\sqrt{|S'\setminus S|} \le \tau\sqrt{s'-\ell}$. Combining everything and returning to (27), we have

$$\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} \le \frac{\|y_{S'\setminus S}\|_2\cdot\tau\sqrt{s'-\ell}}{\|y_{S'\setminus S}\|_2^2 + \tau^2\cdot(s-\ell)} \le \max_{t\ge 0}\frac{t\sqrt{s'-\ell}}{t^2+s-\ell},$$

where for the last step we consider $t = \frac{\|y_{S'\setminus S}\|_2}{\tau}$. This quantity is maximized at $t = \sqrt{s-\ell}$, so we obtain

$$\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} \le \frac{\sqrt{s-\ell}\cdot\sqrt{s'-\ell}}{2(s-\ell)} = \frac{1}{2}\sqrt{\frac{s'-\ell}{s-\ell}}.$$

Finally, by definition, we must have $\ell \in \{0, 1, \ldots, s'\}$, so we obtain

$$\frac{\langle y - x, z - x\rangle}{\|y - x\|_2^2} \leq \max_{\ell \in \{0,1,\ldots,s'\}} \frac{1}{2}\sqrt{\frac{s' - \ell}{s - \ell}} = \frac{1}{2}\sqrt{\rho},$$

where the maximum is obtained at $\ell = 0$. This proves that $\gamma_{s,\rho}(\Psi_s^{\mathrm{HT}}) \leq \frac{\sqrt{\rho}}{2}$.

To prove a matching lower bound, consider $z = \mathbf{1}_d$. Then $x = \Psi_s^{\mathrm{HT}}(z) = \mathbf{1}_S$, for some subset $S \subset \{1, \ldots, d\}$ of cardinality $|S| = s$. Let $S' \subset \{1, \ldots, d\} \backslash S$ be a disjoint set of cardinality $|S'| = s'$ (recall that we have assumed $s + s' \leq d$), and let $y = \frac{1}{\sqrt{\rho}} \cdot \mathbf{1}_{S'}$. Then

$$\frac{\langle y - x, z - x\rangle}{\|y - x\|_2^2} = \frac{\frac{1}{\sqrt{\rho}} \cdot s'}{\frac{1}{\rho} \cdot s' + s} = \frac{\sqrt{\rho}}{2},$$

thus proving that $\gamma_{s,\rho}(\Psi_s^{\mathrm{HT}}) \geq \frac{\sqrt{\rho}}{2}$. □

*Proof of Lemma 2.* We consider two cases. If there exists $z \neq 0$ such that $\Psi_s(z) = 0$, then fix any such $z$ and fix an index $i$ such that $z_i \neq 0$. Let $y = \epsilon \cdot \mathrm{sign}(z_i) \cdot \mathbf{e}_i$, where $\mathbf{e}_i$ is the vector with a 1 in entry $i$ and zeros elsewhere. $y$ is $s'$-sparse since $s' \geq 1$. Then

$$\frac{\langle y - \Psi_s(z), z - \Psi_s(z)}{\|y - \Psi_s(z)\|_2^2} = \frac{\langle y, z\rangle}{\|y\|_2^2} = \frac{\epsilon |z_i|}{\epsilon^2} = \frac{|z_i|}{\epsilon}.$$

Since $|z_i| > 0$ and $\epsilon > 0$ can be taken to be arbitrarily small, this shows that $\gamma_{s,\rho}(\Psi_s) = \infty \geq 1$.

On the other hand, if $\Psi_s(z) \neq 0$ for any $z \neq 0$, then define $g : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ by $g(x) = \frac{\Psi_s(x)}{\|\Psi_s(x)\|_2}$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. Since $\|\Psi_s(x)\|_2$ is a continuous function on a compact space and takes only positive values, $\|\Psi_s(x)\|_2$ is lower-bounded by a positive value, which then implies $g$ is continuous. Since $g(x)$ inherits the sparsity of $\Psi_s(x)$ for all $x$, we see that $g(\mathbb{S}^{d-1}) \subset \mathbb{S}^{d-1} \backslash \{x_0\}$, where $x_0 = \mathbf{1}_d/\sqrt{d}$ is a dense point on the sphere. Now let $h$ be a homeomorphism from $\mathbb{S}^{d-1} \backslash \{x_0\}$ to $\mathbb{R}^{d-1}$ (for example, take $h$ to be the stereographic projection from the point $x_0$). Then $h \circ g : \mathbb{S}^{d-1} \to \mathbb{R}^{d-1}$ is continuous. By the Borsuk-Ulam theorem, there exist two antipodal point being mapped to the same point, i.e. there exists $z \in \mathbb{S}^{d-1}$ such that $h \circ g(z) = h \circ g(-z)$, and thus $g(z) = g(-z)$ since $h$ is bijective. Now, there are two possibities—either $\langle z, g(z)\rangle \leq 0$, or alternately $\langle z, g(z)\rangle > 0$ in which case $\langle -z, g(z)\rangle = \langle -z, g(-z)\rangle < 0$. Replacing $z$ with $-z$ if needed, then, we have some $z \in \mathbb{S}^{d-1}$ such that $\langle z, g(z)\rangle \leq 0$. Then by definition of $g$, we have $\langle z, \Psi_s(z)\rangle \leq 0$. Setting $y = \mathbf{0}_d$, we then calculate

$$\frac{\langle y - \Psi_s(z), z - \Psi_s(z)\rangle}{\|y - \Psi_s(z)\|_2^2} = \frac{\|\Psi_s(z)\|_2^2 - \langle z, \Psi_s(z)\rangle}{\|\Psi_s(z)\|_2^2} \geq \frac{\|\Psi_s(z)\|_2^2 - 0}{\|\Psi_s(z)\|_2^2} = 1,$$

proving that $\gamma_{s,\rho}(\Psi_s) \geq 1$, as desired. □

*Proof of Lemmas 6 and 4.* These lemmas are special cases of the general result, Lemma 5, proved below. □

*Proof of Lemma 3.* Let $z = \mathbf{1}_d$ and let $x = \Psi_s(z)$. Let $S = \mathrm{Support}(x)$, with $|S| \leq s$, and let $S' \subset \{1, \ldots, d\} \backslash S$ be any set disjoint from $S$, with cardinality $|S'| = s'$ (recall that we have assumed $s + s' \leq d$). Let $y = t \cdot \mathbf{1}_{S'}$, where

$$t = \frac{r}{\rho}\left(1 - r + \sqrt{r^2 - 2r + 1 + \rho}\right), \text{ for } r = \frac{\|x\|_2}{\sqrt{s}}.$$

Then $\|y\|_0 = s'$, and we can calculate

$$\begin{aligned}
\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} &= \frac{\langle y,z\rangle - \langle x,z\rangle + \|x\|_2^2}{\|y\|_2^2 + \|x\|_2^2} \text{ since } x,y \text{ have disjoint supports } S \text{ and } S' \\
&= \frac{t\cdot s' - \langle x, \mathbf{1}_d\rangle + \|x\|_2^2}{t^2\cdot s' + \|x\|_2^2} \text{ by definition of } y \text{ and } z \\
&\geq \frac{t\cdot s' - \sqrt{s}\cdot\|x\|_2 + \|x\|_2^2}{t^2\cdot s' + \|x\|_2^2} \text{ since } x \text{ is } s\text{-sparse} \qquad (28) \\
&= \frac{t\cdot s' - s\cdot r + s\cdot r^2}{t^2\cdot s' + s\cdot r^2} \\
&= \frac{t\cdot\rho - r + r^2}{t^2\cdot\rho + r^2}.
\end{aligned}$$

Plugging in the value of $t$ that we chose above, we continue:

$$\begin{aligned}
\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} &\geq \frac{\frac{r}{\rho}\left(1-r+\sqrt{r^2-2r+1+\rho}\right)\cdot\rho - r + r^2}{\frac{r^2}{\rho^2}\left(1-r+\sqrt{r^2-2r+1+\rho}\right)^2\cdot\rho + r^2} \\
&= \frac{r\sqrt{r^2-2r+1+\rho}}{\frac{r^2}{\rho}\left(1-r+\sqrt{r^2-2r+1+\rho}\right)^2 + r^2} \\
&= \frac{\rho}{r\left(2\sqrt{r^2-2r+1+\rho}+2(1-r)\right)},
\end{aligned}$$

where the last few steps are just simplifying the expression. Next, we consider the denominator. It can easily be verified that

$$r\left(2\sqrt{r^2-2r+1+\rho}+2(1-r)\right) \leq 1+\rho$$

for all $r \geq 0$, which we check by verifying that the left-hand side is maximized when $r = \frac{1+\rho}{2}$. Therefore,

$$\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} \geq \frac{\rho}{1+\rho},$$

which proves that

$$\gamma_{s,\rho}(\Psi_s) \geq \frac{\rho}{1+\rho},$$

as desired. □

*Proof of Lemma 5.* We first show the upper bound. Fix any $z \in \mathbb{R}^d$ and any $s'$-sparse $y \in \mathbb{R}^d$. Let $x = \Psi_{s;\sigma}(z)$. Let $S = \text{Support}(x)$ and $S' = \text{Support}(y)$, and let $\ell = |S \cap S'|$. Then we have

$$\begin{aligned}
\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} &= \frac{\langle (y-x)_{S'}, (z-x)_{S'}\rangle - \langle x_{S\setminus S'}, (z-x)_{S\setminus S'}\rangle}{\|(y-x)_{S'}\|_2^2 + \|x_{S\setminus S'}\|_2^2} \\
&\leq \frac{\|(y-x)_{S'}\|_2\|(z-x)_{S'}\|_2 - \langle x_{S\setminus S'}, (z-x)_{S\setminus S'}\rangle}{\|(y-x)_{S'}\|_2^2 + \|x_{S\setminus S'}\|_2^2}.
\end{aligned}$$

Let $\tau = \max_{i\notin S}|z_i|$, i.e. the thresholding level. Due to the definition of $\Psi_{s;\sigma}$ and the assumptions on $\sigma$, it is direct to verify the following bounds: if $i \notin S$, then $|(z-x)_i| = |z_i| \leq \tau$; if $i \in S$, then

$|(z-x)_i| \le \tau\sigma(1)$, $|x_i| \ge \tau(1-\sigma(1))$, and $x_i(z-x)_i \ge \tau^2\sigma(1)\big(1-\sigma(1)\big)$. Plugging these bounds back in our calculation above, we get:

$$\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} \le \frac{\|(y-x)_{S'}\|_2\sqrt{(s'-\ell)\cdot\tau^2 + \ell\cdot\tau^2\sigma(1)^2} - (s-\ell)\cdot\tau^2\sigma(1)\big(1-\sigma(1)\big)}{\|(y-x)_{S'}\|_2^2 + (s-\ell)\cdot\tau^2\big(1-\sigma(1)\big)^2}$$
$$\le \max_{t\ge 0} \frac{t\sqrt{\frac{s'-\ell}{s-\ell} + \frac{\ell}{s-\ell}\cdot\sigma(1)^2} - \sigma(1)\big(1-\sigma(1)\big)}{t^2 + \big(1-\sigma(1)\big)^2},$$

where the last step holds by considering $t = \frac{\|(y-x)_{S'}\|_2}{\tau\sqrt{s-\ell}}$. Next, we can calculate

$$\max_{\ell\in\{0,\dots,s'\}} \sqrt{\frac{s'-\ell}{s-\ell} + \frac{\ell}{s-\ell}\cdot\sigma(1)^2} = \max_{\ell\in\{0,\dots,s'\}} \sqrt{\frac{s' - \big(1-\sigma(1)^2\big)\ell}{s-\ell}} = \sqrt{\frac{\rho}{\min\big\{1, (1-\rho)/\sigma(1)^2\big\}}},$$

where the maximum is attained at $\ell = 0$ if $\rho \le 1-\sigma(1)^2$, and at $\ell = s'$ otherwise. It therefore follows that

$$\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} \le \max_{t\ge 0} \frac{t\sqrt{\frac{\rho}{\min\{1,(1-\rho)/\sigma(1)^2\}}} - \sigma(1)\big(1-\sigma(1)\big)}{t^2 + \big(1-\sigma(1)\big)^2}$$
$$= \frac{\frac{\rho}{\min\{1,(1-\rho)/\sigma(1)^2\}}}{2\sigma(1)\big(1-\sigma(1)\big)\left(1 + \sqrt{1 + \frac{\rho/\sigma(1)^2}{\min\{1,(1-\rho)/\sigma(1)^2\}}}\right)}, \quad (29)$$

where to compute the last step we can check that the maximum is achieved at

$$t = \frac{\sigma(1)\big(1-\sigma(1)\big) + \sqrt{\sigma(1)^2\big(1-\sigma(1)\big)^2 + \frac{\rho\big(1-\sigma(1)\big)^2}{\min\{1,(1-\rho)/\sigma(1)^2\}}}}{\sqrt{\frac{\rho}{\min\{1,(1-\rho)/\sigma(1)^2\}}}}.$$

This proves the upper bound. To prove the lower bound, we simply choose $y$ and $z$ so that the inequalities above become equalities. Set $z = \mathbf{1}_d$ and $x = \Psi_s^{\mathrm{RT}}(z)$, and let $S = \mathrm{Support}(x)$. Due to the definition of $\Psi_{s;\sigma}$, we see that $x = (1-\sigma(1))\cdot\mathbf{1}_S$. To construct $y$, we consider two cases. If $\rho \le 1-\sigma(1)^2$, we let $S' \subset \{1,\dots,d\}\backslash S$ be any set disjoint from $S$ with cardinality $|S'| = s'$ (recall that $s + s' \le d$ by assumption). Then let $y = \frac{t}{\sqrt{\rho}}\cdot\mathbf{1}_{S'}$, where $t \ge 0$ is arbitrary, so that we have

$$\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} = \frac{\frac{t}{\sqrt{\rho}}\cdot s' - \sigma(1)\big(1-\sigma(1)\big)\cdot s}{\frac{t^2}{\rho}\cdot s' + \big(1-\sigma(1)\big)^2\cdot s} = \frac{t\sqrt{\rho} - \sigma(1)\big(1-\sigma(1)\big)}{t^2 + \big(1-\sigma(1)\big)^2}.$$

Alternately, if $\rho > 1-\sigma(1)^2$, let $S' \subset S$ be any set of cardinality $|S'| = s'$, and set $y = \Big(1-\sigma(1) + t\sqrt{\frac{1-\rho}{\rho}}\Big)\cdot\mathbf{1}_{S'}$, where again $t > 0$ is arbitrary. For this second case, we calculate

$$\frac{\langle y-x, z-x\rangle}{\|y-x\|_2^2} = \frac{t\sigma(1)\sqrt{\frac{1-\rho}{\rho}}\cdot s' - \sigma(1)\big(1-\sigma(1)\big)\cdot(s-s')}{t^2\cdot\frac{1-\rho}{\rho}\cdot s' + \big(1-\sigma(1)\big)^2\cdot(s-s')} = \frac{t\sqrt{\frac{\rho}{(1-\rho)/\sigma(1)^2}} - \sigma(1)\big(1-\sigma(1)\big)}{t^2 + \big(1-\sigma(1)\big)^2}.$$

Combining the two cases, and recalling that $t \ge 0$ is arbitrary, we see that

$$\gamma_{s,\rho}(\Psi_{s;\sigma}) \ge \max_{t\ge 0} \frac{t\sqrt{\frac{\rho}{\min\{1,(1-\rho)/\sigma(1)^2\}}} - \sigma(1)\big(1-\sigma(1)\big)}{t^2 + \big(1-\sigma(1)\big)^2},$$

which matches the upper bound calculated in (29) above. □

*Proof of Lemma 7.* Without loss of generality, let $n \geq m$. Let $Z = \begin{pmatrix} \mathbf{I}_m \\ \mathbf{0}_{(n-m)\times m} \end{pmatrix}$, and let $X = \widetilde{\Psi}_s(Z)$. Let $X = UDV^\top$ be a singular value decomposition of $X$, with $U \in \mathbb{R}^{n\times s}, V \in \mathbb{R}^{m\times s}$. Let $V_\perp \in \mathbb{R}^{m\times s'}$ be an orthonormal matrix that is orthogonal to $V$ (recall that $s + s' \leq m$ by assumption), and let $Y = t \cdot \begin{pmatrix} V_\perp V_\perp^\top \\ \mathbf{0}_{(n-m)\times m} \end{pmatrix}$, for some $t \geq 0$. Then $\mathrm{rank}(Y) = s'$, and we can calculate

$$\begin{aligned}
\frac{\langle Y - X, Z - X\rangle}{\|Y - X\|_{\mathrm{F}}^2} &= \frac{\langle Y, Z\rangle - \langle X, Z\rangle + \|X\|_{\mathrm{F}}^2}{\|Y\|_{\mathrm{F}}^2 + \|X\|_{\mathrm{F}}^2} \text{ since } X \text{ and } Y \text{ have orthogonal row spaces by def. of } V_\perp \\
&= \frac{t \cdot s' - \|X\|_* \|Z\| + \|X\|_{\mathrm{F}}^2}{t^2 \cdot s' + \|X\|_{\mathrm{F}}^2} \text{ by def. of } Y \text{ and } Z \text{ (here } \|\cdot\|_* \text{ is the nuclear norm)} \\
&\geq \frac{t \cdot s' - \sqrt{s} \cdot \|X\|_{\mathrm{F}} + \|X\|_{\mathrm{F}}^2}{t^2 \cdot s' + \|X\|_{\mathrm{F}}^2} \text{ since } \mathrm{rank}(X) \leq s.
\end{aligned}$$

Comparing to (28), we see that the remainder of the argument is identical to the proof of Lemma 3. □

*Proof of Lemma 8.* First, fix any $y, z \in \mathbb{R}^d$. Let $Y = \mathrm{diag}(y)$ and $Z = \mathrm{diag}(z)$, so that $\widetilde{\Psi}_s(Z) = \mathrm{diag}(\Psi_s(z))$ and $\mathrm{rank}(Y) = \|y\|_0$. Then we trivially have $\frac{\langle Y - \widetilde{\Psi}_s(Z), Z - \widetilde{\Psi}_s(Z)\rangle}{\|Y - \widetilde{\Psi}_s Z\|_{\mathrm{F}}^2} = \frac{\langle y - \Psi_s(z), z - \Psi_s(z)\rangle}{\|y - \Psi_s z\|_2^2}$, and maximizing over all $y, z$ yields the restricted concavity, $\gamma_{s,\rho}(\Psi_s)$. This proves that $\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) \geq \gamma_{s,\rho}(\Psi_s)$.

Next we show the reverse inequality. Consider any $Y, Z \in \mathbb{R}^{n\times m}$ with $\mathrm{rank}(Y) \leq s'$, and let $X = \widetilde{\Psi}_s(Z) = U \cdot \mathrm{diag}(\Psi_s(d)) \cdot V^\top$, where $Z = U \cdot \mathrm{diag}(d) \cdot V^\top$ is the singular value decomposition. We want to prove the claim that

$$\langle Y - X, Z - X\rangle \leq \gamma_{s,\rho}(\Psi_s)\|Y - X\|_{\mathrm{F}}^2.$$

In other words, defining

$$\mathsf{h}(Y) = \gamma_{s,\rho}(\Psi_s)\|Y - X\|_{\mathrm{F}}^2 - \langle Y - X, Z - X\rangle = \gamma_{s,\rho}(\Psi_s)\left\|Y - \left(X + \frac{Z - X}{2\gamma_{s,\rho}(\Psi_s)}\right)\right\|_{\mathrm{F}}^2 - \frac{\|Z - X\|_{\mathrm{F}}^2}{4\gamma_{s,\rho}(\Psi_s)},$$

we'd like to show that $\mathsf{h}(Y) \geq 0$ for all rank-$s'$ matrices $Y$. Now, by definition of $X$, we can see that $U$ and $V$ are the left and right singular vector matrices for $X + \frac{Z-X}{2\gamma_{s,\rho}(\Psi_s)}$, and therefore $\mathsf{h}(Y)$ is minimized by some matrix $Y$ of the form $Y = U \cdot \mathrm{diag}(y) \cdot V^\top$, for some $s'$-sparse vector $y$. Now, for any matrix of this form, we have

$$\begin{aligned}
\langle Y - X, Z - X\rangle &= \langle U \cdot \mathrm{diag}(y) \cdot V^\top - U \cdot \mathrm{diag}(\Psi_s(d)) \cdot V^\top, U \cdot \mathrm{diag}(d) \cdot V^\top - U \cdot \mathrm{diag}(\Psi_s(d)) \cdot V^\top\rangle \\
&= \langle y - \Psi_s(d), d - \Psi_s(d)\rangle \leq \gamma_{s,\rho}(\Psi_s)\|y - \Psi_s(d)\|_2^2 = \gamma_{s,\rho}(\Psi_s)\|Y - X\|_{\mathrm{F}}^2,
\end{aligned}$$

by using the definition of relative concavity for sparse vectors. This proves that

$$\min_{\mathrm{rank}(Y)\leq s'} \mathsf{h}(Y) = \min_{Y = U\cdot\mathrm{diag}(y)\cdot V^\top, \|y\|_0 \leq s} \mathsf{h}(Y) \geq 0,$$

thus proving that $\widetilde{\gamma}_{s,\rho}(\widetilde{\Psi}_s) \leq \gamma_{s,\rho}(\Psi_s)$, as desired. □

## A.3 Proofs for prediction error in linear regression

In this section we prove our prediction error bounds for the linear regression setting.

*Proof of Theorem 5.* Since $s = C\kappa s_0$ and so our sparsity ratio is $\rho = \frac{1}{C\kappa} \le \frac{1}{2}$, Lemma 5 with the conditions on $\sigma$ proves that $\gamma_{s,\rho}(\Psi_{s;\sigma}) \le \rho = \frac{1}{C\kappa}$. Since this is strictly smaller than $\frac{1}{2\kappa}$, Theorem 1 proves that

$$\mathsf{f}(\tilde{\theta}_t) \le \mathsf{f}(\theta_0) + \left(\frac{1-1/\kappa}{1-2/C\kappa}\right)^t \cdot \frac{\beta}{2}\|\widehat{\theta}_0 - \theta_0\|_2^2.$$

Next, recalling the definition of $\mathsf{f}(\theta)$, this is equivalent to

$$\frac{1}{2n}\|\sigma z - X(\tilde{\theta}_t - \theta_0)\|_2^2 \le \frac{1}{2n}\|\sigma z\|_2^2 + \left(\frac{1-1/\kappa}{1-2/C\kappa}\right)^t \cdot \frac{\beta}{2}\|\widehat{\theta}_0 - \theta_0\|_2^2,$$

where $z \sim N(0, \mathbf{I}_n)$ and $y = X\theta_0 + \sigma z$. Rearranging terms,

$$\frac{1}{2n}\|X(\tilde{\theta}_t - \theta_0)\|_2^2 \le \frac{\sigma}{n}\langle z, X(\widehat{\theta}_t - \theta_0)\rangle + \left(\frac{1-1/\kappa}{1-2/C\kappa}\right)^t \cdot \frac{\beta}{2}\|\widehat{\theta}_0 - \theta_0\|_2^2.$$

Now, by Lemma 9 below, with probability at least $1-\delta$, we have

$$\begin{aligned}\langle z, X(\tilde{\theta}_t - \theta_0)\rangle &\le \|X(\tilde{\theta}_t - \theta_0)\|_2 \cdot \sqrt{7s\log(d) + 3\log(1/\delta)} \\ &\le \frac{1}{4\sigma}\|X(\tilde{\theta}_t - \theta_0)\|_2^2 + \sigma(7s\log(d) + 3\log(1/\delta)),\end{aligned}$$

and so combining everything,

$$\frac{1}{2n}\|X(\tilde{\theta}_t - \theta_0)\|_2^2 \le \frac{1}{4n}\|X(\tilde{\theta}_t - \theta_0)\|_2^2 + \sigma^2 \cdot \frac{7s\log(d) + 3\log(1/\delta)}{n} + \left(\frac{1-1/\kappa}{1-2/C\kappa}\right)^t \cdot \frac{\beta}{2}\|\widehat{\theta}_0 - \theta_0\|_2^2.$$

Rearranging terms, then,

$$\frac{1}{n}\|X(\tilde{\theta}_t - \theta_0)\|_2^2 \le \sigma^2 \cdot \frac{28s\log(d) + 12\log(1/\delta)}{n} + \left(\frac{1-1/\kappa}{1-2/C\kappa}\right)^t \cdot 2\beta\|\widehat{\theta}_0 - \theta_0\|_2^2.$$

Plugging in $s = C\kappa s_0$, this proves the theorem. □

**Lemma 9.** *Fix any sparsity level $s$, dimension $d \ge 3$, and sample size $n$. Fix any $s$-sparse $\theta_0 \in \mathbb{R}^d$, and any matrix $X \in \mathbb{R}^{n\times d}$ such that $X \in \mathcal{X}(\alpha, \beta, s)$ for some parameters $1 \le \alpha \le \beta$. Let $z \sim N(0, \mathbf{I}_n)$. Then for any $\delta > 0$,*

$$\mathbb{P}\left\{\langle z, X(\theta - \theta_0)\rangle \le \|X(\theta - \theta_0)\|_2 \cdot \sqrt{7s\log(d) + 3\log(1/\delta)} \text{ for all } s\text{-sparse } \theta \in \mathbb{R}^d\right\} \ge 1 - \delta.$$

*Proof of Lemma 9.* Let $A_0 \subset \{1, \dots, d\}$ be the support of $\theta_0$. We take a union bound over all sets $A \subset \{1, \dots, d\}$ of size $|A| = s$. First, for any fixed $A$, let $U^A \in \mathbb{R}^{n\times|A\cup A_0|}$ be an orthogonal basis for the column space of $X_{A\cup A_0} = (X_{ij})j \in A \cup A_0 \in \mathbb{R}^{n\times|A\cup A_0|}$. Then

$$\begin{aligned}\langle z, X(\theta - \theta_0)\rangle &= \langle z, X_{A\cup A_0}(\theta - \theta_0)_{A\cup A_0}\rangle = \langle U^A U^{A\top} z, X_{A\cup A_0}(\theta - \theta_0)_{A\cup A_0}\rangle \\ &\le \|U^A U^{A\top} z\|_2 \|X_{A\cup A_0}(\theta - \theta_0)_{A\cup A_0}\|_2 = \|U^{A\top} z\|_2 \|X_{A\cup A_0}(\theta - \theta_0)_{A\cup A_0}\|_2.\end{aligned}$$

Next, $\|U^{A\top} z\|_2^2 \sim \chi^2_{|A\cup A_0|} \le \chi^2_{2s}$. By Laurent and Massart [2000, Lemma 1], then,

$$\mathbb{P}\left\{\|U^{A\top} z\|_2^2 \ge 2s + 2\sqrt{2st} + 2t\right\} \le e^{-t}.$$

Taking $t = \log(d^s/\delta)$, we see that

$$\max_{|A|=s} \|U^{A\top} z\|_2^2 \le 2s + 2\sqrt{2s\log(d^s/\delta)} + 2\log(d^s/\delta) \le 7s\log(d) + 3\log(1/\delta)$$

with probability at least $1-\delta$, proving the lemma. □