# Fast and Reliable Parameter Estimation from Nonlinear Observations

Samet Oymak[*] and Mahdi Soltanolkotabi[†]

October 2016

**Abstract**

In this paper we study the problem of recovering a structured but unknown parameter $\boldsymbol{\theta}^*$ from $n$ nonlinear observations of the form $y_i = f(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle)$ for $i = 1, 2, \ldots, n$. We develop a framework for characterizing time-data tradeoffs for a variety of parameter estimation algorithms when the nonlinear function $f$ is unknown. This framework includes many popular heuristics such as projected/proximal gradient descent and stochastic schemes. For example, we show that a projected gradient descent scheme converges at a linear rate to a reliable solution with a near minimal number of samples. We provide a sharp characterization of the convergence rate of such algorithms as a function of sample size, amount of a-prior knowledge available about the parameter and a measure of the nonlinearity of the function $f$. These results provide a precise understanding of the various tradeoffs involved between statistical and computational resources as well as a-prior side information available for such nonlinear parameter estimation problems.

## 1 Introduction

Parameter estimation is fundamental to many supervised learning tasks in signal processing and machine learning. Given training data consisting of $n$ pairs of input features (a.k.a. measurements) $\boldsymbol{x}_i \in \mathbb{R}^p$ and desired outputs $\boldsymbol{y}_i \in \mathbb{R}$ we wish to infer a function that best explains the training data. The simplest functions are linear ones where the outputs are linear functions of the features $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle$ with $\boldsymbol{\theta}^* \in \mathbb{R}^p$ an unknown parameter to be learned from data. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be a feature matrix with rows containing the $n$ features $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ and the vector $\boldsymbol{y} \in \mathbb{R}^n$ containing $n$ output values $y_1, y_2, \ldots, y_n$. The parameter $\boldsymbol{\theta}^*$ is typically estimated by solving an optimization problem of the form

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \quad \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_{\ell_2}^2 \quad \text{subject to} \quad \mathcal{R}(\boldsymbol{\theta}) \le R. \tag{1.1}$$

Here, $\mathcal{R}(\boldsymbol{\theta})$ is a regularization function used to avoid overfitting specially when the number of samples $n$ is significantly smaller than the number of parameters $p$. For example when the parameter $\boldsymbol{\theta}^*$ is believed to be sparse a typical regularization function is $\mathcal{R}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_{\ell_1}$. Over the last few years there has been significant progress in understanding the properties of the optimization problem

[*] Google Inc. 1600 Amphitheatre Parkway, Mountain View, CA

[†] Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA

(1.1) and when it is successful in recovering the unknown parameter $\boldsymbol{\theta}^*$ and in turn predicting future outcomes given a new feature vector. We shall review some of this literature in Section 5.

Even though linear regression models are widely used they rarely capture the feature-output relation in the data precisely. For example in signal processing, such linear models are often first order approximation of typically unknown nonlinear mappings. In this paper we study the sensitivity of various iterative shrinkage schemes used for solving (1.1) to such modeling mismatch. More specifically, we assume that the output is related to the features via the nonlinear equations

$$y_i = f(\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle) \quad \text{for } i = 1, 2, \ldots, n. \tag{1.2}$$

Here, $f : \mathbb{R} \to \mathbb{R}$ is an unknown function and $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is the unknown parameter we wish to estimate. This model is also known as the *single index model* in the statistics literature. Throughout this paper we shall use $\boldsymbol{y} = f(\boldsymbol{X}\boldsymbol{\theta}^*)$ as a shorthand for (1.2). With $f$ unknown, it is natural to try to find the unknown parameter $\boldsymbol{\theta}^*$ via the optimization problem (1.1). In this paper we study the effectiveness of various optimization heuristics used for solving (1.1) under this nonlinear modeling assumption. Our results are very general and hold even when the regularization function $\mathcal{R}$ is nonconvex. This is perhaps surprising as it is not clear that when $\mathcal{R}$ is nonconvex the global optimum to (1.1) can be found. We precisely characterize the run time of projected/proximal gradient and stochastic gradient methods for solving such problems as a function of the number of outputs, the ability of the function $\mathcal{R}$ to enforce prior information and a measure of the nonlinearity of the function. Our results provide an accurate understanding of the various computational and statistical tradeoffs involved when solving such nonlinear parameter estimation problems.

## 2 Precise measures for statistical resources

To arrive at precise tradeoffs between computational and statistical resources we need to quantify the various resources. Computation is easily measured in terms of time/iterations. We measure data size or *sample complexity* in terms of the number of samples $n$. Naturally the required number of samples for reliable parameter estimation depends on how well the regularization function $\mathcal{R}$ can capture the properties of the underlying parameter $\boldsymbol{\theta}^*$. For example if we know our unknown parameter is approximately sparse naturally using an $\ell_1$ norm for the regularizer is superior to using an $\ell_2$ regularizer. To quantify this capability we first need a couple of standard definitions which we adapt from [20].

**Definition 2.1 (Descent set and cone)** *The* set of descent *of a function $t$ at a point $\boldsymbol{\theta}^*$ is defined as*

$$\mathcal{D}_{\mathcal{R}}(\boldsymbol{\theta}^*) = \Big\{\boldsymbol{h} : \ \mathcal{R}(\boldsymbol{\theta}^* + \boldsymbol{h}) \le \mathcal{R}(\boldsymbol{\theta}^*)\Big\}.$$

*The* cone of descent *is defined as a closed cone $\mathcal{C}_{\mathcal{R}}(\boldsymbol{\theta}^*)$ that contains the descent set, i.e. $\mathcal{D}_{\mathcal{R}}(\boldsymbol{\theta}^*) \subset \mathcal{C}_{\mathcal{R}}(\boldsymbol{\theta}^*)$. The* tangent cone *is the conic hull of the descent set. That is, the smallest closed cone $\mathcal{C}_{\mathcal{R}}(\boldsymbol{\theta})$ obeying $\mathcal{D}_{\mathcal{R}}(\boldsymbol{\theta}^*) \subset \mathcal{C}_{\mathcal{R}}(\boldsymbol{\theta}^*)$.*

We note that the capability of the regularizer $\mathcal{R}$ in capturing the properties of the unknown parameter $\boldsymbol{\theta}^*$ depends on the size of the descent cone $\mathcal{C}_{\mathcal{R}}(\boldsymbol{\theta}^*)$. The smaller this cone is the more suited the function $\mathcal{R}$ is at capturing the properties of $\boldsymbol{\theta}^*$. To quantify the size of this set we shall use the notion of mean-width.

**Definition 2.2 (Gaussian width)** *The Gaussian width of a set $\mathcal{C} \in \mathbb{R}^p$ is defined as:*

$$\omega(\mathcal{C}) := \mathbb{E}_{\boldsymbol{g}}[\sup_{\boldsymbol{z} \in \mathcal{C}} \langle \boldsymbol{g}, \boldsymbol{z} \rangle],$$

*where the expectation is taken over $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)$.*

We now have all the definitions in place to quantify the capability of the function $\mathcal{R}$ in capturing the properties of the unknown parameter $\boldsymbol{\theta}^*$. This naturally leads us to the definition of the minimum required number of samples.

**Definition 2.3 (minimal number of samples)** *Let $\mathcal{C}_{\mathcal{R}}(\boldsymbol{\theta}^*)$ be a cone of descent of $\mathcal{R}$ at $\boldsymbol{\theta}^*$ and set $\omega = \omega(\mathcal{C}_{\mathcal{R}}(\boldsymbol{\theta}^*) \cap \mathcal{B}^n)$. Also let $\phi(t) = \sqrt{2}\frac{\Gamma(\frac{t+1}{2})}{\Gamma(\frac{t}{2})} \approx \sqrt{t}$. We define the minimal sample function as*

$$\mathcal{M}(\mathcal{R}, \boldsymbol{\theta}^*, t) = \phi^{-1}(\omega + t) \approx (\omega + t)^2.$$

*We shall often use the short hand $n_0 = \mathcal{M}(\mathcal{R}, \boldsymbol{\theta}^*, t)$ with the dependence on $\mathcal{R}, \boldsymbol{\theta}^*, t$ implied. We note that for convex functions $\mathcal{R}$ based on [2, 10] $n_0$ is exactly the minimum number of samples required for the estimator (1.1) to succeed in recovering an unknown parameter $\boldsymbol{\theta}^*$ with high probability from linear measurements $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^*$. With some overloading, even for non-convex functions $\mathcal{R}$, we shall refer to $n_0$ as the "minimum number of samples".*

We pause to note that prior literature [2, 29] indeed shows that $n_0$ is a good notion of complexity demonstrating that when the feature-response relationship is linear (i.e. $f$ is a linear map) and the regularizer $\mathcal{R}$ is convex the properties of the estimator (1.1) can be precisely characterized in terms of $n_0$. Please also see [6, 20, 30] for some extensions to the non-convex case as well as the role this quantity plays in the computational complexity of projected gradient schemes.

Finally, to analyze algorithm performance as a function of the nonlinearity of the map $f$, we shall use three parameters. Two of these parameters are essentially the intrinsic mean and variance associated with the nonlinear map $f(\cdot)$ and the final term captures a non-asymptotic deviation from linearity.

**Definition 2.4 (nonlinearity parameters)** *Let $g \in \mathbb{R}$ be a standard normal random variable. Define,*

- **Mean term:** $\mu = \mathbb{E}[f(g)g]$,
- **Variance term:** $\sigma^2 = \mathbb{E}[(f(g) - \mu g)^2]$,
- **Deviation term:** $\gamma^2 = \mathbb{E}[g^2(f(g) - \mu g)^2]$.

To see that these nonlinearity measures conform with our intuition, note that in the linear case $f(\boldsymbol{X}\boldsymbol{\theta}) = \boldsymbol{X}\boldsymbol{\theta}$ so that one can take $f(g) = g$ and so the mean term is equal to one and the variance and deviation terms are equal to zero. When $f$ is a nonlinear function like $f(g) = \text{sgn}(g)$ we have $\mu = \sqrt{2/\pi} < 1$ and $\sigma^2 = \gamma = 1 - 2/\pi > 0$. While these examples are beneficial to give intuition, we shall see that for fast and reliable estimation of the vector $\boldsymbol{\theta}^*$ we do not require explicit knowledge of these parameter values.

We can view nonlinear measurements of the form $\boldsymbol{y} = f(\boldsymbol{X}\boldsymbol{\theta}^*)$ as linear measure from the vector $\mu\boldsymbol{\theta}^*$ with an effective noise term $\boldsymbol{w} = f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*$. Intuitively, when the features $\boldsymbol{X}$

are sufficiently randomized, we expect this effective noise to behave similar to $f(\boldsymbol{g}) - \mu \boldsymbol{g}$ where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Consequently, we expect $\mathbb{E}[\|\boldsymbol{w}\|_{\ell_2}^2] \propto n\sigma^2$, which further justifies the definition of $\sigma^2$ as a sort of "variance" from nonlinearity. However, control of this variance term is not enough to show reliable estimation of the parameter with high probability. To be able to make probabilistic statements, we need the Euclidean norm of the effective noise ($\|\boldsymbol{w}\|_{\ell_2}^2$) to concentrate around its expected value. In particular when $\boldsymbol{w} = f(\boldsymbol{g}) - \mu \boldsymbol{g}$, the quantity $\|\boldsymbol{w}\|_{\ell_2}^2$ is the sum of $n$ i.i.d. random variables and exponentially concentrates under mild conditions. For simplicity we will state our results in terms of the concentration probability function defined below.

**Definition 2.5 (Concentration probability function)** *Assume $\boldsymbol{g} \in \mathbb{R}^n$ is distributed as a standard normal random vector $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$ and set $b_n = \mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}] \approx \sqrt{n}$. We define the concentration probability function as*

$$p(\eta) = \mathbb{P}(\|f(\boldsymbol{g}) - \mu \boldsymbol{g}\|_{\ell_2} > \eta b_n \sigma) + \mathbb{P}(|\boldsymbol{g}^T (f(\boldsymbol{g}) - \mu \boldsymbol{g})| > \eta \frac{b_n^2}{\sqrt{n}} \gamma),$$

*with $\mu$, $\sigma$, and $\gamma$ as defined in Definition 2.4.*

We note that Markov inequality implies $p(\eta) \le \frac{2}{\eta^2} \frac{n^2}{b_n^4} \approx \frac{2}{\eta^2}$. However, for many nonlinear functions one can obtain much sharper concentration bounds. For example, when the function $f$ is bounded or Lipschitz, standard concentration of sub-exponential random variables show that $p(\eta) \le e^{-c \min(\eta, \eta^2) n}$ with $c$ a constant depending only on the upper bound/Lipshitz constant of the function $f$. Now that we have precise measures for the various statistical resources we are ready to state our results.

# 3 Precise convergence rates for iterative shrinkage schemes

As mentioned earlier we wish to understand the convergence rates of different iterative shrinkage schemes used for solving nonlinear parameter estimation problems. Throughout this paper we assume that the features $\boldsymbol{x}_i$ are i.i.d. random Gaussian vectors with distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Furthermore, without loss of generality we assume the unknown parameter $\boldsymbol{\theta}^*$ has unit Euclidean norm.

## 3.1 Projected Gradient Descent

Perhaps the simplest algorithm for solving the nonlinear equations (1.2) is *Project Gradient Descent (PGD)* where we use gradient descent on the least squares cost

$$\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|_{\ell_2}^2,$$

followed by projection on the constraint set $\mathcal{K} = \{\boldsymbol{\theta} \in \mathbb{R}^p : \mathcal{R}(\boldsymbol{z}) \le R\}$. More specifically, starting from an initial vector $\boldsymbol{\theta}_0$, PGD iteratively applies the update

$$\boldsymbol{\theta}_{\tau+1} = \mathcal{P}_{\mathcal{K}} \left( \boldsymbol{\theta}_\tau + \alpha_\tau \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_\tau) \right). \tag{3.1}$$

Here, $\mathcal{P}_{\mathcal{K}}(\boldsymbol{z})$ denotes the Euclidean projection of the vector $\boldsymbol{z}$ onto the set $\mathcal{K}$ and $\alpha_\tau$ is the step size. Throughout this paper we will assume that the tuning parameter is perfectly tuned so that

$R = \mathcal{R}(\mu\boldsymbol{\theta}^*)$ where $\mu$ is the mean term per Definition 2.4. However, we can extend our arguments to the case where $R \neq \mathcal{R}(\mu\boldsymbol{\theta}^*)$ by utilizing the sensitivity analysis developed in [20, Theorem 2.6].

Our first result shows that projected gradient descent allows for fast and reliable parameter estimation from nonlinear observations.

**Theorem 3.1** *Let $f : \mathbb{R} \to \mathbb{R}$ be an unknown nonlinear function. Also, let $\boldsymbol{y} = f(\boldsymbol{X}\boldsymbol{\theta}^*) \in \mathbb{R}^n$ be $n$ nonlinear observations from $\boldsymbol{\theta}^*$ with the feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ consisting of i.i.d. $\mathcal{N}(0,1)$ entries. Furthermore, let $\kappa_{\mathcal{R}}$ be a constant that is equal to one for convex regularizers $\mathcal{R}$ and equal to two for nonconvex ones. Furthermore, let $\boldsymbol{n}_0 = \mathcal{M}(\mathcal{R}, \mu\boldsymbol{\theta}^*, t)$ be the minimal number of data samples as per Definition 2.3 and assume*

$$n \geq 8\kappa_{\mathcal{R}}^2 n_0. \tag{3.2}$$

*Also let $\mu, \sigma$, and $\eta$ be the nonlinearity parameters per Definition 2.4. Then, starting from any initial estimate $\boldsymbol{\theta}_0$ the PGD iterates (3.1) with step size $\alpha_\tau = 1/b_n^2 \approx 1/n$ and tuning parameter $R = \mathcal{R}(\mu\boldsymbol{\theta}^*)$ obeys*

$$\|\boldsymbol{\theta}_\tau - \mu\boldsymbol{\theta}_0\|_{\ell_2} \leq \left(\sqrt{8\kappa_{\mathcal{R}}^2\frac{n_0}{n}}\right)^\tau \|\boldsymbol{\theta}_0 - \mu\boldsymbol{\theta}^*\|_{\ell_2} + \frac{\kappa_{\mathcal{R}}}{1-\left(\sqrt{8\kappa_{\mathcal{R}}^2\frac{n_0}{n}}\right)}\frac{\eta\left(\sigma\sqrt{n_0}+\gamma\right)}{\sqrt{n}}, \tag{3.3}$$

*for all $\tau$ with probability at least $1-p(\eta)-10e^{-\frac{t^2}{8}}$. Here, $p(\eta)$ is the concentration probability function as per Definition 2.5.*

Note that $\eta\left(\sigma\sqrt{n}\right)$ is roughy the Euclidean norm of the effective noise $\boldsymbol{w} = f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*$ induced by replacing the nonlinear equation $\boldsymbol{y} = f(\boldsymbol{X}\boldsymbol{\theta}^*)$ with the linear equation $\boldsymbol{y} = \mu\boldsymbol{X}\boldsymbol{\theta}^*$. Thus, the theorem above shows that the projected gradient updates converge at a linear rate to a small neighborhood around the "true" solution $\mu\boldsymbol{\theta}^*$. The radius of this neighborhood decreases with an increase in the number of samples $n$. The size of this radius is near-optimal and comparable to recent results [19,22,28] where the estimate is obtained by solving the convex optimization problem in (1.1) which applies only when the regularization function $\mathcal{R}$ is convex. Indeed, the size of this radius scales like $\sqrt{n_0/n}\,\|\boldsymbol{w}\|_{\ell_2}$ which up to a small constant is exactly the same as the result one would get when the model is linear of the form $\boldsymbol{y} = \mu\boldsymbol{X}\theta^* + \boldsymbol{w}$. This is perhaps unexpected as it demonstrates that our nonlinear model exactly behaves like a fictitious linear model with the same effective noise!

The convergence rate in (3.3) is linear and proportional to $1/\sqrt{n}$. This shows that the convergence rate improves with an increase in the sample size which in turn leads to faster convergence with more data samples. This implies that the more samples we have not only does the quality of the solution improve but also that we arrive at this solution with less computational effort. Thus, more samples is beneficial both in terms of statistical reliability and computational efficiency.[1]

Finally, another interesting aspect of Theorem 3.1 is that it applies to both convex and nonconvex regularizers. Showing that one can obtain statistically reliable solutions with nonconvex

[1] We note that the latter statement about computational efficiency is true only to a certain extent. In particular when the feature matrix $\boldsymbol{X}$ does not have a fast vector-matrix multiply the improvemed convergence rate is soon dominated by the increased cost of the matrix-vector multiplies in each iteration due to the increase in the number of samples and hence the size of the feature matrix. However, we expect (3.3) to be correct for many random feature models that do have fast vector-matrix multiply e.g. Fourier type matrices in signal processing or sparse features in machine learning.

regularizers is perhaps unexpected as it is not a-priori clear that the global solution to (1.1) can be found using a computationally tractable algorithm.[2]

### 3.2 Stochastic gradient schemes

In this section we provide convergence guarantees for stochastic gradient schemes. Our result concerns a Projected variation of the Stochastic Gradient Descent algorithm [5] which we refer to as PSGD. Let $\{\psi_\tau\}_{\tau=1}^{\infty}$ denote random integer taking the value $i \in \{1, 2, \dots, n\}$ with probability $\frac{\|\boldsymbol{x}_i\|_{\ell_2}^2}{\|\boldsymbol{X}\|_F^2}$. Starting from an initial estimate $\boldsymbol{\theta}_0$ the PSGD algorithm iteratively applies the updates

$$\boldsymbol{\theta}_{\tau+1} = \mathcal{P}_{\mathcal{K}}\left(\boldsymbol{\theta}_\tau + \frac{(\boldsymbol{y}_{\psi_\tau} - \langle \boldsymbol{x}_{\psi_\tau}, \boldsymbol{\theta}_\tau \rangle)}{\left\|\boldsymbol{x}_{\psi_\tau}\right\|_{\ell_2}^2} \boldsymbol{x}_{\psi_\tau}\right). \tag{3.4}$$

We will show that it is possible to obtain guarantees that are on par with the PGD guarantees derived in the previous section. This suggests that nonlinear parameter estimation problems can be solved by highly parallel and asynchronous algorithms. As in the previous case our error bounds will again only depend on the effective noise terms $\sigma$ and $\gamma$.

**Theorem 3.2** *Let $f : \mathbb{R} \to \mathbb{R}$ be an unknown nonlinear function. Also, let $\boldsymbol{y} = f(\boldsymbol{X}\boldsymbol{\theta}^*) \in \mathbb{R}^n$ be $n$ nonlinear observations from $\boldsymbol{\theta}^*$ with the feature matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ consisting of i.i.d. $\mathcal{N}(0,1)$ entries. Also, let $\mathcal{R}$ be a convex regularizer. Furthermore, let $\boldsymbol{n}_0 = \mathcal{M}(\mathcal{R}, \mu\boldsymbol{\theta}^*, t)$ be the minimal number of data samples as per Definition 2.3 and assume*

$$n > n_0. \tag{3.5}$$

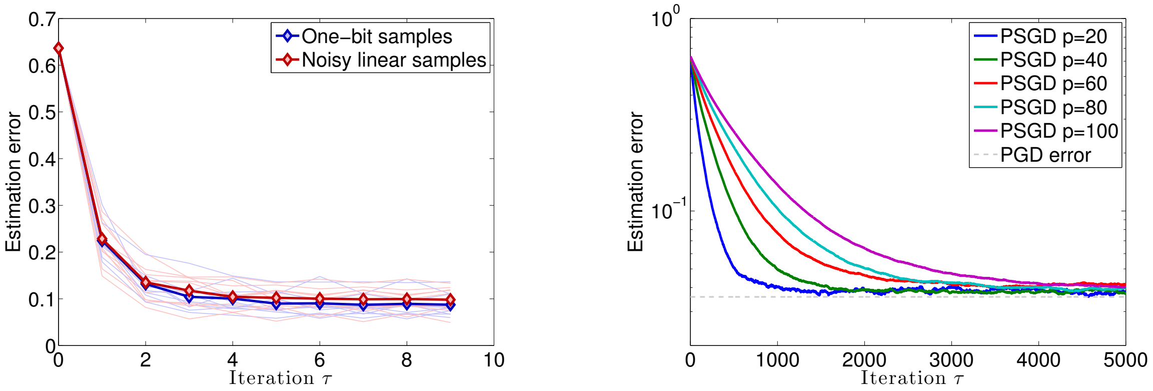
*Also let $\mu, \sigma$, and $\eta$ be the nonlinearity parameters per Definition 2.4. Then, starting from any initial estimate $\boldsymbol{\theta}_0$ the PSGD iterates (3.4) with tuning parameter $R = \mathcal{R}(\mu\boldsymbol{\theta}^*)$ obey*

$$\mathbb{E}[\|\boldsymbol{\theta}_\tau - \mu\boldsymbol{\theta}^*\|_{\ell_2}^2] \le \left(1 - \frac{\left(1 - \sqrt{\frac{n_0}{n}}\right)^2}{2p}\right)^\tau \|\boldsymbol{\theta}_0 - \mu\boldsymbol{\theta}^*\|_{\ell_2}^2 + \frac{1.01}{\left(1 - \sqrt{\frac{n_0}{n}}\right)^2}\eta^2\sigma^2, \tag{3.6}$$

*for all $\tau$ with probability at least $1 - (n+1)e^{-cp} - p(\eta)$. Here, the expectation is over the random variables $\{\psi_s\}_{s=1}^{\tau}$.*

Similar to our results for the PGD iterations the results of Theorem 3.2 above demonstrates that PSGD also converges at a geometric rate to a neighborhood of the true solution $\mu\boldsymbol{\theta}^*$. However, comparing (3.6) with its counterpart in (3.3) the PSGD results are weaker in two ways. First, while the convergence is geometric it is no longer linear and slower than its PGD counter part. However, we should point out that the cost of each iteration of PSGD is roughly $1/n$ times the cost of a PGD update. Second, the radius of the neighborhood of the true solution is larger in the PSGD case compared with the PGD case by a factor of roughly size $\sqrt{n/n_0}$. We believe this to be an artifact of our proof technique and we expect that for $\tau$ sufficiently large the the second term should be of the form $\frac{n_0}{n}\eta^2\sigma^2$ in lieu of $\eta^2\sigma^2$.

[2] This statement is of course only true if the projection onto the sub-level sets of the regularization function is computationally efficient. This is true for many non-convex regularizers including some $\ell_p$ norms with $p < 1$. Furthermore, we note that projection onto convex sets may also not be in general tractable a good example of this is projection onto the set of completely positive matrices which is known to be NP-hard.



**(a)** **(b)**

**Figure 1:** (a) Estimation errors ($\|\boldsymbol{\theta}_\tau - \sqrt{2/\pi}\boldsymbol{\theta}^*\|_{\ell_2}$) obtained via running PGD iterates as a function of the number of iterations $\tau$. The plots are for two different observations models: 1) nonlinear one bit measurements $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^*$ and 2) noisy linear observations $\boldsymbol{y} = \sqrt{2/\pi}\boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{w}$ with $\boldsymbol{w} \sim \mathcal{N}(0, (1-2/\pi)\boldsymbol{I}_n)$. The bold colors depict average behavior over 100 trials. None bold color depict the estimation error of some sample trials. (b) Convergence behavior of PSGD for $\frac{n}{p} = 4$ and $\frac{s}{p} = 0.1$ as $p$ varies.

### 3.3 Proximal gradient methods

In Section 3.1 we discussed our results for nonlinear estimation problem by enforcing the constraint $\mathcal{R}(\boldsymbol{\theta}) \le \mathcal{R}(\mu\boldsymbol{\theta}^*)$. Another popular approach for finding structured solutions to linear inverse problems is solving a penalized variant of (5.1). Our framework also provides some insights for convergence of such proximal gradient methods. However, our results for proximal methods require a few additional definitions and modeling assumptions. We therefore defer these results to Appendix A.

## 4 Numerical Experiments

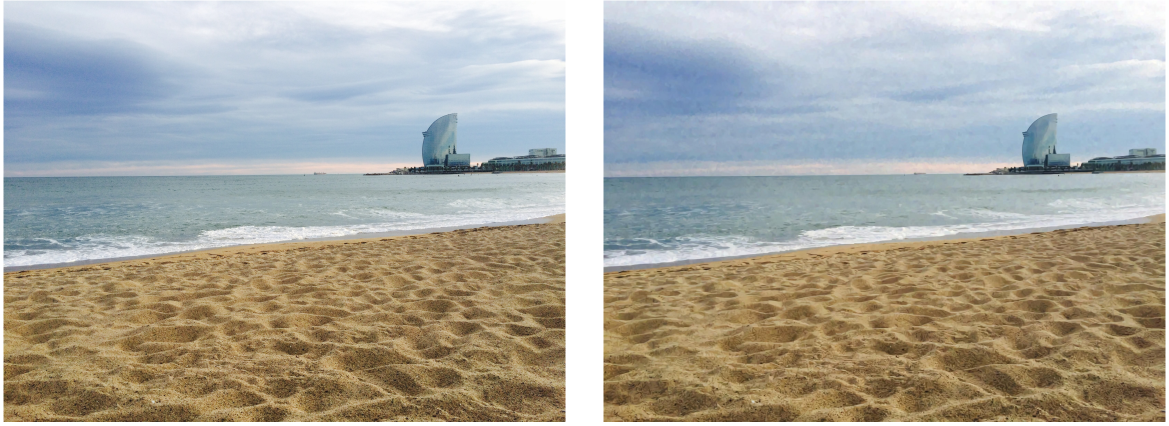
In this section we will discuss a few synthetic numerical experiments to corroborate our theoretical results in the previous sections as well as understand some of their limitations. First we will start with some synthetic experiments followed by some experiments on natural images.

### 4.1 Synthetic experiments

In our first experiment we generate a unit norm sparse vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$ of dimension $p = 500$ containing $s = p/50$ non-zero entries. We also generate a random feature matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with $n = p/2$ and containing i.i.d. $\mathcal{N}(0,1)$ entries. We now take two sets of observations of size $n$ from $\boldsymbol{\theta}^*$:

- One-bit nonlinear observations: the response vector is equal to $\boldsymbol{y} = \text{sgn}(\boldsymbol{X}\boldsymbol{\beta})$.
- Linear observations: the response is $\boldsymbol{y} = \sqrt{\frac{2}{\pi}}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{w}$ with $\boldsymbol{w} \sim \mathcal{N}(0, (1-2/\pi)\boldsymbol{I}_n)$.

We wish to infer the vector $\sqrt{2/\pi}\boldsymbol{\theta}^*$ from these set of observations. Note that while the observation models are different, the effective noise levels in both problems are roughly of the same size i.e. the Euclidean norm of $\boldsymbol{w} = \boldsymbol{y} - \sqrt{\frac{2}{\pi}}\boldsymbol{X}\boldsymbol{\theta}^*$ in both cases is roughly of size $\sqrt{n}\sigma$ with $\sigma = \sqrt{1-\frac{2}{\pi}}$. We



**(a)** Original image

**(b)** Reconstructed image

**Figure 2:** (a) Image of a beach in Barcelona, Spain. (b) Reconstructed image obtained from nonlinear observations (quantized (DCT) coefficients to 16 levels and then subsampled the observations by a factor of two). Reconstructed PSNR=27.5944.

apply the PGD iterations (3.1) to both observations models starting from $\boldsymbol{\theta}_0 = \mathbf{0}$. In Figure 1a the resulting estimation errors ($\|\boldsymbol{\theta}_\tau - \sqrt{2/\pi}\boldsymbol{\theta}^*\|_{\ell_2}$) are depicted as a function of the number of iterations $\tau$. These bold colors depicts average behavior over 100 trials. The estimation error of some sample trials are also depicted in none bold colors. The plots in Figure 1a clearly show that PGD iterates applied to nonlinear observations converge quickly to an estimate which is of the same size as the effective noise induced by the nonlinearity. In this sense, PGD iterates converge quickly to a reliable solution which has exactly the same quality as the optimal results obtained in [22, 28].[3] Figure 1a also clearly demonstrates that the behavior of the PGD iterates applied to both models are essentially the same further corroborating the results of Theorem 3.1. This leads to the striking conclusion that there is essentially no difference between the convergence of linear and nonlinear samples when the effective noise is of the same size! In addition Figure 1a shows that one iteration of PGD updates applied to nonlinear observations is not sufficient for reaching a statistically reliable solution and further iterations lead to further improvements. This demonstrates the advantage of our framework over other computational methods [23, 33] as the error bounds obtained in these papers are on par with the error bounds obtained by applying the first iteration of PGD.

In the next experiment we again consider the nonlinear one-bit observation model $\boldsymbol{y} = \text{sgn}(\boldsymbol{X}\boldsymbol{\theta}^*)$ with $\boldsymbol{\theta}^*$ a sparse vector with $s$ nonzero entries. In this experiment we fix the quantities $\frac{n}{p} = 4$, $\frac{s}{p} = 0.1$ (which in turn fixes $\frac{n_0}{n}$) and vary $p$. We apply the PSGD iteration of (3.4) and depict the estimation error as a function of the number of iterations $\tau$ in Figure 1b for different values of $p$. These plots show that after PSGD converges, the estimation error is the same as that of PGD. This is not consistent with the second term in (3.6) of Theorem 3.2. As mentioned earlier we conjecture that the result of (3.6) holds with the second term divided by $\sqrt{n/n_0}$. Such a result would be consistent with the numerical simulation of Figure 1b.

### 4.2 Experiments on images

In this section we demonstrate the utility of an image denoiser for image recovery from quantized nonlinear observations. Our nonlinear observations consists of modulating each pixel of the image by a random i.i.d. ±1 mask, applying a two dimensional Discrete Fourier Transform (DCT), and then picking a random subset of size $m$ of these DCT coefficients and then quantizing the results to 16 different values (4 bits). Since we have a color photograph we apply such nonlinear observation to each color band for a total of $3m$ nonlinear observations.

To recover the original image from such under-sampled nonlinear observations we start from $\boldsymbol{\theta}_0 = \mathbf{0}$ and iteratively apply the updates in (A.2). However, in liu of the prox function we use a nonlinear mapping $\mathcal{S}$ with a tunning parameter $\lambda_\tau$. We note that many nonlinear mapping can also be thought of as the prox of another function. We shall use the CBM3D denoiser [11] as the nonlinear mapping $\mathcal{S}$ so that $\mathcal{S}$ is the denoising procedure with a tunning parameter $\lambda_\tau$ which is tuned based on the assumed variance of the noise. We shall use $\lambda_\tau = \max(\lambda_0\rho^\tau, \lambda_{\min})$ in our experiments. This choice is based on our theoretical results stated in Theorem A.3 of Appendix A which suggest that this is a good tuning strategy. We now apply this proximal update with the CBM3D denoiser with $\lambda_0 = 14.43375 \left\|\boldsymbol{\theta}^*\right\|_{\ell_2} \frac{\sqrt{p}}{n}$ and $\gamma = 0.95$. We remind the reader that the image has $3n$ total pixels ($n$ pixels in each color band). For $m = 0.5n$ we run 10 iterations of the proximal update (A.2) and record the relative error $\left\|\hat{\boldsymbol{\theta}} - \mu\boldsymbol{\theta}^*\right\|_{\ell_2} / \left\|\mu\boldsymbol{\theta}^*\right\|_{\ell_2}$ (color images are viewed as a large vector). We depict the original image together with the reconstruction in Figure 2. The relative noise induced by the nonlinearity in this case was 0.1977 ($\left\|\mu\boldsymbol{X}\boldsymbol{\theta}^* - f(\boldsymbol{X}\boldsymbol{\theta}^*)\right\|_{\ell_2} / \left\|\mu\boldsymbol{X}\boldsymbol{\theta}^*\right\|_{\ell_2} = 0.1977$). The relative error we obtained by our reconstruction was 0.0757 which is equivalent to a Peak Signal to Noise Ratio (PSNR) of 27.5944. This figure indicates that even though the image is under-sampled by a factor of two and we quantize the image into four bits we get a rather good reconstruction.

## 5 Prior Art

During the last decade, the problem of sparse estimation has been the focus of significant interest. This central problem in high-dimensional statistics is about estimating an unknown sparse parameter from possibly underdetermined observations. This task is often accomplished by the lasso estimator

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\right\|_{\ell_2}^2 + \lambda \left\|\boldsymbol{\theta}\right\|_{\ell_1}, \tag{5.1}$$

where the response vector $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^* + \boldsymbol{w}$ consists of noisy linear observations. In recent years major developments in the theory of sparse estimation [8,9,12,25] have emerged. In particular, the main optimization (5.1) has been generalized in multiple ways.

- **Design matrix:** There is now a good understanding of the design matrices $\boldsymbol{X}$ that allow near-optimal estimation of $\boldsymbol{\theta}^*$. Researchers have characterized useful conditions such as restricted strong convexity, null space property and incoherence [7,19].

- **Parameter structure:** $\boldsymbol{\theta}^*$ does not have to be sparse and with carefully choice of regularizers many other structures can be recovered [2,10] with a minimal number of samples.

[3] We note that these papers do not provide any computational convergence guarantees.

- **Response variable:** The response vector $\boldsymbol{y}$ does not have to be a linear function of $\boldsymbol{X}\boldsymbol{\theta}^*$. Recent papers [17, 24, 28, 32, 33] allow for a much more general model $\boldsymbol{y} = f(\boldsymbol{X}\boldsymbol{\beta})$ where $f$ applies entrywise on the elements of $\boldsymbol{X}\boldsymbol{\beta}$. Despite these interesting results most of the literature in this direction such as [24,28] do not address computational issues or only address them for particular structures (e.g. sparsity) and algorithms such as [33].

In this paper, we propose and analyze iterative algorithms that solve nonlinear estimation problems of the form (1.2). Our contributions can be summarized as follows.

• **Nonlinear Observations:** We allow for an unknown nonlinear relationship between the unknown parameter $\boldsymbol{\theta}^*$ and the response variable $\boldsymbol{y}$ of the form $\boldsymbol{y} = f(\boldsymbol{X}\boldsymbol{\theta}^*)$. By viewing nonlinearity as noise, we derive sharp performance guarantees for a variety of fast algorithms in terms of basic statistics of the nonlinearity of the unknown function $f$.

• **Unified analysis:** The same idea applies to several different algorithms with minor modification. As a result, we obtain convergence rates and estimation error bounds for interesting algorithms such as proximal gradient and projected stochastic methods.

• **Optimal error rates:** Our analysis of the proposed algorithms is sharp and the resulting bounds are optimal up to small constants. Our bounds yield error rates comparable to what one would get from studying the properties of the optimum solution to the problem (1.1) in the special case where the regularizer $\mathcal{R}$ is a convex function [28].[4]

•**Nonconvex constraints:** An interesting aspect of our framework is that it applies even when the constraints are nonconvex. As we mentioned our convergence results are for specific algorithms and we do not just study the properties of the optimum solution to (1.1) as in previous publications [28]. Furthermore, in contrast to [28] some of our theorems hold without requiring the regularizer to be convex. This is particularly important when the regularizer is nonconvex as it is not clear that the global optimum to (1.1) can be found via a tractable algorithm.

•**Precise tradeoffs between computational and statistical resources:** In this paper we have provided precise convergence guarantees for a variety of nonlinear parameter estimation algorithms. Thus, our results provides precise tradeoffs between computational resources such as time/iterations and statistical resources such as data size, amount of nonlinearity, amount of available prior knowledge (through the choice of a regularizer) etc. In this vein, the result of this paper can also be seen as a generalization of [1,20] to the nonlinear estimation setup. Indeed, in the absence of nonlinearity we can recover many of the results stated in [1,20]. In comparison to [1,20] our results also apply to a variety of different algorithms: proximal methods, stochastic methods, etc. We note however, that while our results can be applied to a variety of feature distributions such as those studied in [1, 20], in this paper we have focused on Gaussian feature matrices $\boldsymbol{X}$.

## 6 Conclusion and future directions

In this paper we have presented a framework for characterizing time-data tradeoffs for a variety of nonlinear parameter estimation algorithms. Our results provide a precise understanding of the various tradeoffs involved between statistical and computational resources demonstrating that

[4] We note that compared to [28] several works on nonlinear estimation such as [23, 33] suffer from the fact that the resulting estimation error is nonzero even if there is no nonlinearity in the model (i.e. $f(\boldsymbol{X}\boldsymbol{\theta}^*) = \boldsymbol{X}\boldsymbol{\theta}^*$) even if $n$ is sufficiently large. However, our analysis as well as that of [28] precisely recovers the unknown parameter in the linear setting with a minimal number of observations while yielding better or equal guarantees for highly nonlinear problems such as 1-bit compressive sensing.

fast and reliable parameter estimation is possible from nonlinear observations. There are many interesting future direction to pursue:

- **Precise statistical constants:** In this paper we have shown that many iterative algorithms for nonlinear parameter estimation e.g. Projected Gradient Descent (PGD) converge to a solution which is a small constant factor away from the "effective noise" induced by the nonlinearity in these problems. However, Figure 1a suggests that this constant should be exactly equal to one. Recently, Thrampoulidis et al. [28] rigorously argued this for the optimal solution to a convex program [28] when the regularization function is convex. However analyzing the properties of iterative algorithms with sharp convergence guarantees as provided in this paper proves to be more challenging. An interesting future direction is to show whether the "precise" error rates of PGD does indeed depend only on the "effective noise" without any additional constants. Furthermore, as we mentioned in the main text our results for PSGD seems to be off by a factor of $\sqrt{n}/\sqrt{n_0}$ closing this gap is a particularly important future direction.

- **Parallel and lock free schemes:** For nonlinear parameter estimation problems parallel algorithms are efficient and more desirable specially when the design matrix is sparse. There are very interesting recent work for providing guarantees for parallel and lock-free implementations of stochastic algorithms [26]. An interesting future direction is to characterize time-data tradeoffs for such algorithms specialized for use in nonlinear parameter estimation problems.

- **Proximal gradient methods:** Our guarantees for proximal gradient schemes discussed in Appendix A require additional modeling assumption and are not as strong as our results for projected gradient methods. Given the wide use of proximal methods in practice providing optimal guarantees for proximal algorithms is an important future direction.

# 7 Proofs

In this section we will prove all of our results. Throughout we use $\mathcal{B} \in \mathbb{R}^n$ to denote the unit ball of $\mathbb{R}^n$. We begin with stating some preliminary lemmas that we will use throughout the proofs.

## 7.1 Preliminaries

In this section we gather some preliminary lemmas about projections onto sets and certain properties of Gaussian random matrices. Most of the results stated in this section are directly adapted from [20] (we only state the results for the convenience of the reader). The first one is a result concerning projection onto cones.

**Lemma 7.1** *Let $\mathcal{C} \subset \mathbb{R}^n$ be a closed cone and $\boldsymbol{v} \in \mathbb{R}^n$. Then*

$$\|\mathcal{P}_{\mathcal{C}}(\boldsymbol{v})\|_{\ell_2} = \sup_{\boldsymbol{u} \in \mathcal{C} \cap \mathcal{B}^n} \boldsymbol{u}^* \boldsymbol{v}. \tag{7.1}$$

The next lemma just states that translation preserves distances.

**Lemma 7.2** *Suppose $\mathcal{K} \subset \mathbb{R}^n$ is a closed set. The projection onto $\mathcal{K}$ obeys*

$$\mathcal{P}_{\mathcal{K}}(\boldsymbol{x} + \boldsymbol{v}) - \boldsymbol{x} = \mathcal{P}_{\mathcal{K} - \{\boldsymbol{x}\}}(\boldsymbol{v}).$$

The next lemma compares the length of a projection onto a set to the length of projection onto the conic approximation of the set.

**Lemma 7.3 (Comparison of projections)** *Let $\mathcal{D}$ be a closed and nonempty set that contains* $\mathbf{0}$. *Let $\mathcal{C}$ be a nonempty and closed cone containing $\mathcal{D}$ ($\mathcal{D} \subset \mathcal{C}$). Then for all $\boldsymbol{v} \in \mathbb{R}^n$,*

$$\|\mathcal{P}_{\mathcal{D}}(\boldsymbol{v})\|_{\ell_2} \le 2 \|\mathcal{P}_{\mathcal{C}}(\boldsymbol{v})\|_{\ell_2} \tag{7.2}$$

*Furthermore, if $\mathcal{D}$ is a convex set. Then for all $\boldsymbol{v} \in \mathbb{R}^n$,*

$$\|\mathcal{P}_{\mathcal{D}}(\boldsymbol{v})\|_{\ell_2} \le \|\mathcal{P}_{\mathcal{C}}(\boldsymbol{v})\|_{\ell_2}. \tag{7.3}$$

We next state a result about control of set restricted eigenvalues of Gaussian random matrices from [20].

**Lemma 7.4** *Let $\mathcal{C} = \mathcal{C}_{\mathcal{R}}(\mu\boldsymbol{\theta}^*)$ be the cone of descent of the regularizer $\mathcal{R}$ at the point $\mu\boldsymbol{\theta}^*$ as per Definition 2.1. Furthermore, let $\boldsymbol{n}_0 = \mathcal{M}(\mathcal{R}, \mu\boldsymbol{\theta}^*, t)$ be the minimal number of data samples as per Definition 2.3. Then for a random matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with i.i.d. $\mathcal{N}(0,1)$ entries*

$$\sup_{\boldsymbol{v},\boldsymbol{u}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{u}^T\left(\boldsymbol{I} - \frac{1}{b_n^2}\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{v} \le \sqrt{8\frac{n_0}{n}},$$

*holds with probability at least $1 - 9e^{-\frac{t^2}{8}}$. Here, $b_n = \sqrt{2}\frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \approx \sqrt{n}$ with $\Gamma$ denoting the Gamma function.*

The next lemma due to Gordon [16] provides a lower bound on the minimum eigenvalue of a random Gaussian matrix restricted to a cone.

**Lemma 7.5 (Gordon's escape through the mesh [16])** *Assume the same setup and definitions as Lemma 7.4 above. Then for a random matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ with i.i.d. $\mathcal{N}(0,1)$ entries,*

$$\inf_{\boldsymbol{u}\in\mathcal{C}\cap\mathcal{B}^p} \|\boldsymbol{X}\boldsymbol{u}\|_{\ell_2}^2 \ge (b_n - b_0)^2,$$

*holds with probability at least $1 - e^{-\frac{t^2}{2}}$.*

### 7.2 Key Lemma for controlling nonlinear terms

In this section we state and prove a key lemma that is crucial in our analysis and allows us to deal with the nonlinearity of our observations.

**Lemma 7.6 (Controlling the effective noise)** *Let $\mathcal{C} = \mathcal{C}_{\mathcal{R}}(\mu\boldsymbol{\theta}^*)$ be the cone of descent of the regularizer $\mathcal{R}$ at the point $\mu\boldsymbol{\theta}^*$ as per Definition 2.1. Also let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be a matrix of i.i.d. $\mathcal{N}(0,1)$ entries independent of $\boldsymbol{\theta}^*$ and $\boldsymbol{w} = f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*$ be the effective noise. Furthermore, let $\boldsymbol{n}_0 = \mathcal{M}(\mathcal{R}, \mu\boldsymbol{\theta}^*, t)$ be the minimal number of data samples as per Definition 2.3. Also let $\mu, \sigma$, and $\eta$ be the nonlinearity parameters for the function $f$ as per definition 2.4. Then*

$$\left\|\mathcal{P}_{\mathcal{C}}\left(\boldsymbol{X}^T\boldsymbol{w}\right)\right\|_{\ell_2} \le \frac{b_n^2}{\sqrt{n}}\eta\left(\sigma\sqrt{n_0} + \gamma\right).$$

*holds with probability at least $1 - p(\eta) - \exp(-t^2/2)$. Here, $p(\eta)$ is the concentration probability function as per Definition 2.5 and $b_n = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \approx \sqrt{n}$.*

**Proof** We begin by defining $\boldsymbol{X}_{||} = \boldsymbol{X}\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}$ and $\boldsymbol{X}_{\perp} = \boldsymbol{X}\left(\boldsymbol{I}-\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\right)$. Based on these definition we can decompose $\boldsymbol{X}$ into the sum of two matrices $\boldsymbol{X} = \boldsymbol{X}_{||} + \boldsymbol{X}_{\perp}$ where the rows of $\boldsymbol{X}_{||}/\boldsymbol{X}_{\perp}$ are parallel/orthogonal to the direction of $\boldsymbol{\theta}^*$. Now note that since $\boldsymbol{X}$ has independent standard normal entries $\boldsymbol{X}_{||}$ and $\boldsymbol{X}_{\perp}$ are independent. Hence using Lemma 7.1

$$\begin{aligned}
\left\|\mathcal{P}_{\mathcal{C}}\left(\boldsymbol{X}^T\boldsymbol{w}\right)\right\|_{\ell_2} &= \sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{w},\\
&= \sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \left(\boldsymbol{v}^T\boldsymbol{X}_{||}^T\boldsymbol{w} + \boldsymbol{v}^T\boldsymbol{X}_{\perp}^T\boldsymbol{w}\right),\\
&\le \sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{v}^T\boldsymbol{X}_{||}^T\boldsymbol{w} + \sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{v}^T\boldsymbol{X}_{\perp}^T\boldsymbol{w},\\
&= \sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{v}^T\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{w} + \sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{v}^T\boldsymbol{X}_{\perp}^T\boldsymbol{w},\\
&\le \left|\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{w}\right|\left(\sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p}\left|\boldsymbol{v}^T\boldsymbol{\theta}^*\right|\right) + \sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{v}^T\boldsymbol{X}_{\perp}^T\boldsymbol{w},\\
&\le \left|\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{w}\right| + \sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{v}^T\boldsymbol{X}_{\perp}^T\boldsymbol{w}. \qquad (7.4)
\end{aligned}$$

In the last inequality we used the fact that $\left|\boldsymbol{v}^T\boldsymbol{\theta}^*\right| \le \|\boldsymbol{v}\|_{\ell_2}$ together with $\|\boldsymbol{\theta}^*\|_{\ell_2} = 1$. We now proceed by bounding each of these terms. To bound the first term note that $\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{w}$ can be rewritten in the form

$$\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{w} = (\boldsymbol{X}\boldsymbol{\theta}^*)^T\left(f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*\right).$$

Since $\boldsymbol{X}$ has i.i.d. $\mathcal{N}(0,1)$ entries and $\|\boldsymbol{\theta}^*\|_{\ell_2} = 1$, thus $\boldsymbol{g} = \boldsymbol{X}\boldsymbol{\theta}^*$ is a random Gaussian vector with i.i.d. $\mathcal{N}(0,1)$ entries. Therefore, utilizing Definitions 2.4 and 2.5

$$\left|\boldsymbol{\theta}^{*T}\boldsymbol{X}^T\boldsymbol{w}\right| = \left|\boldsymbol{g}^T(f(\boldsymbol{g}) - \mu\boldsymbol{g})\right| \le \frac{b_n^2}{\sqrt{n}}\eta\gamma, \qquad (7.5)$$

holds with probability at least $1 - \mathbb{P}\left(\left|\boldsymbol{g}^T(f(\boldsymbol{g}) - \mu\boldsymbol{g})\right| > \eta\frac{b_n^2}{\sqrt{n}}\gamma\right)$.

To bound the second term in (7.4) let $\tilde{\mathcal{C}}$ denote the projection of the set $\mathcal{C}\cap\mathcal{B}^p$ onto the plane orthogonal to the direction of $\boldsymbol{\theta}^*$, i.e. $\tilde{\mathcal{C}} = (\boldsymbol{I} - \boldsymbol{\theta}^*\boldsymbol{\theta}^{*T})(\mathcal{C}\cap\mathcal{B}^p)$. Now note that for a Gaussian random vector $\boldsymbol{z}\in\mathbb{R}^p$ with i.i.d. $\mathcal{N}(0,1)$ entries by standard Gaussian concentration

$$\sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{v}^T(\boldsymbol{I} - \boldsymbol{\theta}^*\boldsymbol{\theta}^{*T})\boldsymbol{z} = \sup_{\boldsymbol{u}\in\tilde{\mathcal{C}}} \boldsymbol{u}^T\boldsymbol{z} \le \omega(\tilde{\mathcal{C}}) + t \le \omega(\mathcal{C}\cap\mathcal{B}^p) + t = \omega + t, \qquad (7.6)$$

holds with probability at least $1 - e^{-\frac{t^2}{2}}$. Now note that $\boldsymbol{w} = f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^* = f(\boldsymbol{X}_{||}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}_{||}\boldsymbol{\theta}^*$ is only a function of $\boldsymbol{X}_{||}$ and is thus independent of $\boldsymbol{X}_{\perp}$. Thus the vector $\boldsymbol{X}_{\perp}^T\boldsymbol{w}$ has the same distribution as a random vector $\|\boldsymbol{w}\|_{\ell_2}(\boldsymbol{I} - \boldsymbol{\theta}^*\boldsymbol{\theta}^{*T})\boldsymbol{z}$ where $\boldsymbol{z}\in\mathbb{R}^p$ is a Gaussian random vector that is independent of $\boldsymbol{w}$ and has i.i.d. $\mathcal{N}(0,1)$ entries. Thus using (7.6) together with Definitions 2.4 and 2.5 we conclude that

$$\sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p} \boldsymbol{v}^T\boldsymbol{X}_{\perp}^T\boldsymbol{w} \le (\omega + t)\|\boldsymbol{w}\|_{\ell_2} = (\omega + t)\|f(\boldsymbol{g}) - \mu\boldsymbol{g}\|_{\ell_2} \le (\omega + t)\eta b_n\sigma, \qquad (7.7)$$

holds with probability at least $1-e^{-\frac{t^2}{2}}-\mathbb{P}(\|f(\boldsymbol{g})-\mu\boldsymbol{g}\|_{\ell_2}>\eta b_n\sigma)$. From Definition 2.3 we know that $n_0$ is defined via $\omega+t=\sqrt{2}\frac{\Gamma\left(\frac{n_0+1}{2}\right)}{\Gamma\left(\frac{n_0}{2}\right)}$. By [20, Lemma 6.9] $\sqrt{2}\frac{\Gamma\left(\frac{n_0+1}{2}\right)}{\Gamma\left(\frac{n_0}{2}\right)}\le b_n\sqrt{\frac{n_0}{n}}$ implying that $(\omega+t)\le b_n\sqrt{\frac{n_0}{n}}$. Plugging the latter into (7.7) we conclude that

$$\sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p}\boldsymbol{v}^T\boldsymbol{X}_\perp^T\boldsymbol{w}\le\eta b_n^2\sqrt{\frac{n_0}{n}}\sigma, \tag{7.8}$$

holds with probability at least $1-e^{-\frac{t^2}{2}}-\mathbb{P}(\|f(\boldsymbol{g})-\mu\boldsymbol{g}\|_{\ell_2}>\eta b_n\sigma)$. Combining (7.5) and (7.8) via the union bound together with (7.4) completes the proof of this lemma. ■

### 7.3 Proof of Theorem 3.1

Let us denote the error in our updates by $\boldsymbol{h}_\tau=\boldsymbol{\theta}_\tau-\mu\boldsymbol{\theta}^*$ and the "effective noise" by $\boldsymbol{w}=f(\boldsymbol{X}\boldsymbol{\theta})-\mu\boldsymbol{X}\boldsymbol{\theta}^*$. Further let $\mathcal{D}$ denote the descent set of the regularizer $\mathcal{R}$ at $\mu\boldsymbol{\theta}^*$ i.e. $\mathcal{D}=\{\boldsymbol{z}\big|\mathcal{R}(\mu\boldsymbol{\theta}^*+\boldsymbol{z})\le\mathcal{R}(\mu\boldsymbol{\theta}^*)\}$. Using the definition of $\boldsymbol{h}_\tau$ and $\boldsymbol{w}$ together with Lemma 7.2 allows us to conclude that

$$\begin{aligned}\boldsymbol{h}_{\tau+1}&=\mathcal{P}_{\mathcal{K}}\left(\mu\boldsymbol{\theta}^*+\left(\boldsymbol{I}-\alpha_\tau\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{h}_\tau-\alpha_\tau\boldsymbol{X}^T\boldsymbol{w}\right)-\mu\boldsymbol{\theta}^*,\\&=\mathcal{P}_{\mathcal{D}}\left(\left(\boldsymbol{I}-\alpha_\tau\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{h}_\tau-\alpha_\tau\boldsymbol{X}^T\boldsymbol{w}\right).\end{aligned}$$

Let $\kappa_{\mathcal{R}}$ be a constant that is equal to one for convex regularizers and equal to two for nonconvex regularizers. Furthermore, assume $\mathcal{C}$ is the cone of descent of $\mathcal{R}$ at $\mu\boldsymbol{\theta}^*$. Applying Lemmas 7.1 and 7.3 we conclude that

$$\begin{aligned}\|\boldsymbol{h}_{\tau+1}\|_{\ell_2}\le&\kappa_{\mathcal{R}}\left\|\mathcal{P}_{\mathcal{C}}\left(\left(\boldsymbol{I}-\alpha_\tau\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{h}_\tau-\alpha_\tau\boldsymbol{X}^T\boldsymbol{w}\right)\right\|_{\ell_2},\\\le&\kappa_{\mathcal{R}}\cdot\sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p}\boldsymbol{v}^T\left(\left(\boldsymbol{I}-\alpha_\tau\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{h}_\tau-\alpha_\tau\boldsymbol{X}^T\boldsymbol{w}\right),\\\le&\kappa_{\mathcal{R}}\left(\sup_{\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p}\boldsymbol{v}^T\left(\boldsymbol{I}-\alpha_\tau\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{h}_\tau+\alpha_\tau\cdot\sup_{\boldsymbol{v}\in-\mathcal{C}\cap\mathcal{B}^p}\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{w}\right),\\\le&\kappa_{\mathcal{R}}\left(\|\boldsymbol{h}_\tau\|_{\ell_2}\cdot\sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{C}\cap\mathcal{B}^p}\boldsymbol{v}^T\left(\boldsymbol{I}-\alpha_\tau\boldsymbol{X}^T\boldsymbol{X}\right)\boldsymbol{u}+\alpha_\tau\cdot\sup_{\boldsymbol{v}\in-\mathcal{C}\cap\mathcal{B}^p}\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{w}\right).\end{aligned}$$

Using $\alpha_\tau=1/b_n^2$ and combining Lemmas 7.4 and 7.6 to bound the first and second terms in the above inequality implies that

$$\|\boldsymbol{h}_{\tau+1}\|_{\ell_2}\le\sqrt{8\kappa_{\mathcal{R}}^2\frac{n_0}{n}}\|\boldsymbol{h}_\tau\|_{\ell_2}+\kappa_{\mathcal{R}}\frac{\eta\left(\sigma\sqrt{n_0}+\gamma\right)}{\sqrt{n}}.$$

holds with probability at least $1-10e^{-\frac{t^2}{8}}-p(\eta)$ for all $\tau$. By iteratively applying the latter inequality we conclude that with high probability

$$\begin{aligned}\|\boldsymbol{h}_\tau\|_{\ell_2}\le&\left(\sqrt{8\kappa_{\mathcal{R}}^2\frac{n_0}{n}}\right)^\tau\|\boldsymbol{h}_0\|_{\ell_2}+\kappa_{\mathcal{R}}\left(\sum_{k=0}^{\tau-1}\left(\sqrt{8\kappa_{\mathcal{R}}^2\frac{n_0}{n}}\right)^k\right)\frac{\eta\left(\sigma\sqrt{n_0}+\gamma\right)}{\sqrt{n}},\\\le&\left(\sqrt{8\kappa_{\mathcal{R}}^2\frac{n_0}{n}}\right)^\tau\|\boldsymbol{h}_0\|_{\ell_2}+\frac{\kappa_{\mathcal{R}}}{1-\left(\sqrt{8\kappa_{\mathcal{R}}^2\frac{n_0}{n}}\right)}\frac{\eta\left(\sigma\sqrt{n_0}+\gamma\right)}{\sqrt{n}},\end{aligned}$$

concluding the proof.

### 7.4 Proof of Theorem 3.2

Our proofs in this section is in part inspired by the analysis of the randomized Kaczmarz algorithm due to Strohmer and Vershynin [27]. To begin our proof first note that by Lemma 7.5

$$\|\boldsymbol{X}\boldsymbol{h}\|_{\ell_2}^2 = \sum_{i=1}^n (\boldsymbol{x}_i^*\boldsymbol{h})^2 \geq (b_n - b_{n_0})^2 \|\boldsymbol{h}\|_{\ell_2}^2,$$

holds for all $\boldsymbol{h} \in \mathcal{C} := \mathcal{C}_{\mathcal{R}}(\mu\boldsymbol{\theta}^*)$ with probability at least $1 - e^{-\frac{t^2}{2}}$. We now rewrite the latter equation in the alternative form

$$\sum_{i=1}^n \frac{\|\boldsymbol{x}_i\|_{\ell_2}^2}{\|\boldsymbol{X}\|_F^2}\left(\left\langle \boldsymbol{h}, \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_{\ell_2}}\right\rangle\right)^2 \geq \frac{(b_n - b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\|\boldsymbol{h}\|_{\ell_2}^2. \tag{7.9}$$

Now define a random vector $\boldsymbol{z}$ distributed such that $\boldsymbol{z} = \frac{\boldsymbol{x}_i}{\|\boldsymbol{x}_i\|_{\ell_2}}$ with probability $\frac{\|\boldsymbol{x}_i\|_{\ell_2}^2}{\|\boldsymbol{X}\|_F^2}$. Then, (7.9) can alternatively be written in the form

$$\mathbb{E}[(\boldsymbol{z}^T\boldsymbol{h})^2] \geq \frac{(b_n - b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\|\boldsymbol{h}\|_{\ell_2}^2. \tag{7.10}$$

Define $\boldsymbol{h}_\tau = \boldsymbol{\theta}_\tau - \mu\boldsymbol{\theta}^*$ and recall that $\boldsymbol{y}_i = \boldsymbol{w}_i + \mu\langle \boldsymbol{x}_i, \boldsymbol{\theta}^*\rangle$. Similar to the proof of Theorem 3.1 in Section 7.3 using the definition of $\boldsymbol{h}_\tau$ and $\boldsymbol{w}$ together with Lemma 7.2 allows us to conclude that

$$\boldsymbol{h}_{\tau+1} = \mathcal{P}_{\mathcal{D}}\left(\boldsymbol{h}_\tau - \langle \boldsymbol{z}_\tau, \boldsymbol{h}_\tau\rangle \boldsymbol{z}_\tau + \frac{w_{\psi_\tau}}{\|\boldsymbol{x}_{\psi_\tau}\|_{\ell_2}}\boldsymbol{z}_\tau\right),$$

where $\{\boldsymbol{z}_\tau\}_{\tau=1}^\infty$ are independent realizations of $\boldsymbol{z}$. First, note that $\mathbf{0} \in \mathcal{D}$. Now we utilize the fact that projection onto a convex set containing $\mathbf{0}$ can only decrease the Euclidean norm of a vector we conclude that

$$\begin{aligned}\|\boldsymbol{h}_{\tau+1}\|_{\ell_2}^2 &\leq \left\|\boldsymbol{h}_\tau - \langle \boldsymbol{z}_\tau, \boldsymbol{h}_\tau\rangle \boldsymbol{z}_\tau + \frac{w_{\psi_\tau}}{\|\boldsymbol{x}_{\psi_\tau}\|_{\ell_2}}\boldsymbol{z}_\tau\right\|_{\ell_2}^2, \\ &= \|\boldsymbol{h}_\tau\|_{\ell_2}^2 - (\langle \boldsymbol{z}_\tau, \boldsymbol{h}_\tau\rangle)^2 + \frac{w_{\psi_\tau}^2}{\|\boldsymbol{x}_{\psi_\tau}\|_{\ell_2}^2}.\end{aligned}$$

Utilizing (7.10) in the above inequality, conditioned on $\boldsymbol{h}_\tau$ we have

$$\mathbb{E}\left[\|\boldsymbol{h}_{\tau+1}\|_{\ell_2}^2 \,\middle|\boldsymbol{h}_\tau\right] = \mathbb{E}\left[\|\boldsymbol{h}_\tau\|_{\ell_2}^2 - (\langle \boldsymbol{z}_\tau, \boldsymbol{h}_\tau\rangle)^2 \,\middle|\boldsymbol{h}_\tau\right] + \frac{1}{n}\sum_{i=1}^n \frac{w_i^2}{\|\boldsymbol{x}_i\|_{\ell_2}^2} \leq \left(1 - \frac{(b_n - b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\right)\|\boldsymbol{h}_\tau\|_{\ell_2}^2 + \frac{1}{n}\sum_{i=1}^n \frac{w_i^2}{\|\boldsymbol{x}_i\|_{\ell_2}^2}.$$

Using the independence of the random variables $\boldsymbol{z}_\tau$ together with law of total expectation yields

$$\begin{aligned}\mathbb{E}\left[\|\boldsymbol{h}_{\tau+1}\|_{\ell_2}^2\right] &\leq \left(1 - \frac{(b_n - b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\right)\mathbb{E}\left[\|\boldsymbol{h}_\tau\|_{\ell_2}^2\right] + \frac{1}{n}\sum_{i=1}^n \frac{w_i^2}{\|\boldsymbol{x}_i\|_{\ell_2}^2}, \\ &\leq \left(1 - \frac{(b_n - b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\right)\mathbb{E}\left[\|\boldsymbol{h}_\tau\|_{\ell_2}^2\right] + \frac{1}{n}\frac{\|\boldsymbol{w}\|_{\ell_2}^2}{\min_i \|\boldsymbol{x}_i\|_{\ell_2}^2}.\end{aligned} \tag{7.11}$$

Standard concentration of Chi-squared random variables together with the union bound implies that with probability at least $1-ne^{-cp}$, $\min_i \|\boldsymbol{x}_i\|_{\ell_2}^2 \ge 0.996p$. Here, $c$ is a fixed numerical constant. Combining the latter with (7.11) implies that

$$\mathbb{E}\left[\,\|\boldsymbol{h}_{\tau+1}\|_{\ell_2}^2\,\right] \le \left(1-\frac{(b_n-b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\right)\mathbb{E}\left[\,\|\boldsymbol{h}_{\tau}\|_{\ell_2}^2\,\right] + 1.005\frac{\|\boldsymbol{w}\|_{\ell_2}^2}{np},$$

holds with probability at least $1-ne^{-cp}$. Using the fact that $\boldsymbol{g}=\boldsymbol{X}\boldsymbol{\theta}^*$ is a Gaussian random vector with $\mathcal{N}(0,1)$ entries, Definitions 2.4 and 2.5 immediately imply that

$$\mathbb{E}\left[\,\|\boldsymbol{h}_{\tau+1}\|_{\ell_2}^2\,\right] \le \left(1-\frac{(b_n-b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\right)\mathbb{E}\left[\,\|\boldsymbol{h}_{\tau}\|_{\ell_2}^2\,\right] + 1.005b_n^2\frac{\eta^2\sigma^2}{np},$$

holds with probability at least $1-ne^{-cp}-\mathbb{P}(\|f(\boldsymbol{g})-\mu\boldsymbol{g}\|_{\ell_2}>\eta b_n\sigma)$ for all $\tau$. By iteratively applying the latter inequality we conclude that with probability at least $1-ne^{-cp}-\mathbb{P}(\|f(\boldsymbol{g})-\mu\boldsymbol{g}\|_{\ell_2}>\eta b_n\sigma)$ we have

$$\begin{aligned}
\mathbb{E}\left[\,\|\boldsymbol{h}_{\tau}\|_{\ell_2}^2\,\right] &\le \left(1-\frac{(b_n-b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\right)^{\tau}\|\boldsymbol{h}_0\|_{\ell_2}^2 + 1.005\eta^2 b_n^2\left(\sum_{k=0}^{\tau-1}\left(1-\frac{(b_n-b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\right)^{k}\right)\frac{\eta^2\sigma^2}{np},\\
&\le \left(1-\frac{(b_n-b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\right)^{\tau}\|\boldsymbol{h}_0\|_{\ell_2}^2 + 1.005\frac{b_n^2}{1-\left(1-\frac{(b_n-b_{n_0})^2}{\|\boldsymbol{X}\|_F^2}\right)}\frac{\eta^2\sigma^2}{np},\\
&= \left(1-b_n^2\frac{(1-\frac{b_{n_0}}{b_n})^2}{\|\boldsymbol{X}\|_F^2}\right)^{\tau}\|\boldsymbol{h}_0\|_{\ell_2}^2 + 1.005\frac{\|\boldsymbol{X}\|_F^2}{\left(1-\frac{b_{n_0}}{b_n}\right)^2}\frac{\eta^2\sigma^2}{np}.
\end{aligned}\tag{7.12}$$

By [20, Lemma 6.9] we have $\frac{b_{n_0}}{b_n}\le\sqrt{\frac{n_0}{n}}$. Also by standard concentration of Chi-squared random variables with probability at least $1-e^{-cnp}$, $\|\boldsymbol{X}\|_F^2\le 1.0049np$. Furthermore, $b_n^2\ge\frac{1.0049}{2}n$ for all $n\ge1$. Plugging the latter three inequalities into (7.12) implies that

$$\mathbb{E}\left[\,\|\boldsymbol{h}_{\tau}\|_{\ell_2}^2\,\right] \le \left(1-\frac{\left(1-\sqrt{\frac{n_0}{n}}\right)^2}{2p}\right)^{\tau}\|\boldsymbol{h}_0\|_{\ell_2}^2 + \frac{1.01}{\left(1-\sqrt{\frac{n_0}{n}}\right)^2}\eta^2\sigma^2,$$

holds with probability at least $1-(n+1)e^{-cp}-p(\eta)$.

# References

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

[2] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, 2014.

[3] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

[4] M. Bayati and A. Montanari. The lasso risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2012.

[5] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[6] J. J. Bruer, J. A. Tropp, V. Cevher, and S. Becker. Time–data tradeoffs by aggressive smoothing. In *Advances in Neural Information Processing Systems*, pages 1664–1672, 2014.

[7] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[8] E. J. Candes and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

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

[10] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[11] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16(8):2080–2095, 2007.

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

[13] D. L. Donoho, A. Maleki, and A. Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[14] R. Eghbali and M. Fazel. Decomposable norm minimization with proximal-gradient homotopy algorithm. *arXiv preprint arXiv:1501.06711*, 2015.

[15] R. Foygel and L. Mackey. Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247, 2014.

[16] Y. Gordon. *On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$*. Springer, 1988.

[17] P. Li and T. Zhang. Gradient hard thresholding pursuit for sparsity-constrained optimization. 2014.

[18] C. A. Metzler, A. Maleki, and R. G. Baraniuk. From denoising to compressed sensing. *arXiv preprint arXiv:1406.4175*, 2014.

[19] S. Negahban, B. Yu, M. J. Wainwright, and P. 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.

[20] S. Oymak, B. Recht, and M. Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

[21] Samet Oymak and Babak Hassibi. Sharp mse bounds for proximal denoising. *arXiv preprint arXiv:1305.2714*, 2013.

[22] Y. Plan and R. Vershynin. The generalized lasso with non-linear observations. *arXiv preprint arXiv:1502.04071*, 2015.

[23] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[24] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *Information Theory, IEEE Transactions on*, 59(1):482–494, 2013.

[25] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[26] B. Recht, C. Re, S. Wright, and F. Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.

[27] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262–278, 2009.

[28] C. Thrampoulidis, E. Abbasi, and B. Hassibi. The lasso with non-linear measurements is equivalent to one with linear measurements. *arXiv preprint arXiv:1506.02181*, 2015.

[29] C. Thrampoulidis, S. Oymak, and B. Hassibi. Simple error bounds for regularized noisy linear inverse problems. *arXiv preprint arXiv:1401.6578*, 2014.

[30] R. Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance*, pages 3–66. Springer, 2015.

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

[32] Z. Yang, Z. Wang, H. Liu, and T. Eldar, Y. C.and Zhang. Sparse nonlinear regression: Parameter estimation and asymptotic inference. *arXiv preprint arXiv:1511.04514*, 2015.

[33] X. Yi, Z. Wang, C. Caramanis, and H. Liu. Optimal linear estimation under unknown nonlinear transform. *arXiv preprint arXiv:1505.03257*, 2015.

# A Penalized formulation via resampling

In this section we aim to understand the convergence properties of a penalized variant of (5.1) for nonlinear parameter estimation problems. Concretely, we focus on solving the following optimization problem

$$\min_{\theta\in\mathbb{R}^p} \ \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{X}\theta\right\|_{\ell_2}^2+\lambda\mathcal{R}(\theta), \tag{A.1}$$

with $\lambda$ a nonnegative regularization parameter. A common approach to solve the penalized formulation is via Proximal Gradient Descent (ProxGD). Starting from an initial estimate $\theta_0$, ProxGD iteratively applies the update

$$\theta_{\tau+1}=\mathbf{prox}_{\lambda_\tau}\left(\theta_\tau+\alpha_\tau\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\theta_\tau)\right). \tag{A.2}$$

Here, $\alpha_\tau$ is the step size and $\mathbf{prox}_\lambda$ is the proximal function associated with $\mathcal{R}$ and is defined as

$$\mathbf{prox}_\lambda(\boldsymbol{z})=\arg\min_{\bar{\boldsymbol{z}}}\frac{1}{2}\left\|\boldsymbol{z}-\bar{\boldsymbol{z}}\right\|_{\ell_2}^2+\lambda\mathcal{R}(\bar{\boldsymbol{z}}).$$

In this section we wish to understand the properties of the proximal gradient iterations (A.2) for nonlinear parameter estimation problems with Gaussian features. To gain some insights into the performance of the update (A.2) we study a variant of this update where in each iteration we use a fresh set of observations and feature vectors. In particular, we assume that we run the updates

$$\theta_{\tau+1}=\mathbf{prox}_{\lambda_\tau}\left(\theta_\tau+\alpha_\tau\boldsymbol{X}_\tau^T(\boldsymbol{y}_\tau-\boldsymbol{X}_\tau\theta_\tau)\right). \tag{A.3}$$

Here $\{\boldsymbol{X}_\tau\}_{\tau=1}^\infty$ are i.i.d. mini-batches of Gaussian features with $\boldsymbol{X}_\tau\in\mathbb{R}^{n\times p}$ and $\boldsymbol{y}_\tau=f(\boldsymbol{X}_\tau\theta)$ are mini-batches of nonlinear observations. We emphasize that such an approach is not useful in practice as one often wishes to reuse the measurements and samples across all iterations as in the update (A.2). Nevertheless, we hope that such an analysis provides useful insights into the performance of (A.2) and the key parameters involved in its convergence.

When dealing with the penalized version of the problem the definition of minimal number of samples as discussed in Definition 2.3 is no longer adequate as this minimal number of samples would not only depend on the regularization function $\mathcal{R}$ but also the regularization parameter $\lambda$ in (A.1). In this case the minimal sample complexity is no longer based on the notion of Gaussian width but rather a closely related quantity of Gaussian distance defined below.

**Definition A.1 (Gaussian distance)** *Let $\boldsymbol{g} \in \mathbb{R}^p$ be a random Gaussian vector with i.i.d. $\mathcal{N}(0,1)$ entries. Also assume $\mathcal{R} : \mathbb{R}^p \to \mathbb{R}$ is a regularization function with closed sub-level sets. For a regularization function $\mathcal{R}$ at a point $\boldsymbol{\theta}^*$ we define the Gaussian distance at level $\lambda$ as*

$$\mathcal{G}(\mathcal{R}, \boldsymbol{\theta}^*, \lambda) := \sqrt{\mathbb{E}[\text{dist}^2(\boldsymbol{g}, \lambda\partial\mathcal{R}(\boldsymbol{\theta}^*))]}.$$

*Here, $\partial\mathcal{R}(\boldsymbol{\theta}^*)$ is the sub-differential of $\mathcal{R}$ at $\boldsymbol{\theta}^*$. Also for a vector $\boldsymbol{z} \in \mathbb{R}^p$ and a set $\mathcal{C} \in \mathbb{R}^p$ the distance function dist$(\boldsymbol{z}, \mathcal{C})$ is the Eucleadian distance between the point $\boldsymbol{z}$ and the set $\mathcal{C}$ i.e. dist$(\boldsymbol{z}, \mathcal{C}) = \inf_{\bar{\boldsymbol{z}} \in \mathcal{C}} \|\boldsymbol{z} - \bar{\boldsymbol{z}}\|_{\ell_2}$. We will often use the shorthand $\mathcal{G}(\lambda)$ with the dependence on $\mathcal{R}$ and $\boldsymbol{\theta}^*$ implied.*

The Gaussian distance defined above is closely related to the notion of mean width. In fact one can show that $\omega(\mathcal{C}_{\mathcal{R}}(\boldsymbol{\theta}^*) \cap \mathcal{B}^p) \approx \min_\lambda \mathcal{G}(\mathcal{R}, \boldsymbol{\theta}^*, \lambda)$ [2, 15]. With this definition in place we are ready to explain the minimal sample complexity for the regularized case.

**Definition A.2 (minimal number of samples with regularization)** *Let $\partial\mathcal{R}(\boldsymbol{\theta}^*)$ be the sub-differential of the regularization function $\mathcal{R}$ at $\boldsymbol{\theta}^*$ and set $\mathcal{G}(\lambda) = \mathcal{G}(\mathcal{R}, \boldsymbol{\theta}^*, \lambda)$. We define the minimal sample function as*

$$\mathcal{M}_\lambda(\mathcal{R}, \boldsymbol{\theta}^*, t) = \phi^{-1}(\mathcal{G}(\lambda) + 7t + \sqrt{2}) \approx (\mathcal{G}(\lambda) + 7t + \sqrt{2})^2.$$

*We shall often use the short hand $\boldsymbol{n}_0(\lambda) = \mathcal{M}_\lambda(\mathcal{R}, \boldsymbol{\theta}^*, t)$ with the dependence on $\mathcal{R}, \boldsymbol{\theta}^*, t$ implied. We note that for convex functions $\mathcal{R}$ based on [29] $\boldsymbol{n}_0(\lambda)$ is exactly the minimum number of samples required for the estimator (1.1) to succeed in recovering an unknown parameter $\boldsymbol{\theta}^*$ with high probability from linear measurements $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}^*$.*

We note that the regularized variant of the minimal sample complexity $\boldsymbol{n}_0(\lambda)$ is intimately related to the version without regularization $n_0$. In fact $n_0 \approx \min_\lambda \boldsymbol{n}_0(\lambda)$.

Before we state our result for the proximal iterations we would like to mention that the proximal gradient algorithm will not work if we set the shrinkage parameter $\lambda_\tau$ to be a constant in each iteration rather $\lambda_\tau$ should decay with the iterations.[5] In this paper we recommend a particular strategy for updating $\lambda_\tau$. Starting with tuning parameters $M_0$ and $\rho$ we set the shrinkage parameters $\lambda_\tau$ via the following set of recursions

$$\lambda_\tau = \frac{((1 + \frac{t}{b_n})M_\tau + \eta\sigma)\lambda}{b_n} \quad \text{where} \quad M_{\tau+1} = \rho M_\tau + \frac{\eta(\sigma\sqrt{\boldsymbol{n}_0(\lambda)} + \gamma)}{\sqrt{n}}. \tag{A.4}$$

Observe that $M_\tau$ satisfies the bound

$$M_\tau \le \rho^\tau M_0 + \frac{\eta(\sigma\sqrt{\boldsymbol{n}_0(\lambda)} + \gamma)}{\sqrt{n}}. \tag{A.5}$$

We note that such tuning strategies our quite common in the optimization literature. In particular a related tuning strategy is utilized in the updates of the AMP algorithm [4].

Now that we have described our strategy for tuning the shrinkage parameter we are ready to state our main result for the proximal gradient scheme.

[5] The reason for this will become clear to the reader by consider the case when $f$ is a linear function and there is no noise.

**Theorem A.3** *Let $f:\mathbb{R}\rightarrow\mathbb{R}$ be an unknown nonlinear function. Also, for $\tau=1,2,\ldots$ let $\boldsymbol{y}_\tau=f(\boldsymbol{X}_\tau\boldsymbol{\theta}^*)\in\mathbb{R}^n$ be $n$ nonlinear observations from $\boldsymbol{\theta}^*$ with the feature matrix $\boldsymbol{X}_\tau\in\mathbb{R}^{n\times p}$ consisting of i.i.d. $\mathcal{N}(0,1)$ entries. Furthermore, let $\mathcal{R}$ be a convex regularizer and let $\boldsymbol{n}_0(\lambda)=\mathcal{M}_\lambda(\mathcal{R},\mu\boldsymbol{\theta}^*,t)$ be the minimal number of data samples as per Definition A.2 and assume $0\le t\le b_n$. Assume that*

$$n>\boldsymbol{n}_0(\lambda).$$

*Also let $\mu,\sigma$, and $\eta$ be the nonlinearity parameters per definition 2.4 and let $\lambda$ be the regularization parameter in (A.1). Assume we start from an initial estimate and utilize the tuning strategy discussed in (A.4) with tuning parameters obeying the following conditions*

$$M_0\ge\|\boldsymbol{\theta}_0-\mu\boldsymbol{\theta}^*\|_{\ell_2}\quad\text{and}\quad\rho\ge\sqrt{\frac{\boldsymbol{n}_0(\lambda)}{n}}.$$

*Then, the proximal gradient iterations (A.3) with step size $\alpha_\tau=1/b_n^2\approx1/n$ obeys*

$$\|\boldsymbol{\theta}_\tau-\mu\boldsymbol{\theta}_0\|_{\ell_2}\le M_\tau,\tag{A.6}$$

*for all $\tau$ with probability at least $1-\tau\left(2p(\eta)+7\exp(-t^2/2)\right)$. Here, $p(\eta)$ is the concentration probability function as per Definition 2.5.*

Theorem A.3 can be connected to our main results, for example Theorem 3.1. Observe that exponentially decaying error bounds (A.5) and (3.3) have essentially identical forms. In particular, as discussed previously if $\lambda$ is chosen to be the minimizer of $\boldsymbol{n}_0(\lambda)$ then $n_0\approx\boldsymbol{n}_0(\lambda_{\min})$ and if we set $\rho=\sqrt{n_0/n}$ in the proximal iterations the convergence results of the two theorem are the same up to constant factors. In fact, the convergence rate of the theorem above is sharper and has a constant of one. This is due to the fact that we can provide sharper constants with the resampling framework. We again emphasize that while we make use of a resampling strategy such an approach is not practical and one usually wishes to run the update (A.2). The reader is referred to [3, 13, 14, 31] for related works in this direction. In particular, Xiao and Zhang [31] study the particular case of $\mathcal{R}(\boldsymbol{x})=\|\boldsymbol{x}\|_{\ell_1}$ and very recently, Erghbali and Fazel [14] have more general results that apply to the case where $\mathcal{R}(\boldsymbol{x})$ is a decomposable norm and $\lambda$ is sufficiently large. We note that in contrast with our results these publications do not provide sharp constants with the exception of the AMP algorithm [3, 13, 18]. However, rigorous results for the AMP algorithm are limited to linear estimation and separable regularizers $\mathcal{R}$.

# B Proofs for penalized formulation (Proof of Theorem A.3)

In this section we will prove our result for the penalized formulation, mainly Theorem A.3. Before diving into the details of the proof of Theorem A.3 we state and prove a few useful lemmas in Section B.1. We then utilize these results to complete the proof of Theorem A.3 in Section B.2.

## B.1 Preliminary lemmas for regularized estimation

In this section we state and prove a few results about proximal functions and Gaussian distances.

**Lemma B.1 (e.g. [21])** *Let* $\textbf{prox}_\lambda$ *be the proximal function associated with $\mathcal{R}$ defined as*

$$\textbf{prox}_\lambda(\boldsymbol{z})=\arg\min_{\bar{\boldsymbol{z}}}\frac{1}{2}\|\boldsymbol{z}-\bar{\boldsymbol{z}}\|_{\ell_2}^2+\lambda\mathcal{R}(\bar{\boldsymbol{z}}).$$

*Then,*

$$\|\textbf{prox}_\lambda(\boldsymbol{x}+\boldsymbol{h})-\boldsymbol{x}\|_{\ell_2}\le\text{dist}(\boldsymbol{h},\lambda\partial\mathcal{R}(\boldsymbol{x})).$$

The next lemma proves a useful property about the Gaussian distance.

**Lemma B.2** *Let $\mathcal{C} \subset \mathbb{R}^n$ be a compact and convex set. Then for a Gaussian random vector $\boldsymbol{g}$ with i.i.d. $\mathcal{N}(0,1)$ entries, $\mathbb{E}[dist^2(\alpha\boldsymbol{g},\mathcal{C})]$ is a nondecreasing function of $\alpha$ on $[0,\infty)$.*

**Proof** Since $\boldsymbol{g}$ has a symmetric distribution around 0, it suffices to show $\mathbb{E}[\text{dist}^2(\alpha\boldsymbol{g},\mathcal{C})] + \mathbb{E}[\text{dist}^2(-\alpha\boldsymbol{g},\mathcal{C})]$ is nondecreasing. Differentiating the square of the distance with respect to $\alpha$, we have

$$\frac{\partial\,\mathbb{E}[\text{dist}^2(\alpha\boldsymbol{g},\mathcal{C})]}{\partial\alpha} = 2\,\mathbb{E}[\langle\alpha\boldsymbol{g} - \mathcal{P}_{\mathcal{C}}(\alpha\boldsymbol{g}),\boldsymbol{g}\rangle].$$

Now, observe that

$$\langle\alpha\boldsymbol{g} - \mathcal{P}_{\mathcal{C}}(\alpha\boldsymbol{g}),\boldsymbol{g}\rangle + \langle-\alpha\boldsymbol{g} - \mathcal{P}_{\mathcal{C}}(-\alpha\boldsymbol{g}),-\boldsymbol{g}\rangle = 2\alpha\left\|\boldsymbol{g}\right\|_{\ell_2}^2 - \langle\mathcal{P}_{\mathcal{C}}(\alpha\boldsymbol{g}) - \mathcal{P}_{\mathcal{C}}(-\alpha\boldsymbol{g}),\boldsymbol{g}\rangle\,.$$

Since $\mathcal{C}$ is convex, $\langle\mathcal{P}_{\mathcal{C}}(\alpha\boldsymbol{g}) - \mathcal{P}_{\mathcal{C}}(-\alpha\boldsymbol{g}),\boldsymbol{g}\rangle \le \|\mathcal{P}_{\mathcal{C}}(\alpha\boldsymbol{g}) - \mathcal{P}_{\mathcal{C}}(-\alpha\boldsymbol{g})\|_{\ell_2}\|\boldsymbol{g}\|_{\ell_2} \le 2\alpha\left\|\boldsymbol{g}\right\|_{\ell_2}^2$. Hence

$$\frac{\partial\,\mathbb{E}[\text{dist}^2(\alpha\boldsymbol{g},\mathcal{C})]}{\partial\alpha} + \frac{\partial\,\mathbb{E}[\text{dist}^2(-\alpha\boldsymbol{g},\mathcal{C})]}{\partial\alpha} \ge 0,$$

concluding the proof. ■

Next we state a lemma that is useful for understanding the distance and projection properties of a Gaussian vector onto a subspace.

**Lemma B.3** *Assume $\boldsymbol{g}$ is a Gaussian random vector with i.i.d. $\mathcal{N}(0,1)$ entries. Let $\mathcal{C} \in \mathbb{R}^n$ be an arbitrary set that contains the origin and let $S \in \mathbb{R}^n$ be a subspace of dimension $n-d$. Then the following inequalities hold*

$$\mathbb{E}[\text{dist}(\mathcal{P}_S(\boldsymbol{g}),\mathcal{C})] \le \mathbb{E}[\text{dist}(\boldsymbol{g},\mathcal{C})] + \sqrt{d}, \tag{B.1}$$

$$\mathbb{E}[\sup_{\boldsymbol{v}\in\mathcal{C}}\langle\mathcal{P}_S(\boldsymbol{g}),\boldsymbol{v}\rangle] \le \mathbb{E}[\sup_{\boldsymbol{v}\in\mathcal{C}}\langle\boldsymbol{g},\boldsymbol{v}\rangle]. \tag{B.2}$$

**Proof** First note that $\mathbb{E}[\|\boldsymbol{g} - \mathcal{P}_S(\boldsymbol{g})\|_{\ell_2}] \le \sqrt{d}$. Consequently, by the triangular inequality for distance to sets

$$\mathbb{E}[\text{dist}(\mathcal{P}_S(\boldsymbol{g}),\mathcal{C})] \le \mathbb{E}[\text{dist}(\boldsymbol{g},\mathcal{C})] + \mathbb{E}[\|\boldsymbol{g} - \mathcal{P}_S(\boldsymbol{g})\|_{\ell_2}] \le \mathbb{E}[\text{dist}(\boldsymbol{g},\mathcal{C})] + \sqrt{d}.$$

Now let $\hat{\boldsymbol{v}} = \arg\sup_{\boldsymbol{v}\in\mathcal{C}}\langle\mathcal{P}_S(\boldsymbol{g}),\boldsymbol{v}\rangle$. Since $\mathcal{P}_S(\boldsymbol{g})$ and $\boldsymbol{g} - \mathcal{P}_S(\boldsymbol{g})$ are independent, $\hat{\boldsymbol{v}}$ and $\boldsymbol{g} - \mathcal{P}_S(\boldsymbol{g})$ are independent as well. Consequently

$$\mathbb{E}[\sup_{\boldsymbol{v}\in\mathcal{C}}\langle\mathcal{P}_S(\boldsymbol{g}),\boldsymbol{v}\rangle] = \mathbb{E}[\langle\mathcal{P}_S(\boldsymbol{g}),\hat{\boldsymbol{v}}\rangle] = \mathbb{E}[\langle\boldsymbol{g},\hat{\boldsymbol{v}}\rangle] \le \mathbb{E}[\sup_{\boldsymbol{v}\in\mathcal{C}}\langle\boldsymbol{g},\boldsymbol{v}\rangle],$$

concluding the proof. ■

## B.2 Proof of Theorem A.3

We first provide the convergence result for a single step of the proximal iteration in the lemma below whose proof is differed to Section B.2.1.

**Lemma B.4 (Single step estimator)** *Let $\boldsymbol{\theta}_\tau$ be the estimate at the $\tau$th iteration with the associated error $\boldsymbol{h}_\tau = \boldsymbol{\theta}_\tau - \mu\boldsymbol{\theta}^*$ obeying $\|\boldsymbol{h}_\tau\|_{\ell_2} \le M$ for some $M \ge 0$. Given $\lambda \ge 0$ and $0 \le t \le b_n$, pick $\lambda_\tau = \frac{\lambda((b_n+t)M + b_n\eta\sigma)}{b_n^2}$. In order to estimate $\mu\boldsymbol{\theta}^*$ from $\boldsymbol{y} = f(\boldsymbol{X}\boldsymbol{\theta}^*)$ consider the following update*

$$\boldsymbol{\theta}_{\tau+1} = \mathbf{prox}_{\lambda_\tau}\left(\boldsymbol{\theta}_\tau + \frac{1}{b_n^2}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_\tau)\right).$$

*Let* $\boldsymbol{n}_0(\lambda) = \mathcal{M}_\lambda(\mathcal{R}, \mu\boldsymbol{\theta}^*, t)$ *be the minimal number of data samples as per Definition A.2 and assume* $0 \le t \le b_n$*. If* $\boldsymbol{X}$ *has i.i.d.* $\mathcal{N}(0,1)$ *entries, then with probability at least* $1 - 2p(\eta) - 7\exp(-t^2/2)$*,* $\boldsymbol{\theta}_{\tau+1}$ *obeys*

$$\|\boldsymbol{\theta}_{\tau+1} - \mu\boldsymbol{\theta}^*\|_{\ell_2} \le \sqrt{\frac{\boldsymbol{n}_0(\lambda)}{n}} M + \eta\sqrt{\frac{\boldsymbol{n}_0(\lambda)}{n}}\sigma + \eta\gamma/\sqrt{n}.$$

With this lemma in hand we are ready to prove Theorem A.3. We shall show this by induction. Suppose the residual obeys $\|\boldsymbol{h}_\tau\|_{\ell_2} \le M_\tau$ with probability at least $1 - \tau P$ where $P = 2p(\eta) + 7\exp(-\frac{t^2}{2})$. Now, observe that the particular choice of $\lambda_\tau$ (as a function of $M_\tau$) makes Proposition B.4 applicable.

Using the fact that new samples are independent of the rest, Lemma B.4 implies that

$$\|\boldsymbol{h}_{\tau+1}\|_{\ell_2} \le \sqrt{\frac{\boldsymbol{n}_0(\lambda)}{n}} M_\tau + \eta\sqrt{\frac{\boldsymbol{n}_0(\lambda)}{n}}\sigma + \eta\gamma/\sqrt{n} \le M_{\tau+1},$$

holds with probability at least $1 - P$. Applying the union bound, $\|\boldsymbol{h}_i\|_{\ell_2} \le M_i$ holds for all $0 \le i \le \tau + 1$ with probability at least $1 - (\tau+1)P$, completing the proof. All that remains now is to complete the proof of Lemma B.4 which is the subject of the next section.

### B.2.1 Proof of Lemma B.4

**Proof** Define $\boldsymbol{w} = f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*$. The term inside the proximal operator can be rewritten as

$$\begin{aligned}\boldsymbol{\theta}_\tau + \frac{1}{b_n^2}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}_\tau) &= \mu\boldsymbol{\theta}^* + \left(\boldsymbol{h}_\tau - \frac{1}{b_n^2}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{h}_\tau\right) + \frac{1}{b_n^2}\boldsymbol{X}^T(\boldsymbol{y} - \mu\boldsymbol{X}\boldsymbol{\theta}^*).\\ &= \mu\boldsymbol{\theta}^* + \left(\boldsymbol{h}_\tau - \frac{1}{b_n^2}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{h}_\tau\right) + \frac{1}{b_n^2}\boldsymbol{X}^T\boldsymbol{w}. \end{aligned} \tag{B.3}$$

Note that $\mu\boldsymbol{\theta}^*$ is the term we wish our iterates to converge to and the remaining terms can be viewed as noise. Define $\mathcal{S}$ to be the $n-2$ dimensional subspace perpendicular to $\boldsymbol{\theta}^*, \boldsymbol{h}_\tau$. Given $\boldsymbol{v} \in \mathcal{S}^\perp$, let $\boldsymbol{v}^\perp$ be the projection of $\boldsymbol{v}$ onto the direction perpendicular to $\boldsymbol{v}$ and $\mathcal{S}$. The noise terms will be split into three terms by using $\boldsymbol{X}^T = \mathcal{P}_\mathcal{S}(\boldsymbol{X}^T) + \mathcal{P}_{\boldsymbol{v}^\perp}(\boldsymbol{X}^T) + \mathcal{P}_{\boldsymbol{v}}(\boldsymbol{X}^T)$.

- $\boldsymbol{e}_1 = \mathcal{P}_\mathcal{S}(\boldsymbol{X}^T)(-\boldsymbol{X}\boldsymbol{h}_\tau + (f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*))$.
- $\boldsymbol{e}_2 = \mathcal{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X}^T)(f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*) + \mathcal{P}_{(\boldsymbol{\theta}^*)^\perp}(\boldsymbol{X}^T)(f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*)$.
- $\boldsymbol{e}_3 = b_n^2\boldsymbol{h}_\tau - \mathcal{P}_{\boldsymbol{h}_\tau}(\boldsymbol{X}^T)\boldsymbol{X}\boldsymbol{h}_\tau - \mathcal{P}_{\boldsymbol{h}_\tau^\perp}(\boldsymbol{X}^T)\boldsymbol{X}\boldsymbol{h}_\tau$.

With this notation

$$\boldsymbol{\theta}_{\tau+1} = \mathbf{prox}_{\lambda_\tau}\left(\mu\boldsymbol{\theta}^* + \frac{1}{b_n^2}(\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3)\right).$$

We next relate the proximal estimator to the subdifferential via Lemma B.1. This yields

$$\begin{aligned}\|\boldsymbol{\theta}_{\tau+1} - \mu\boldsymbol{\theta}^*\|_{\ell_2} &\le \operatorname{dist}\left(\frac{1}{b_n^2}(\boldsymbol{e}_1 + \boldsymbol{e}_2 + \boldsymbol{e}_3), \lambda_\tau\partial\mathcal{R}(\mu\boldsymbol{\theta}^*)\right).\\ &\le \frac{1}{b_n^2}\left(\operatorname{dist}\left(\boldsymbol{e}_1, b_n^2\lambda_\tau\partial\mathcal{R}(\mu\boldsymbol{\theta}^*)\right) + \|\boldsymbol{e}_2\|_{\ell_2} + \|\boldsymbol{e}_3\|_{\ell_2}\right).\end{aligned}$$

We now proceed by estimating each of these terms. We will show that $\boldsymbol{e}_2$ and $\boldsymbol{e}_3$ are fairly small and we will obtain a bound for the term involving $\boldsymbol{e}_1$.

**Estimating $\boldsymbol{e}_2$:** To estimate $\boldsymbol{e}_2$ we use the fact that $\|f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*\|_{\ell_2} \le \eta b_n\sigma$ and

$$\left\|\mathcal{P}_{\boldsymbol{\theta}^*}(\boldsymbol{X}^T)\left(f(\boldsymbol{X}\boldsymbol{\theta}^*) - \mu\boldsymbol{X}\boldsymbol{\theta}^*\right)\right\|_{\ell_2} \le \eta b_n^2\gamma/\sqrt{n}, \tag{B.4}$$

holds with probability at least $1-p(\eta)$. Next we use the fact that $(\boldsymbol{\theta}^*)^{\perp^T}\boldsymbol{X}^T$ is an i.i.d. normal random vector and is independent of $\boldsymbol{X}\boldsymbol{\theta}^*$ to conclude that

$$\left\|\mathcal{P}_{(\boldsymbol{\theta}^*)^\perp}(\boldsymbol{X}^T)\left(f(\boldsymbol{X}\boldsymbol{\theta}^*)-\mu\boldsymbol{X}\boldsymbol{\theta}^*\right)\right\|_{\ell_2}\le t\sigma\eta b_n, \tag{B.5}$$

holds with probability at least $1-\exp(-t^2/2)$. Combining (B.4) and (B.5) together with the union bound yields

$$\|\boldsymbol{e}_2\|_{\ell_2}\le\eta b_n(t\sigma+\gamma b_n/\sqrt{n}). \tag{B.6}$$

**Estimating $\boldsymbol{e}_3$:** We now bound the term involving $\boldsymbol{e}_3$. For $t\le b_n$, with probability at least $1-3\exp(-t^2/2)-\exp(-b_n^2/2)\ge 1-4\exp(-t^2/2)$, the followings identities hold. First, using an independence argument again $\left\|\mathcal{P}_{\boldsymbol{h}_\tau^\perp}(\boldsymbol{X}^T)\boldsymbol{X}\boldsymbol{h}_\tau\right\|_{\ell_2}\le 2tb_n\|\boldsymbol{h}_\tau\|_{\ell_2}$. Second, $|\,\|\boldsymbol{X}\boldsymbol{h}_\tau\|_{\ell_2}-b_n\|\boldsymbol{h}_\tau\|_{\ell_2}|\le t\|\boldsymbol{h}_\tau\|_{\ell_2}$. Using this, it follows that

$$\left\|b_n^2\boldsymbol{h}_\tau-\mathcal{P}_{\boldsymbol{h}_\tau}(\boldsymbol{X}^T)\boldsymbol{X}\boldsymbol{h}_\tau\right\|_{\ell_2}\le\|\boldsymbol{h}_\tau\|_{\ell_2}(2b_nt+t^2)\le 3tb_n\|\boldsymbol{h}_\tau\|_{\ell_2}.$$

Combining these identities we arrive at

$$\|\boldsymbol{e}_3\|_{\ell_2}\le 5tb_n\|\boldsymbol{h}_\tau\|_{\ell_2}. \tag{B.7}$$

Combining bounds (B.6) and (B.7) which involve $\mathrm{e}_2$ and $\mathrm{e}_3$, with probability at least $1-5\exp(-t^2/2)-p(\eta)$ we have

$$b_n^{-2}(\|\mathrm{e}_2\|_{\ell_2}+\|\mathrm{e}_3\|_{\ell_2})\le b_n^{-1}[5t\|\boldsymbol{h}_\tau\|_{\ell_2}+\eta(t\sigma+\gamma b_n/\sqrt{n})]\le(\eta/b_n)[t(5\eta^{-1}\|\boldsymbol{h}_\tau\|_{\ell_2}+\sigma)+\gamma b_n/\sqrt{n}]. \tag{B.8}$$

We note that this bound grows as $b_n^{-1}$.

**Estimating $\boldsymbol{e}_1$:** The remaining term is $\boldsymbol{e}_1$. Define $\boldsymbol{a}:=\boldsymbol{X}\boldsymbol{h}_\tau+f(\boldsymbol{X}\boldsymbol{\theta}^*)-\mu\boldsymbol{X}\boldsymbol{\theta}^*$. Observe that with probability at least $1-\exp(-t^2/2)-p(\eta)$

$$\sigma_{tot}:=\|\boldsymbol{a}\|_{\ell_2}\le(b_n+t)\|\boldsymbol{h}_\tau\|_{\ell_2}+\eta b_n\sigma\le(b_n+t)M+\eta b_n\sigma:=\sigma_{up}.$$

Note that $\boldsymbol{h}_\tau,\boldsymbol{\theta}^*\in\mathcal{S}^\perp$. Thus, conditioned on $\boldsymbol{a}$, $\mathcal{P}_{\mathcal{S}}(\boldsymbol{X}^T)\boldsymbol{a}$ is statistically identical to $\boldsymbol{g}'=\mathcal{P}_{\mathcal{S}}(\boldsymbol{g})$ with $\boldsymbol{g}\sim\mathcal{N}(0,\sigma_{tot}^2\boldsymbol{I}_n)$. Now, applying Lemma B.3

$$\mathbb{E}[\operatorname{dist}(\mathcal{P}_{\mathcal{S}}(\boldsymbol{X}^T)[\boldsymbol{X}\boldsymbol{h}_\tau+f(\boldsymbol{X}\boldsymbol{\theta}^*)-\mu\boldsymbol{X}\boldsymbol{\theta}^*],b_n^2\lambda_\tau\partial f(\boldsymbol{x}))]\le\mathbb{E}[\operatorname{dist}(\boldsymbol{g},b_n^2\lambda_\tau\partial f(\boldsymbol{x}))]+\sqrt{2}\sigma_{tot}. \tag{B.9}$$

The problem is reduced to upper bounding $\mathbb{E}[\operatorname{dist}(\boldsymbol{g},m\lambda_\tau\partial f(\boldsymbol{x}))]$. Using Lemma B.2

$$\begin{aligned}
\mathbb{E}[\operatorname{dist}(\boldsymbol{g},b_n^2\lambda_\tau\partial f(\boldsymbol{x}))]&\le\sqrt{\mathbb{E}[\operatorname{dist}(\boldsymbol{g},b_n^2\lambda_\tau\partial f(\boldsymbol{x}))^2]}\\
&=\sigma_{up}\sqrt{\mathbb{E}[\operatorname{dist}(\sigma_{up}^{-1}\boldsymbol{g},\sigma_{up}^{-1}b_n^2\lambda_\tau\partial f(\boldsymbol{x}))^2]}\\
&=\sigma_{up}\sqrt{\mathbb{E}[\operatorname{dist}(\sigma_{up}^{-1}\boldsymbol{g},\lambda\partial f(\boldsymbol{x}))^2]}\\
&\le\sigma_{up}\sqrt{\mathcal{G}(\lambda)}, \qquad\qquad (\text{B.10})
\end{aligned}$$

where (B.10) follows from the definition of $\lambda_\tau$ and Lemma B.2. Merging this with (B.9), together with the union bound implies that with probability at least $1-2\exp(-t^2/2)-p(\eta)$

$$\begin{aligned}
\operatorname{dist}(\mathcal{P}_{\mathcal{S}}(\boldsymbol{X}^T)[\boldsymbol{X}\boldsymbol{h}_\tau+f(\boldsymbol{X}\boldsymbol{\theta}^*)-\mu\boldsymbol{X}\boldsymbol{\theta}^*],b_n^2\lambda_\tau\partial f(\boldsymbol{x}))&\le\sigma_{up}(\sqrt{\mathcal{G}(\lambda)}+\sqrt{2}+t)\\
&\le(b_nM+\eta b_n\sigma)(\sqrt{\mathcal{G}(\lambda)}+\sqrt{2}+2t). \qquad (\text{B.11})
\end{aligned}$$

Here, the last line follows from $b_n\ge\sqrt{\mathcal{G}(\lambda)}+\sqrt{2}+2t$. Merging (B.8) and (B.11) and recalling the definition of the convergence rate $\rho$ the cumulative error takes the form

$$\begin{aligned}
\|\boldsymbol{\theta}_\tau-\mu\boldsymbol{\theta}^*\|_{\ell_2}&\le b_n^{-1}[M(\sqrt{\mathcal{G}(\lambda)}+\sqrt{2}+7t)]+\eta b_n^{-1}[\gamma b_n/\sqrt{n}+\sigma(\sqrt{\mathcal{G}(\lambda)}+\sqrt{2}+3t)]\\
&\le b_n^{-1}M(\sqrt{\mathcal{G}(\lambda)}+\sqrt{2}+7t)+\eta b_n^{-1}[\gamma b_n/\sqrt{n}+\sigma(\sqrt{\mathcal{G}(\lambda)}+\sqrt{2}+7t)]\\
&\le\sqrt{\frac{\boldsymbol{n}_0(\lambda)}{n}}M+\eta\sqrt{\frac{\boldsymbol{n}_0(\lambda)}{n}}\sigma+\eta\gamma/\sqrt{n},
\end{aligned}$$

which is the advertised error bound. ■