Fast, blind, and accurate: Tuning-free sparse regression with global linear convergence

Claudio Mayrink Verdun [*] CLAUDIOVERDUN@SEAS.HARVARD.EDU
Harvard University, Munich Center for Machine Learning

Oleh Melnyk [*] OLEH.MELNYK@TU-BERLIN.DE
TU Berlin, Helmholtz Munich, TU Munich

Felix Krahmer FELIX.KRAHMER@TUM.DE
TU Munich, Munich Center for Machine Learning

Peter Jung PETER.JUNG@TU-BERLIN.DE
TU Berlin, German Aerospace Center (DLR)

Editors: Shipra Agrawal and Aaron Roth

Abstract

Many algorithms for high-dimensional regression problems require the calibration of regularization hyperparameters. This, in turn, often requires the knowledge of the unknown noise variance in order to produce meaningful solutions. Recent works show, however, that there exist certain estimators that are pivotal, i.e., the regularization parameter does not depend on the noise level; the most remarkable example being the square-root lasso. Such estimators have also been shown to exhibit strong connections to distributionally robust optimization. Despite the progress in the design of pivotal estimators, the resulting minimization problem is challenging as both the loss function and the regularization term are non-smooth. To date, the design of fast, robust, and scalable algorithms with strong convergence rate guarantees is still an open problem. This work addresses this problem by showing that an iteratively reweighted least squares (IRLS) algorithm exhibits global linear convergence under the weakest assumption available in the literature. We expect our findings will also have implications for multi-task learning and distributionally robust optimization.

Keywords: Sparse Regression, Square-root LASSO, Iteratively Reweighted Least Squares, Linear Convergence Rate, Majorization-Minimization, Global Convergence, Convex Optimization

1. Introduction

High-dimensional regression problems are ubiquitous in machine learning, statistics, and signal processing. In such scenarios, the number of attributes or features present in each data point exceeds the number of samples. This is the case in many applications, including computer vision, econometrics, or genomics. A large amount of work in the last decades has been devoted to understanding which solutions to such problems are meaningful and how to find them. A common strategy is to design regularizers that bias the variable selection process towards simpler models, limiting the number of active variables within a larger family of predefined admissible features. This approach became known as sparse regression. Enforcing such a parsimonious principle in the regression model reduces overfitting, improves generalization, and enhances the interpretability of the results Hastie et al. (2015). In particular, it is possible to identify which few predictors best

[*] Equal contribution.

explain a certain phenomenon Foucart and Rauhut (2013); Wright and Ma (2022). Mathematically speaking, we consider the linear model given by

$$y_i = \langle x_i, \beta \rangle + e_i, \tag{1}$$

where $X \in \mathbb{R}^{n \times p}$ is the design matrix, $y \in \mathbb{R}^n$ is the data, $e \in \mathbb{R}^n$ is mean-zero random noise and $\beta \in \mathbb{R}^p$ is an unknown vector of coefficients. We consider the case that $p \gg n$, so there are many coefficient vectors that fit this model. Driven by the guiding intuition of enhancing simplicity, one often prefers solutions with a small number of non-zero coefficients, which is closely tied to a small ℓ_1-norm, as it can be interpreted as a convex relaxation of the support size.

Arguably, the most studied method for regression tasks with such an objective is the LASSO Chen and Donoho (1995); Tibshirani (1996). It is an estimator that achieves simultaneous variable selection and estimation by penalizing large ℓ_1-norm among the estimated regression coefficients given as a solution to the following optimization problem

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^p} ||X\beta - y||_2^2 + \lambda ||\beta||_1, \tag{2}$$

where, without loss of generality, the design matrix X is normalized by $1/\sqrt{n}$.

A lot of work has been done over the last years to understand this type of estimator from the computational and statistical point of view. It was shown that this estimator attains optimal minimax rates for the prediction error (Bickel et al., 2009, Section 6) as well as for the reconstruction error in the ℓ_∞-norm, e.g., Lounici (2008); Bellec and Zhang (2022). See also Bunea et al. (2007); Koltchinskii (2009); Ye and Zhang (2010); Raskutti et al. (2011); Dalalyan et al. (2017). The support size of the LASSO minimizer was studied by Foucart et al. (2022), and the consistency of the LASSO in terms of variable selection was established in Zhao and Yu (2006) and Wainwright (2009). Moreover, scalable algorithms were proposed to minimize this objective, e.g., Li et al. (2018); Kümmerle et al. (2021) and debiased versions of the LASSO or its unrolled version were proposed and analyzed in Javanmard and Montanari (2014, 2018); van de Geer et al. (2014); Hoppe et al. (2022, 2023); Bellec and Zhang (2022); Bellec and Tan (2024).

However, such optimal results depend on a regularization choice that relies on oracle knowledge about the noise variance, which is usually not available and hard to estimate in many applications (Giraud, 2015, Chapter 5). Estimating the error variance for LASSO-type problems is a non-trivial problem that still attracts significant interest. See Giraud et al. (2012); Reid et al. (2016); Yu and Bien (2019). The suboptimally tuned LASSO, however, can yield suboptimal recovery guarantees. Besides that, the LASSO estimator lacks some important properties such as scale invariance, see, e.g., (Giraud, 2015, Section 5.1), or asymptotic normality, see, e.g., (Javanmard and Montanari, 2018, Section 1) and references therein for a discussion.

To overcome some of the aforementioned issues, the seminal paper by Belloni et al. (2011) proposed the *square-root LASSO*[1]. This new estimator, in the authors' words *"handles the unknown scale, heteroscedasticity, and (drastic) non-Gaussianity of the noise"*. Its main feature is that the tuning parameter λ that leads to minimax oracle inequalities is independent of the noise level. It is mathematically described by

$$\hat{\beta}_\lambda \in \arg\min_{\beta \in \mathbb{R}^p} ||X\beta - y||_2 + \lambda ||\beta||_1, \tag{3}$$

[1] Also called ℓ_2-lasso in the signal processing literature Oymak et al. (2013).

which can be computed in polynomial time via second-order conic programming. The square-root LASSO initiated a line of research on estimators for sparse regression that exhibit noise-blindness under comparable assumptions, so-called pivotal estimators – see Section 3 for a detailed review.

Besides its statistical properties, the importance of this estimator lies in its interpretation as a distributionally robust optimization (DRO) problem, which, roughly speaking, deals with the problem of finding a regression vector that minimizes the worst-case loss over an uncertainty set (see, e.g., Xu et al. (2010); Bertsimas and Copenhaver (2018); Blanchet and Kang (2017); Olea et al. (2022); Blanchet et al. (2024)). Such a problem has connections with optimal transport (Chu et al., 2022, Section 2).

Although the square-root LASSO and its generalizations exhibit theoretical characteristics comparable to the original LASSO without requiring knowledge of the noise level and also inherit many additional interesting theoretical properties, to our knowledge, none of them admits algorithmic solutions with a provable global linear convergence rate as it has recently been established by Kümmerle et al. (2021) for the original LASSO and its constrained variant, Basis Pursuit. The reason why a computational solution to the square-root LASSO poses a greater challenge is the *non-differentiability* of the loss term that is coupled with the non-smooth terms appearing in the ℓ_1 penalty. This coupling poses a challenge when designing optimization algorithms. While several scalable alternative strategies have been proposed for noise-blind sparse regression - see Section 3 for a discussion - **for essentially all of them, only empirical or, at best, local convergence guarantees are available**. In addition, these local guarantees usually rely on assumptions such as restricted strong convexity and smoothness that are more restrictive than what is required in oracle inequalities for the square-root LASSO; see, e.g., van de Geer (2016). Lastly, many of these algorithms, while noise blind, still require some parameter tuning, such as a suitable initialization or a smoothing strategy that is not universal.

Our contribution: We devise an Iteratively Reweighted Least Squares (IRLS) algorithm, which minimizes non-smooth functions by solving several least squares problems in an iterative way that **provably solve the problem of (non-differentiable) noise-blind sparse regression with a global linear rate**. Our proof requires only the *compatibility condition*, which is the most general condition to analyze this type of estimator; see, e.g., van de Geer and Bühlmann (2009); Stucky and van de Geer (2017). This means that **the devised algorithm exhibits linear convergence while at the same time inheriting the noise-blindness and the oracle inequalities established for the square-root LASSO. Furthermore, it comes with a universal smoothing strategy and does not require any parameter tuning.** Only the desired sparsity level is required as input. An informal version of our main theorem reads as follows.

Theorem 1 *Consider the linear model $y = X\beta_* + e$ and assume that the design matrix X satisfies the compatibility condition. Let β_0^* be the minimizer of Equation (3). There exists a sequence $\{\varepsilon_k\}$, such that the sequence β^k obtained by minimizing a smoothed version $f_{\varepsilon_k}(\beta)$ of the square-root LASSO objective satisfies the following performance guarantee*

1. *Initially, when the objective is still far from its minimum, as quantified by the condition $k \leqslant \hat{k} := \min\left\{k \in \mathbb{R}^p : f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) > 3\lambda(p+1)\varepsilon_k/4\right\}$, the gap of the objective is decreasing at a linear rate*

$$f_{\varepsilon_{k+1}}(\beta^{k+1}) - f_0(\beta_0^*) \leqslant [1 - C_1]\left[f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)\right].$$

2. *When the objective is close to its minimum, $0 \leqslant f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) \leqslant 3\lambda(p+1)\varepsilon_k/4$, the parameter vectors are also close to the true parameter value*

$$\|\beta^k - \beta_*\|_1 \leqslant C_2 \left[\lambda \sigma_s(\beta_*)_{\ell_1} + \|e\|_2\right],$$

where $C_1, C_2 > 0$ are absolute constants, $\sigma_s(\beta)_{\ell_1}$ denotes the ℓ_1-error of the best s-term approximation *of a vector $\beta \in \mathbb{R}^p$, i.e., $\sigma_s(\beta)_{\ell_1} = \inf\{\|\beta - z\|_1 : z \in \mathbb{R}^p$ is s-sparse$\}$.*

We also obtain a sublinear convergence rate when no assumption is imposed on the design matrix. A numerical comparison of the proposed algorithm with other state-of-the-art solutions confirms our findings.

2. Minimizing the square-root LASSO objective with IRLS

In this paper, we devise a majorization-minimization strategy for the non-smooth objective function in (3), i.e., we introduce a smoothed objective $f_\varepsilon(\beta)$ that mitigates the non-smoothness of the $\|\cdot\|_1$-norm as well as the non-smoothness of the $\|\cdot\|_2$-norm but that, at the same time, majorizes the sqrt-LASSO objective function to be minimized. After that, we establish quadratic upper bounds for $f_\varepsilon(\beta)$ that can be optimized efficiently using iterative methods for least squares. Note that the objective function (3) is non-smooth both at the points where the data fidelity term vanishes and wherever an argument entry takes the value zero. Thus, smoothing the objective around these points will yield a smooth function. As Beck and Teboulle (2012), we use a scaled Huber loss function represented by j_γ with parameter $\gamma > 0$ as a proxy to the vanishing terms (which is related to the η-trick, e.g., (Bach et al., 2012, Chapter 5)). This function is defined as

$$j_\gamma(x) := \begin{cases} |x|, & \text{if } |x| \geqslant \gamma, \\ \frac{1}{2}\left(\frac{x^2}{\gamma} + \gamma\right), & \text{if } |x| < \gamma. \end{cases}.$$

The smooth approximation function $j_\gamma(x)$ is continuously differentiable with the Lipschitz gradient with constant $1/\gamma$. As noted by Beck and Sabach (2015), this function is a translation of the Huber function $H_\gamma(x)$, which, in turn, is the Moreau envelope of the absolute value function $|x|$. Motivated by Nesterov (2005), Beck and Teboulle (2012) employed this smoothing to establish accelerated minimization schemes for certain max-type objectives, improving from a rate of $\mathcal{O}(1/k)$ to $\mathcal{O}(1/k^2)$. This smoothing was leveraged by Kümmerle et al. (2021) to solve the basis pursuit problem, a constrained variant of the LASSO, with a linear convergence rate, using an IRLS approach. The key idea is to approximate the objective by the quadratic majorant given by

$$Q_\varepsilon(z,\beta) = f_\varepsilon(\beta) + \langle \nabla f_\varepsilon(\beta), z - \beta\rangle + \frac{1}{2}\langle z - \beta, W_\varepsilon(\beta)(z-\beta)\rangle, \tag{4}$$

where f_ε is the smoothed objective, and W_ε is a weight matrix tailored to its curvature. Exploiting the fact that away from the points of non-smoothness, $\nabla f_\varepsilon(\beta)$ is a linear function, this majorant can be transformed in a pure quadratic function whose minimization properties can then be exploited.

For the square-root LASSO, we work with smoothed objectives of the form

$$f_\varepsilon(\beta) = j_\xi(\|X\beta - y\|_2) + \lambda \sum_{i=1}^{p} j_\delta(\beta_i), \quad \varepsilon = (\xi, \delta), \quad \xi, \delta \geqslant 0 \tag{5}$$

where due to the different nature of the terms, we allow for two different smoothing parameters ξ and δ. That is, for $\xi, \delta > 0$ the gradient is given by

$$\nabla f_\varepsilon(\beta) = \frac{X^T(X\beta - y)}{\max\{\|X\beta - y\|_2, \xi\}} + \lambda \sum_{j=1}^{p} \frac{\beta_j e_j}{\max\{|\beta_j|, \delta\}},$$

which is affine but no longer linear away from the points of non-smoothness. Consequently, one does not obtain a pure quadratic function, and the IRLS strategy by Kümmerle et al. (2021) is not directly applicable. To overcome this obstacle, we introduce the change of variables

$$X\beta - y = \begin{bmatrix} -y & X \end{bmatrix} \begin{bmatrix} 1 \\ \beta \end{bmatrix} = \tilde{X}\tilde{\beta}.$$

The embedding $\tilde{v} := (1, v^T)^T$ for vectors $v \in \mathbb{R}^p$ will be used throughout this paper with the added dimension indexed by zero, i.e., $\tilde{v}_0 = 1$. With this change of variables, our objective function becomes

$$f_\varepsilon(\beta) = j_\xi(\|X\beta - y\|_2) + \lambda \sum_{i=1}^{p} j_\delta(\beta_i) = j_\xi(\|\tilde{X}\tilde{\beta}\|_2) + \lambda \sum_{i=1}^{p} j_\delta(\beta_i) =: \tilde{f}_\varepsilon(\tilde{\beta})$$

and the unconstrained optimization of the proposed smoothed objective becomes

$$\min f_\varepsilon(\beta) = \min_{\tilde{\beta} \in \mathbb{R}^{p+1},\ \tilde{\beta}_0 = 1} \tilde{f}_\varepsilon(\tilde{\beta}). \tag{6}$$

The gradient of $\tilde{f}_\varepsilon(\tilde{\beta})$ is now linear away from the points of non-smoothness so that an IRLS strategy can be devised with a majorant of the form

$$\begin{aligned} Q_\varepsilon(\tilde{z}, \tilde{\beta}) :&= \tilde{f}_\varepsilon(\tilde{\beta}) + \langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}), \tilde{z} - \tilde{\beta} \rangle + \frac{1}{2}\langle \tilde{z} - \tilde{\beta}, W_\varepsilon(\tilde{\beta})(\tilde{z} - \tilde{\beta}) \rangle \\ &= \tilde{f}_\varepsilon(\tilde{\beta}) + \frac{1}{2}\langle \tilde{z}, W_\varepsilon(\tilde{\beta})\tilde{z} \rangle - \frac{1}{2}\langle \tilde{\beta}, W_\varepsilon(\tilde{\beta})\tilde{\beta} \rangle. \end{aligned} \tag{7}$$

where the positive semidefinite weight matrix W_ε is given by

$$W_\varepsilon(\tilde{\beta}) = \frac{\tilde{X}^T\tilde{X}}{\max\{\|\tilde{X}\tilde{\beta}\|_2, \xi\}} + \lambda \begin{bmatrix} 0 & 0 \\ 0 & \operatorname{diag}(\{\max^{-1}\{|\tilde{\beta}_i|, \delta\}\}_{i=1,\dots,p}) \end{bmatrix}. \tag{8}$$

In line with the IRLS strategy, this choice of W_ε ensures that $Q_\varepsilon(\tilde{z}, \tilde{\beta})$ is a majorizer via the following lemma, whose proof can be found in Appendix A.

Lemma 2 *The function in Equation (7) with W_ε as in (8) admits*

i. $W_\varepsilon(\tilde{\beta})\tilde{\beta} = \nabla \tilde{f}_\varepsilon(\tilde{\beta})$, *ii.* $Q_\varepsilon(\tilde{\beta}, \tilde{\beta}) = \tilde{f}_\varepsilon(\tilde{\beta})$, *iii.* $Q_\varepsilon(\tilde{z}, \tilde{\beta}) \geqslant \tilde{f}_\varepsilon(\tilde{z})$.

The minimizer of $\tilde{f}_\varepsilon$ can now be approximated by iteratively solving the least squares problem

$$\tilde{\beta}^{k+1} := \operatorname*{arg\,min}_{\tilde{z} \in \mathbb{R}^{p+1},\ \tilde{z}_0 = 1} Q_\varepsilon(\tilde{z}, \tilde{\beta}^k). \tag{9}$$

Indeed, the objective function is decreasing:

$$0 \leqslant \tilde{f}_{\varepsilon}(\tilde{\beta}^{k+1}) \leqslant Q_{\varepsilon}(\tilde{\beta}^{k+1}, \tilde{\beta}^{k}) \leqslant Q_{\varepsilon}(\tilde{\beta}^{k}, \tilde{\beta}^{k}) = \tilde{f}_{\varepsilon}(\tilde{\beta}^{k}).$$

As the value $\tilde{\beta}_0^k = 1$ is enforced by the constraint, we can then invert the change of variables to obtain an approximating sequence β^k to the minimizer of f_ε. We note that some previous works considered local majorization-minimization strategies, i.e., majorization only in the neighborhood of the current iterate β^k, see Chouzenoux et al. (2023). However, it is challenging to establish a convergence rate theory for such methods.

As discussed in the context of matrix completion, e.g., by Kümmerle and Mayrink Verdun (2020); Kümmerle and Mayrink Verdun (2021), such a method can be seen as a variable metric proximal gradient descent. There are some general convergence results available for such approaches Park et al. (2020); Tran-Dinh et al. (2015), but mainly for the class of self-concordant functions, which does not include the Huber loss. Despite (9) being a constrained optimization, its translation back to the original coordinates can be seen as a classical least squares problem, as the following lemma shows.

Lemma 3 *The $k+1$-st iterate β^{k+1} of the approximating sequence arising from (9) is the minimizer of the unconstrained least squares problem*

$$\min_{z \in \mathbb{R}^p} \frac{\|Xz - y\|_2^2}{\max\{\|X\beta^k - y\|_2, \xi\}} + \lambda \sum_{j=1}^{p} \frac{|z_j|^2}{\max\{|\beta_j^k|, \delta\}}.$$

Iterating these least squares problems will give rise to a solution to the smoothed objective. The convergence speed can be characterized under very general conditions on the design matrix; see Theorem 6 below.

In general, however, the smoothed objective will approach the non-smooth objective only for $\xi, \delta \to 0$. Hence, to minimize the non-smooth objective, one needs to update the smoothing parameters ξ, δ according to an appropriate decay rule, which constitutes Algorithm 1. More details can be found in Appendix E.

Algorithm 1 Quadratic minimization for square-root LASSO

```
Input: Design matrix $X \in \mathbb{R}^{n \times p}$, data vector $y \in \mathbb{R}^n$, initial $\varepsilon_0 = (\xi_0, \delta_0)$.
for $k = 0, 1, 2, \ldots$ do
    $\beta^{k+1} := \arg\min_{z \in \mathbb{R}^p} \frac{\|Xz - y\|_2^2}{\max\{\|X\beta^k - y\|_2, \xi_k\}} + \lambda \sum_{j=1}^{p} \frac{|z_j|^2}{\max\{|\beta_j^k|, \delta_k\}}.$    (10)
    Update $\xi_{k+1}, \delta_{k+1}$ such that $0 < \xi_{k+1} \leqslant \xi_k$ and $0 < \delta_{k+1} \leqslant \delta_k$    (11)
end for
return Sequence $(\beta^k)_{k \geqslant 1}$.
```

Note that due to $f_{\varepsilon_1}(\beta) \leqslant f_{\varepsilon_2}(\beta)$ whenever $0 \leqslant \xi_1 \leqslant \xi_2$ and $0 \leqslant \delta_1 \leqslant \delta_2$, the monotonicity argument above is still applicable, that is, $0 \leqslant f_{\varepsilon_{k+1}}(\beta^{k+1}) \leqslant f_{\varepsilon_k}(\beta^{k+1}) \leqslant f_{\varepsilon_k}(\beta^k)$. While this guarantees that Algorithm 1 eventually converges to a fixed value of the loss function, from this qualitative argument, it is neither possible to infer the properties of the limit point nor quantify the convergence speed. Thus, a more involved analysis is needed. In this paper, we establish that a suitably chosen decay rate allows for (i) a $\mathcal{O}(1/k)$ rate for very general design matrices and, *our main result*, (ii) global linear convergence under the assumption that the design matrix X satisfies the compatibility condition.

3. Related works

Pivotal estimators. Before the proposal of the square-root LASSO (3), there were other attempts to design scale-invariant and pivotal estimators for the sparse regression problem. For example, the work Städler et al. (2010) proposed an estimator that is scaling invariant and simultaneously estimates the noise level to the sparse vector. However, this estimator called *Scaled LASSO*, still relies on parameter tuning, and no algorithmic solutions were discussed. Inspired by robust regression techniques (Huber, 1981, Chapter 7), Antoniadis (2010) proposed to minimize the scaling invariant and jointly convex objective function

$$(\hat{\beta}_\lambda, \hat{\sigma}_\lambda) \in \underset{\beta\in\mathbb{R}^p, \sigma>0}{\arg\min} \frac{1}{2\sigma}\|y - X\beta\|_2^2 + \frac{\sigma}{2} + \lambda\|\beta\|_1, \tag{12}$$

for simultaneous sparse regression and noise estimation. The paper suggested that using the estimator (12) instead of the original Scaled LASSO proposed by Städler et al. (2010) could admit a more efficient solution algorithm. In his own words *“Do the authors think that such a parametrization could lead to a more efficient optimization algorithm?”*. Later, Sun and Zhang (2012) referred to the estimator (12), as the Scaled LASSO, which led to some confusion in the literature as highlighted in (van de Geer, 2016, Section 3.1). See also (Owen, 2007, Equation 8). Subsequent works have referred to it as the Concomitant LASSO by Ndiaye et al. (2017) (in the following, we will also use this name) or as SPICE (SParse Iterative Covariance-based Estimation) by Babu and Stoica (2014). As noted in van de Geer (2016), the Concomitant LASSO is equivalent to the square-root LASSO.

Theory for the square-root LASSO. The most celebrated result for the square-root LASSO estimator is in the form of noise-blind sharp recovery guarantees. Namely, it is possible to establish that the square-root LASSO estimator attains optimal error for a certain choice of regularization parameter λ that is *independent* from the noise level. For example, for Gaussian noise, it was proven in (Belloni et al., 2011, Theorem 1), (Derumigny, 2018, Theorem 3.1) and (Stucky and van de Geer, 2017, Corollary 13) that one can choose λ independently of σ, and obtain a solution of the square-root LASSO β^λ with an error of the order of $s/n \log(p/s)$, which is known to be sharp (see Bellec et al. (2018) and Raskutti et al. (2011)). In particular, Stucky and van de Geer (2017) established this result for design matrices satisfying the *compatibility condition*, the weakest possible assumption for establishing oracle inequalities. Later, this was generalized to adversarial (worst case) noise for design matrices satisfying again the compatibility condition or, equivalently, the so-called *robust null space property* (for a discussion of the equivalence, see Petersen and Jung (2021)). In this work, we will also focus on this framework of adversarial noise. Under slightly stronger assumptions on the design matrix, also bounds on the support size for the solutions of Equation (3) are available Foucart (2023). Recently, Berk et al. (2023) studied the well-posedness and parameter sensitivity of Equation (3).

Generalizations. Since its introduction, the square-root LASSO estimator has been extended to encompass several variants of the sparse regression problem such as group sparse regression Bunea et al. (2014), multivariate response linear regression Liu et al. (2015); Molstad (2022), square-root fused LASSO Jiang et al. (2021), matrix completion Klopp (2014), square-root sorted ℓ_1 penalized estimation (SLOPE) Stucky and van de Geer (2017); Minsker et al. (2024), square-root Principal Component Pursuit Zhang et al. (2021), and more broadly, any regression problem regularized by a norm that fulfills the weak decomposability condition Stucky and van de Geer (2017).

Minimization algorithms and convergence rates. A study by Belloni et al. (2011) explores minimizing (3) using the SDT3 implementation of an interior-point method by Toh et al. (2012) or the first-order method for conic programs TFOCS by Becker et al. (2011). Later, an ADMM-based solver was proposed by Li et al. (2015). To mitigate the costly sub-steps of ADMM, a Primal-Dual Hybrid Gradient with a $O(1/k)$ convergence rate was developed by Goldstein et al. (2015). The equivalent Concomitant LASSO was addressed using a combination of gradient descent and alternating minimization by Sun and Zhang (2012), and Ndiaye et al. (2017) proposed a coordinate descent strategy, coupled with smoothing and a pathwise optimization for the tuning parameter to enhance the empirical speed of convergence. Poon and Peyré (2023) proposed an algorithm based on an overparametrized variational formulation that applies to (3), but did not establish a convergence rate. Li et al. (2020) proposed proximal gradient descent and proximal Newton methods for the square-root LASSO, admitting local linear and local quadratic convergence guarantees, respectively, but required the strong assumption of a locally restricted strongly smooth condition for their validity. More recently, Tang et al. (2020) presented a semismooth Newton-based method for a class of minimization problems, including (3), with local superlinear convergence. There were other attempts to define alternative pivotal estimators in high-dimensional problems specifically designed for speed and scalability, e.g., the self-normalized conic estimator by Belloni et al. (2017) and the self-tuned Dantzig estimator by Gautier and Tsybakov (2013). The latter admits a linear programming formulation, which is solvable in polynomial time, but we are not aware of a thorough numerical evaluation of these methods, and their convergence rates remain elusive.

IRLS algorithms. The method proposed in this paper is an *Iteratively Reweighted Least Squares* (IRLS) algorithm. IRLS algorithms are an active area of research and have been successfully applied for many problems beyond high-dimensional regression, such as in subspace prototype learning Mankovich et al. (2022) and point cloud alignment problems Aftab and Hartley (2015), manifold-valued image restoration Bergmann et al. (2016), system identification Brouillon et al. (2022), joint learning of neural networks Zhang et al. (2019), numerical methods for elliptic PDEs Diening et al. (2020), design of FIR filters Burrus et al. (1994), time-harmonic motion tracking Melnyk et al. (2024), learning sparse and low-rank priors for image problems Lefkimmiatis and Koshelev (2023) and the recovery of low-rank matrices Mohan and Fazel (2012); Fornasier et al. (2011); Kümmerle and Mayrink Verdun (2021). The work Ba et al. (2013) established a correspondence between certain IRLS-type algorithms and a class of Expectation-Maximization algorithms. The IRLS strategy can be traced back to the solution of the *Fermat-Weber problem*, i.e., the problem of finding the geometric median of a discrete set of points in a Euclidean space, proposed by Weiszfeld (1937) in the 1930s, see also Beck and Sabach (2015). Great advantages of the IRLS strategy include that it is tuning-free, does not require a sophisticated initialization, and relies on efficient and simple linear algebra. In each iteration, one only needs to solve a linear system arising from a quadratic problem. Key to its performance is a well-designed sequence of smoothing parameters adapted to the geometry of the problem (but not the data). At the same time, however, the iterative nature with varying degrees of smoothing can make the analysis more complicated, which is why IRLS and related problems have been of continuous interest in the optimization community.

4. Global convergence with fixed smoothing parameter

In this section, we study the case that one does not update ε as in Equation (11), but it is fixed to be a very small positive number. The analysis of this case does not need any structural assumption on the design matrix X, and its proofs only require tools from convex analysis. Our first result establishes global sublinear convergence of IRLS for the smoothed problem.

Theorem 4 *Let $\xi, \delta > 0$ and $k \geqslant 2$ and denote by β^*_ε the minimizer of f_ε. Then the iterates β^k of Algorithm 1 satisfy the inequality*

$$f_\varepsilon(\beta^k) - f_\varepsilon(\beta^*_\varepsilon) \leqslant \max\left\{2^{-\frac{k-1}{2}}(f_\varepsilon(\beta^0) - f_\varepsilon(\beta^*_\varepsilon)), \frac{8(\xi^{-1}\|X\|^2 + \lambda\delta^{-1})(f_\varepsilon(\beta^0) + f_0(\beta^*_\varepsilon))^2}{\lambda^2(k-1)}\right\}.$$

The rate presented in Theorem 4 can be divided into two distinct components. While the first term indicates a geometric decay for large objective gaps, it is unclear when the sublinear term starts to dominate. The following theorem clarifies this issue, establishing linear convergence until the objective gap falls below a fixed threshold.

Theorem 5 *Let $\xi, \delta > 0$ and denote by β^*_ε the minimizer of f_ε. If $f_\varepsilon(\beta^k) - f_\varepsilon(\beta^*_\varepsilon) \geqslant \gamma$ for some constant $\gamma > 0$, then the iterates β^k of Algorithm 1 satisfy the inequality*

$$f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta^*_\varepsilon) \leqslant \Big(f_\varepsilon(\beta^k) - f_\varepsilon(\beta^*_\varepsilon)\Big)\left[1 - \frac{\gamma}{2\max\{\frac{1}{\xi}, \frac{1}{\lambda\delta}\}f_0^2(\beta^*_\varepsilon) + 4\gamma}\right].$$

The necessity to step away from the minimum by γ can be observed in the linear rate in Theorem 5, which deteriorates if γ is chosen smaller. We also note the similar dependency on the smoothing parameters ξ and δ, which is why this result does not translate exactly to IRLS with decaying smoothing parameters discussed in the next section.

5. Global convergence with decay of the smoothing parameter

This preceding analysis assumed a fixed regularization parameter and derived convergence results towards the solution β^*_ε of the regularized problem. However, the true potential of IRLS-type algorithms lies in constructing a sequence of objective functions with decaying regularization parameters $\varepsilon_k = (\xi_k, \delta_k)$, as captured in Algorithm 1, allowing the iterates to converge towards the solution of the original non-smooth function (3).

In analogy to constant ε, we again establish global sublinear convergence of $f_{\varepsilon_k}(\beta^k)$ to $f_0(\beta^*_0)$. However, the most general version of this statement is more technical, which is why we refer it to Appendix C, see Theorem 19. Its essence is that by using appropriately tuned parameters, one can achieve a convergence rate of $k^{-1/2}$. Recent lower bounds by Chizat (2022) for a related smooth problem class suggest that this rate may be optimal in the general framework that we are considering in this section. Instead, we state a tuning-free variant of this theorem, also proved in Appendix C, which still achieves a rate of $k^{-1/3}$.

Theorem 6 *Consider the sequences $\xi_k = \lambda\delta_k$ and*

$$\delta_k = \frac{2\min_{s=0,\dots,k} f_0(\beta^s)}{\lambda\sqrt{p+1}\sqrt{k+1}}, \quad k \geqslant 0.$$

For $k \geqslant 1$, the sequence β^k generated by Algorithm 1 admits

$$f_{\varepsilon_{k+1}}(\beta^{k+1}) - f_0(\beta_0^*) \leqslant \max\left\{4\sqrt{p+1} f_{\varepsilon_0}(\beta^0), 10e(f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*))\right\} k^{-\frac{1}{3}}.$$

6. Global convergence under the compatibility condition

The square-root LASSO is mainly used to retrieve sparse regressors in a noise-blind way. It is, therefore, standard in the literature to assume that the design matrix $X \in \mathbb{R}^{n\times p}$ is, in some sense, well-conditioned on the set of sparse vectors or that the kernel of such matrices has a benign geometry. The common concept to capture this is the compatibility condition, which is also the sharpest condition to obtain oracle inequalities for estimation and prediction, see (van de Geer and Bühlmann, 2009, Figure 1).

Definition 7 *A matrix $X \in \mathbb{R}^{n\times p}$ is said to satisfy the* (L,S)-compatibility condition *if there exists $L \in (1,\infty)$ such that for the set $\Delta_{L,S} := \left\{v \in \mathbb{R}^N : \|v_{S^c}\|_1 \leqslant L\|v_S\|_1 \text{ and } \|v_S\|_1 \neq 0\right\}$ the condition $\inf_{v\in\Delta_{L,S}} \frac{|S|\|Xv\|_2^2}{\|v_S\|_1^2} > 0$ holds true.*

Note that many random designs have this property, even for heavy-tailed distributions (Dirksen et al., 2018, Theorem 5.1). An equivalent formulation of this condition, more commonly used in adversarial noise models (see (Petersen and Jung, 2021, Proposition 6.1) for the proof of equivalence), is the robust null space property (NSP).

Definition 8 *(Foucart and Rauhut, 2013, Definition 4.17) A matrix $X \in \mathbb{R}^{n\times p}$ is said to satisfy the* robust null space property (NSP) *of order $s \in [p]$ with constants $0 < \rho < 1$ and $\tau > 0$ if for any set $S \subset [p]$ of cardinality $|S| \leqslant s$, it holds that*

$$\|v_S\|_1 \leqslant \rho\|v_{S^c}\|_1 + \tau\|Xv\|_2, \text{ for all } v \in \mathbb{R}^p. \tag{13}$$

In particular, this property implies that sparse recovery via Basis Pursuit, the constrained equivalent of the LASSO, is robust with respect to adversarial perturbations (Foucart and Rauhut, 2013, Chapter 4); in the noiseless case, it has even been shown to be a necessary condition for the success of Basis pursuit. See also (Petersen and Jung, 2021, Section 3.3) for an extensive discussion.

Our main theorem, proved in Appendix D, establishes the global linear convergence rate of IRLS for (3) under NSP:

Theorem 9 *Consider the linear system $y = X\beta_* + e$, where matrix $X \in \mathbb{R}^{n\times p}$ satisfies the null space property with constants $0 < \rho \leqslant \frac{1}{6}$ and $0 < \tau \leqslant \frac{7}{6}$. Let $\lambda \leqslant 1/7$ and let $\delta_k = \min\left\{\delta_{k-1}, \frac{\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}}{\lambda(p+1)}\right\}$ and $\xi_k = \lambda\delta_k$ for $k \geqslant 0$ and $\delta_{-1} = +\infty$ be the sequence of the smoothing parameters. Then, for*

$$k \leqslant \hat{k} := \min\left\{k \in \mathbb{N} : f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) > 3\lambda(p+1)\delta_k/4\right\}$$

it holds that the iterates β^k of Algorithm 1 admit

$$f_{\varepsilon_{k+1}}(\beta^{k+1}) - f_0(\beta_0^*) \leqslant \left[1 - \frac{1}{1250(p+1)}\right]\left[f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)\right].$$

Moreover, for $0 \leqslant f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) \leqslant 3\lambda(p+1)\delta_k/4$, we have

$$\|\beta^k - \beta_*\|_1 \leqslant 18\left(\sigma_s(\beta_*)_{\ell_1} + \lambda^{-1}\|e\|_2\right).$$

We consider it a crucial feature of our result that it does not impose a parameter choice but works for all values of $\lambda \leqslant \frac{1}{7}$. The reason is that for a random noise model, one may also consider smaller values of λ: it has been shown that for the square-root LASSO, there is a trade-off between the parameter λ and the success probability. To achieve a success probability of $1-\alpha$, one should choose $\lambda = \sqrt{\frac{2\log(2p/\alpha)}{n-1}}$, see (van de Geer, 2016, Lemma 8.2). This is in line with our worst-case bound above. If one uses a covering argument over the set of admissible error vectors – as the covering size is exponential in n, one needs an exponentially small α, which yields a constant λ.

7. Numerical results

In this section, we numerically investigate the convergence properties of our method and compare it with other state-of-the-art algorithms for solving the square-root LASSO. First, we examine the convergence rate for multiple decay rates of the smoothing parameter ε_k. Then, we compare our approach against different state-of-the-art algorithms (reviewed in Appendix F below) for minimizing the square-root LASSO objective function in terms of accuracy and running time. The design matrix $X \in \mathbb{R}^{200\times5000}$ is a standard Gaussian random matrix normalized by $1/\sqrt{n}$. We generate data y that admits a linear model $y = X\beta_* + e$ for some sparse parameter vector β_* with $s = 20$ and e being Gaussian noise. The final version of the code is available in a Github repository[2], and for further implementation details on IRLS, we refer the reader to Appendix E.

7.1. Comparing decay rates of the smoothing parameter

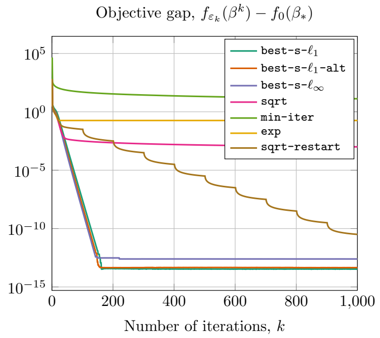
In our first set of experiments, we consider the noiseless scenario and put Theorem 9 into perspective by investigating the convergence to the true parameter vector β_* for $\lambda = \frac{1}{7}$ and different decay strategies for the parameters $\varepsilon_k = (\xi_k, \delta_k)$. We compare seven different rules to decrease δ_k and always choose $\xi_k = \lambda\delta_k$ in line with Theorem 9. For better readability, we only provide a qualitative description and refer to the table in Figure 1 for the precise formulations of the rules.

The first choice is the theoretical decay rate that leads to our linear convergence rate results, Theorem 9 based on the best-s-term approximation error in ℓ_1, here denoted by `best-s-`ℓ_1. As it does not impact the validity of Theorem 9 and only affects constants in it as discussed in Remark 28, we also explore modifying the constant by a multiplicative factor, rule `best-s-`ℓ_1`-alt`. Inspired by Daubechies et al. (2010), we also include the corresponding rule with the best-s-term approximation error in ℓ_∞, denoted `best-s-`ℓ_∞.

The fourth and fifth strategies, `sqrt` and `min-iter`, guarantee sublinear convergence in a more general context, see Appendix C and Theorem 6. The sixth rule suggested by Chartrand and Yin (2008) implementing geometric decay was the first parameter decay strategy for IRLS suggested in the context of sparse recovery. Lastly, we consider a basic restarting scheme `sqrt + restart`, where the algorithm is restarted after every $K = 100$ iteration. As it has been known since the seminal work by Nemirovskii and Nesterov (1985) that restarting strategies can improve the convergence rate, e.g., O'donoghue and Candes (2015); Roulet and d'Aspremont (2017); Renegar and Grimmer (2022), we find it an interesting follow-up question whether or not it theoretically improves the more general sublinear IRLS guarantees.

Each algorithm was executed for 1000 iterations, and the resulting objective function gap $f_\varepsilon(\beta^t) - f_0(\beta^*)$ is depicted in Figure 1. The numerical results confirm that the IRLS version with rules

[2] https://github.com/claudioverdun/sqrt-lasso



Name	Rule	Reference
best-s-ℓ_1	$\delta_k = \min\left\{\delta_{k-1}, \frac{\|\|X\beta^k - y\|\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}}{\lambda(p+1)}\right\}$	Theorem 9
best-s-ℓ_1-alt	$\delta_k = \min\left\{\delta_{k-1}, \frac{1}{2}\frac{\|\|X\beta^k - y\|\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}}{\lambda(p+1)}\right\}$	Remark 28 with $c = \frac{1}{2}$
best-s-ℓ_∞	$\delta_k = \min\left\{\delta_{k-1}, \frac{1}{2}\frac{\|\|X\beta^k - y\|\|_2 + \lambda\sigma_s(\beta^k)_{\ell_\infty}}{\lambda(p+1)}\right\}$	See Daubechies et al. (2010).
sqrt	$\delta_k = \delta_0/\sqrt{1+k}$	Theorem 19 with $\theta = 1/2$
min-iter	$\delta_k = \frac{2\min_{\ell=0,\dots,k-1} f_0(\beta^\ell)}{\lambda(p+1)^{1/2}(1+k)^{1/2}}$	Theorem 6
exp	$\delta_k = \delta_0 2^{-k}$	See Chartrand and Yin (2008).
sqrt-restart	$\delta_{t,k} = \delta_{t,0}/\sqrt{1+k}, \delta_{t,0} = \delta_{t-1,K}$	t = 1,. . . , 10.

Figure 1: Convergence of IRLS with different strategies to decrease δ_k. $n = 5000$, $m = 200$, $s = 20$. The table illustrates the different decays of the smoothing parameter. $\delta_0 = 10^{-4}$, $\delta_{0,K} = 10^{-3}$.

best-s-ℓ_1, best-s-ℓ_1-alt and best-s-ℓ_∞ attain linear convergence. Moreover, the restarted sqrt-restart improves the sublinear convergence of sqrt to a linear rate.

7.2. Comparison with alternative methods

We compare the IRLS algorithm with some of the parameter choice rules described above, namely, sqrt, best-s-ℓ_1-alt, and sqrt-restart, against other state-of-the-art algorithms in the presence of noise. We refer the reader to Appendix F for more details on the alternative methods.

We run the methods both for the parameter $\lambda = \frac{1}{7}$ and the smaller parameter $\lambda = \frac{1}{100}$ (cf. the discussion after Theorem 9) Every point is an average of 30 trials, and the stopping criterion was a runtime of 60 seconds, or a relative step size of $\|\beta^{k+1} - \beta^k\|_2/\|\beta^k\|_2 \leqslant 10^{-5}$. The only exception was ITEM, where we enforced a relative error of 10^{-8}, which works best for smaller step sizes.

We report the relative error of the parameter vector $\|\beta^k - \beta_*\|_2/\|\beta_*\|_2$, the relative prediction error $\|X\beta^k - y\|_2/\|y\|_2$, the runtime, the effective sparsity $\|\beta^k\|_1^2/\|\beta^k\|_2^2$ and the support failure rate $SFR = 1 - |\mathcal{J} \cap \mathcal{S}|/|\mathcal{S}|$, where $\mathcal{S}$ is the support of β_* and $\mathcal{J}$ are the indices of the s entries of β^k largest in magnitude. Figure 2, depicts all these measures as a function of the signal-to-noise ratio $SNR = 10\log_{10}(\|y\|_2^2/\nu^2)$, where $\nu^2 I_n$ is the covariance matrix of the noise.

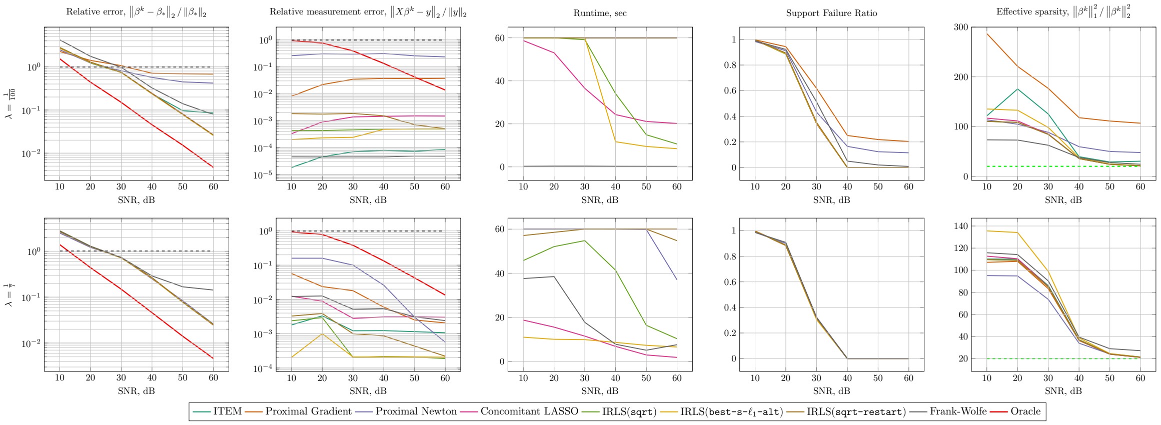


Figure 2: Impact of noise on the reconstruction. $p = 5000$, $s = 20$, $n = 200$.

We observe that IRLS is comparable to the Concomitant LASSO in all metrics and outperforms all other methods in terms of recovery performance. As the gradient methods require more iterations to converge, this may be partly due to the time limit of 60 seconds. Yet all the methods consistently agree in terms of the same support recovery rate. The Frank-Wolfe algorithm, in particular, has a remarkably short runtime. However, it has a worse relative error as the algorithm reaches the relative step size stopping criterion. While decreasing the threshold value improves the accuracy, the runtime deteriorates as the exhibited convergence rate is sublinear. Thus, we see the great potential of Frank-Wolfe as an initialization for other methods. One of the remarkable advantages of IRLS that is not visible in this graph is that we had to tune the parameters for many of the algorithms here presented, including the concomitant LASSO, for each λ individually, while our method (with `best-s-`ℓ_1`-alt` decay strategy) is tuning-free.

8. Conclusion

We presented the first global linear convergence guarantees for a robust and scalable algorithm solving the square-root LASSO problem. Numerical experiments confirmed the linear rate and showed that our method is on par with state-of-the-art methods in terms of accuracy, convergence speed, and noise robustness. Among them, our approach sticks out as the only method with the desired sparsity as the sole variable parameter that does not require any additional parameter tuning. Our theoretical result highlights a linear rate under minimal assumptions.

We believe that our result may serve as a role model for obtaining global fast rates for other minimization problems with coupled non-smooth summands in the objective. Additional interesting topics for follow-up work include the analysis of restarting strategies, as well as the application to multitask problems, distributionally robust optimization, and out-of-sample analysis.

Acknowledgments

C. Mayrink Verdun and Felix Krahmer were supported by the Germany Science Foundation (DFG) in the context of the Emmy Noether junior research group KR 4512/1-1, by the German Federal Ministry of Education and Research in the grant *SparseMRI3D+* (FZK 05M20WOA) and by the Munich Center for Machine Learning. In addition, Claudio Verdun was supported by the DFG within Germany's Excellence Strategy EXC-2111 390814868, and Felix Krahmer was supported by the DFG in the context of the priority program *Theoretical Foundations of Deep Learning* (project KR 4512/6-1). O. Melnyk acknowledges support from the Helmholtz Association under the contracts No. ZT-I-0025 (Ptychography 4.0), No. ZT-I-PF-4-018 (AsoftXm), No. ZT-I-PF-5-28 (EDARTI), No. ZT-I-PF-4-024 (BRLEMMM).

References

Khurrum Aftab and Richard Hartley. Convergence of iteratively re-weighted least squares to robust m-estimators. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 480–487. IEEE, 2015.

Anestis Antoniadis. Comments on: ℓ_1-Penalization for Mixture Regression Models by N. Stadler, P. Bühlmann and S. van de Geer. *TEST*, 19(2):257–258, 2010.

Aleksandr Aravkin, James V Burke, and Daiwei He. IRLS for sparse recovery revisited: Examples of failure and a remedy. *arXiv preprint arXiv:1910.07095*, 2019.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality. *Mathematics of operations research*, 35(2):438–457, 2010.

Demba Ba, Behtash Babadi, Patrick L Purdon, and Emery N Brown. Convergence and stability of iteratively re-weighted least squares algorithms. *IEEE Transactions on Signal Processing*, 62(1): 183–195, 2013.

Prabhu Babu and Petre Stoica. Connection between SPICE and square-root LASSO for sparse parameter estimation. *Signal Processing*, 95:10–14, 2014.

Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

Amir Beck. On the convergence of alternating minimization for convex programming with applications to iteratively reweighted least squares and decomposition schemes. *SIAM J. Optim.*, 25(1): 185–209, 2015.

Amir Beck. *First-order methods in optimization*. SIAM, 2017.

Amir Beck and Shoham Sabach. Weiszfeld's method: Old and new results. *J. Optim. Theory Appl.*, 164:1–40, 2015.

Amir Beck and Shimrit Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1):1–27, 2017.

Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22(2):557–580, 2012.

Stephen Becker, Emmanuel J. Candès, and Michael C. Grant. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput.*, 3(3):165–218, 2011.

Pierre C Bellec and Kai Tan. Uncertainty quantification for iterative algorithms in linear models with application to early stopping. *arXiv preprint arXiv:2404.17856*, 2024.

Pierre C Bellec and Cun-Hui Zhang. De-biasing the lasso with degrees-of-freedom adjustment. *Bernoulli*, 28(2):713–743, 2022.

Pierre C Bellec, Guillaume Lecué, and Alexandre B Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642, 2018.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 12 2011.

Alexandre Belloni, Abhishek Kaul, and Mathieu Rosenbaum. Pivotal estimation via self-normalization for high-dimensional linear models with error in variables. *arXiv preprint arXiv:1708.08353*, 2017.

Ronny Bergmann, Raymond H. Chan, Ralf Hielscher, Johannes Persch, and Gabriele Steidl. Restoration of manifold-valued images by half-quadratic minimization, 2016.

Aaron Berk, Simone Brugiapaglia, and Tim Hoheisel. Square root LASSO: well-posedness, Lipschitz stability and the tuning trade off. *arXiv preprint arXiv:2303.15588*, 2023.

Dimitris Bertsimas and Martin S Copenhaver. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3):931–942, 2018.

Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705 – 1732, 2009.

Jose Blanchet and Yang Kang. Distributionally robust groupwise regularization estimator. In *Asian Conference on Machine Learning*, pages 97–112. PMLR, 2017.

Jose Blanchet, Jiajin Li, Sirui Lin, and Xuhui Zhang. Distributionally robust optimization and robust statistics. *arXiv preprint arXiv:2401.14655*, 2024.

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165:471–507, 2017.

Jérôme Bolte and Edouard Pauwels. Majorization-minimization procedures and convergence of SQP methods for semi-algebraic and tame programs. *Mathematics of Operations Research*, 41 (2):442–465, 2016.

Jean-Sébastien Brouillon, Keith Moffat, Florian Dörfler, and Giancarlo Ferrari-Trecate. Robust online joint state/input/parameter estimation of linear systems. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 2153–2158. IEEE, 2022.

Florentina Bunea, Alexandre Tsybakov, and Marten Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

Florentina Bunea, Johannes Lederer, and Yiyuan She. The group square-root lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inf. Theory*, 60(2):1313–1325, 2014.

Charles Sidney Burrus, Jose Antonio Barreto, and Ivan W Selesnick. Iterative reweighted least-squares design of FIR filters. *IEEE Transactions on Signal Processing*, 42(11):2926–2936, 1994.

Rick Chartrand and Wotao Yin. Iteratively reweighted algorithms for compressive sensing. In *2008 IEEE international conference on acoustics, speech and signal processing*, pages 3869–3872. IEEE, 2008.

Scott Chen and David L Donoho. Examples of basis pursuit. In *Wavelet Applications in Signal and Image Processing III*, volume 2569, pages 564–574. SPIE, 1995.

Farah Cherfaoui, Valentin Emiya, Liva Ralaivola, and Sandrine Anthoine. Recovery and convergence rate of the Frank–Wolfe algorithm for the m-exact-sparse problem. *IEEE Transactions on Information Theory*, 65(11):7407–7414, 2019.

Lénaïc Chizat. Convergence rates of gradient methods for convex optimization in the space of measures. *Open Journal of Mathematical Optimization*, 3:1–19, 2022.

Emilie Chouzenoux, Ségolène Martin, and Jean-Christophe Pesquet. A local MM subspace method for solving constrained variational problems in image recovery. *Journal of Mathematical Imaging and Vision*, 65(2):253–276, 2023.

Hong TM Chu, Kim-Chuan Toh, and Yangjing Zhang. On regularized square-root regression problems: distributionally robust interpretation and fast computations. *The Journal of Machine Learning Research*, 23(1):13885–13923, 2022.

Arnak Dalalyan, Mohamed Hebiri, and Johannes C Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017.

I. Daubechies, R. DeVore, M. Fornasier, and C.S. Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Comm. Pure Appl. Math.*, 63:1–38, 2010.

Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 2019.

Alexis Derumigny. Improved bounds for square-root lasso and square-root slope. *Electronic Journal of Statistics*, 12(1):741 – 766, 2018.

Lars Diening, Massimo Fornasier, Tomasi Roland, and Maximilian Wank. A relaxed Kačanov iteration for the p-poisson problem. *Numer. Math.*, 145(2):1–34, 2020.

Sjoerd Dirksen, Guillaume Lecué, and Holger Rauhut. On the gap between restricted isometry properties and sparse recovery conditions. *IEEE Transactions on Information Theory*, 64(8): 5478–5487, 2018.

M. Fornasier, H. Rauhut, and R. Ward. Low-rank matrix recovery via iteratively reweighted least squares minimization. *SIAM Journal on Optimization*, 21:1614–1640, 2011.

S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer New York, 2013.

Simon Foucart. The sparsity of LASSO-type minimizers. *Applied and Computational Harmonic Analysis*, 62:441–452, 2023.

Simon Foucart, Eitan Tadmor, and Ming Zhong. On the sparsity of LASSO minimizers in sparse data recovery. *Constructive Approximation*, pages 1–19, 2022.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

Eric Gautier and Alexandre B Tsybakov. Pivotal estimation in high-dimensional regression via linear programming. In *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, pages 195–204. Springer, 2013.

Walter Gautschi. Some elementary inequalities relating to the gamma and incomplete gamma function. *Journal of Mathematics and Physics*, 38(1-4):77–81, 1959.

Christophe Giraud. *Introduction to high-dimensional statistics*. Monographs on statistics and applied probability ; 139. CRC Press, 2015.

Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. High-dimensional regression with unknown variance. *Statistical Science*, 27(4):500–518, 2012.

Tom Goldstein, Min Li, and Xiaoming Yuan. Adaptive primal-dual splitting methods for statistical learning and image processing. *Advances in neural information processing systems*, 28, 2015.

Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for machine learning. In *NIPS Workshop on Optimization for ML*, volume 3, pages 3–2, 2012.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC, 2015.

Frederik Hoppe, Felix Krahmer, Claudio Mayrink Verdun, Marion I Menzel, and Holger Rauhut. Uncertainty quantification for sparse fourier recovery. *arXiv preprint arXiv:2212.14864*, 2022.

Frederik Hoppe, Claudio Mayrink Verdun, Hannah Laus, Felix Krahmer, and Holger Rauhut. Uncertainty quantification for learned ista. In *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2023.

P. J. Huber. *Robust statistics*. Wiley New York, 1981.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Adrian Jarret, Julien Fageot, and Matthieu Simeoni. A fast and scalable polyatomic Frank-Wolfe algorithm for the LASSO. *IEEE Signal Processing Letters*, 29:637–641, 2022.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15(1):2869–2909, January 2014. ISSN 1532-4435.

Adel Javanmard and Andrea Montanari. Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593 – 2622, 2018.

He Jiang, Shihua Luo, and Yao Dong. Simultaneous feature selection and clustering based on square root optimization. *European Journal of Operational Research*, 289(1):214–231, 2021.

Donghwan Kim and Jeffrey A Fessler. Optimized first-order methods for smooth convex minimization. *Mathematical programming*, 159:81–107, 2016.

Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20 (1):282–303, 2014.

Vladimir Koltchinskii. Sparsity in penalized empirical risk minimization. In *Annales de l'IHP Probabilités et statistiques*, volume 45, pages 7–57, 2009.

Christian Kümmerle and Claudio Mayrink Verdun. Escaping saddle points in ill-conditioned matrix completion with a scalable second order method. In *Workshop on Beyond First Order Methods in ML Systems at the* 37^{th} *International Conference on Machine Learning*, 2020.

Christian Kümmerle and Claudio Mayrink Verdun. A scalable second order method for ill-conditioned matrix completion from few samples. In *Proceedings of 2021 International Conference on Machine Learning (ICML'21)*, 2021.

Christian Kümmerle, Claudio Mayrink Verdun, and Dominik Stöger. Iteratively reweighted least squares for ℓ_1-minimization with global linear convergence rate. In *Thirty-fifth Conference on Neural Information Processing Systems*, 2021.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

Stamatios Lefkimmiatis and Iaroslav Sergeevich Koshelev. Learning sparse and low-rank priors for image recovery via iterative reweighted least squares minimization. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingguo Li, Tuo Zhao, Xiaoming Yuan, and Han Liu. The flare package for high dimensional linear regression and precision matrix estimation in R. *Journal of Machine Learning Research*, 16(18): 553–557, 2015.

Xinguo Li, Haoming Jiang, Jarvis Haupt, Raman Arora, Han Liu, Mingyi Hong, and Tuo Zhao. On fast convergence of proximal algorithms for SQRT-Lasso optimization: Don't worry about its nonsmooth loss function. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 49–59, Tel Aviv, Israel, 22–25 Jul 2020. PMLR.

Xudong Li, Defeng Sun, and Kim-Chuan Toh. A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM J. Optim.*, 28(1):433–458, 2018.

Charles HC Little, Kee L Teo, and Bruce Van Brunt. *An Introduction to Infinite Products*. Springer, 2022.

Han Liu, Lie Wang, and Tuo Zhao. Calibrated multivariate regression with application to neural semantic basis discovery. *Journal of Machine Learning Research*, 2015.

Karim Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

Nathan Mankovich, Emily J King, Chris Peterson, and Michael Kirby. The flag median and flagIRLS. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10339–10347, 2022.

Oleh Melnyk, Michael Quellmalz, Gabriele Steidl, Noah Jaitner, Jakob Jordan, and Ingolf Sack. Time-harmonic optical flow with applications in elastography. *arXiv preprint arXiv:2405.15507*, 2024.

Stanislav Minsker, Mohamed Ndaoud, and Lang Wang. Robust and tuning-free sparse linear regression via square-root slope. *SIAM Journal on Mathematics of Data Science*, 6(2):428–453, 2024.

K. Mohan and M. Fazel. Iterative reweighted algorithms for matrix rank minimization. *Journal of Machine Learning Research*, 13:3441–3473, 2012.

Aaron J Molstad. New insights for the multivariate square-root lasso. *The Journal of Machine Learning Research*, 23(1):2878–2929, 2022.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, Vincent Leclère, and Joseph Salmon. Efficient smoothed concomitant lasso estimation for high dimensional regression. *Journal of Physics: Conference Series*, 904:012006, 2017.

Arkaddii S Nemirovskii and Yurii E Nesterov. Optimal methods of smooth convex minimization. *USSR Computational Mathematics and Mathematical Physics*, 25(2):21–30, 1985.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103: 127–152, 2005.

Brendan O'donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15:715–732, 2015.

José Luis Montiel Olea, Cynthia Rush, Amilcar Velez, and Johannes Wiesel. The out-of-sample prediction error of the square-root-lasso and related estimators. *arXiv preprint arXiv:2211.07608*, 2022.

A. B. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 19(2): 59–72, 2007.

Samet Oymak, Christos Thrampoulidis, and Babak Hassibi. The squared-error of generalized LASSO: a precise analysis. In *51st Annual Allerton Conference on Communication, Control, and Computing*, pages 1002–1009. IEEE, 2013.

Christopher C Paige and Michael A Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1):43–71, 1982.

Youngsuk Park, Sauptik Dhar, Stephen Boyd, and Mohak Shah. Variable metric proximal gradient method with diagonal Barzilai-Borwein stepsize. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3597–3601. IEEE, 2020.

Hendrik Bernd Petersen and Peter Jung. Robust instance-optimal recovery of sparse signals at unknown noise levels. *Information and Inference: A Journal of the IMA*, 11(3):845–887, 2021.

Clarice Poon and Gabriel Peyré. Smooth over-parameterized solvers for non-smooth structured optimization. *Mathematical Programming*, pages 1–56, 2023.

Feng Qi. Bounds for the ratio of two gamma functions. *Journal of Inequalities and Applications*, 2010(1):493058, 2010.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE transactions on information theory*, 57(10): 6976–6994, 2011.

Stephen Reid, Robert Tibshirani, and Jerome Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.

James Renegar and Benjamin Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *foundations of computational mathematics*, 22(1):211–256, 2022.

Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.

Jack Sherman and Winifred J Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950.

Nicolas Städler, Peter Bühlmann, and Sara van de Geer. Rejoinder: ℓ_1-penalization for mixture regression models (with discussion). *TEST*, 19(2):209–285, 2010.

Benjamin Stucky and Sara van de Geer. Sharp oracle inequalities for square root regularization. *Journal of Machine Learning Research*, 18(67):1–29, 2017.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 09 2012.

Peipei Tang, Chengjing Wang, Defeng Sun, and Kim-Chuan Toh. A sparse semismooth Newton based proximal majorization-minimization algorithm for nonconvex square-root-loss regression problem. *The Journal of Machine Learning Research*, 21(1):9253–9290, 2020.

Adrien Taylor and Yoel Drori. An optimal gradient method for smooth strongly convex minimization. *Mathematical Programming*, 199(1-2):557–594, 2023.

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

Kim-Chuan Toh, Michael J. Todd, and Reha H. Tütüncü. On the implementation and usage of SDPT3 - a Matlab software package for semidefinite-quadratic-linear programming, version 4.0. chapter Chapter 25, pages 715–754. Springer, 2012.

Quoc Tran-Dinh, Anastasios Kyrillidis, and Volkan Cevher. Composite self-concordant minimization. *J. Mach. Learn. Res.*, 16(1):371–416, 2015.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3), 2014. ISSN 0090-5364.

Sara van de Geer. *Estimation and Testing Under Sparsity: École d'Été de Probabilités de Saint-Flour XLV - 2015*. Springer Publishing Company, Incorporated, 1st edition, 2016. ISBN 3319327739.

Sara A van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Martin J Wainwright. Sharp thresholds for High-Dimensional and noisy sparsity recovery using ℓ_1-Constrained Quadratic Programming (Lasso). *IEEE transactions on information theory*, 55 (5):2183–2202, 2009.

E. Weiszfeld. Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tohoku Mathematical Journal, First Series*, 43:355–386, 1937.

John Wright and Yi Ma. *High-dimensional data analysis with low-dimensional models: Principles, computation, and applications*. Cambridge University Press, 2022.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. *IEEE Transactions on Information Theory*, 7(56):3561–3574, 2010.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

Guo Yu and Jacob Bien. Estimating the error variance in a high-dimensional linear model. *Biometrika*, 106(3):533–546, 2019.

Peiran Yu, Guoyin Li, and Ting Kei Pong. Kurdyka-Łojasiewicz exponent via inf-projection. *Foundations of Computational Mathematics*, 22(4):1171–1217, 2022.

Junhui Zhang, Jingkai Yan, and John Wright. Square root principal component pursuit: tuning-free noisy robust matrix recovery. *Advances in Neural Information Processing Systems*, 34:29464–29475, 2021.

Zaiwei Zhang, Xiangru Huang, Qixing Huang, Xiao Zhang, and Yuan Li. Joint learning of neural networks via iterative reweighted least squares. In *CVPR Workshops*, pages 18–26, 2019.

Peng Zhao and Bin Yu. On model selection consistency of Lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Tuo Zhao, Han Liu, and Tong Zhang. Pathwise coordinate optimization for sparse learning: Algorithm and theory. *The Annals of Statistics*, 46(1):180 – 218, 2018.

Supplementary material of *Fast, blind, and accurate: Noise-blind sparse regression with global linear convergence*.

Appendix A. Proof of Theorem 2

In this appendix, we establish the majorization property of the function $Q_\varepsilon(\tilde{z}, \tilde{\beta})$ described in Theorem 2 as well as the connection between the weight matrix $W_\varepsilon(\tilde{\beta})$ and the derivative of the function $f_\varepsilon(\tilde{\beta})$. Here, we repeat the theorems' statement for the reader's convenience.

Lemma 10 *The function in Equation (7) with W_ε as in (8) admits*

i. $W_\varepsilon(\tilde{\beta})\tilde{\beta} = \nabla \tilde{f}_\varepsilon(\tilde{\beta})$, *ii.* $Q_\varepsilon(\tilde{\beta}, \tilde{\beta}) = \tilde{f}_\varepsilon(\tilde{\beta})$, *iii.* $Q_\varepsilon(\tilde{z}, \tilde{\beta}) \geqslant \tilde{f}_\varepsilon(\tilde{z})$.

Proof: The first derivative of the smoothed objective function is given by

$$\nabla \tilde{f}_\varepsilon(\tilde{\beta}) = \frac{\tilde{X}^T \tilde{X} \tilde{\beta}}{\max\{\|\tilde{X}\tilde{\beta}\|_2, \xi\}} + \lambda \sum_{j=1}^{p} \frac{\beta_j e_j}{\max\{|\beta_j|, \delta\}}, \tag{14}$$

with $\{e_i\}_{i=0,\dots,p}$ being the standard basis vectors in $\mathbb{R}^{1+p}$. In view of the first condition, it is natural to define $W_\varepsilon(\tilde{\beta})$ as in (8). The second condition follows directly from Equation (7). To show that $Q_\varepsilon(\tilde{z}, \tilde{\beta})$ majorizes $f_\varepsilon(\tilde{z})$ for all $z \in \mathbb{R}^p$, we first rewrite Equation (7) as

$$\begin{aligned} Q_\varepsilon(\tilde{z}, \tilde{\beta}) &:= \tilde{f}_\varepsilon(\tilde{\beta}) + \langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}), \tilde{z}\rangle - \langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}), \tilde{\beta}\rangle + \frac{1}{2}\langle \tilde{z}, W_\varepsilon(\tilde{\beta})\tilde{z}\rangle + \frac{1}{2}\langle \tilde{\beta}, W_\varepsilon(\tilde{\beta})\tilde{\beta}\rangle - \langle \tilde{z}, W_\varepsilon(\tilde{\beta})\tilde{\beta}\rangle \\ &= \tilde{f}_\varepsilon(\tilde{\beta}) + \frac{1}{2}\langle \tilde{z}, W_\varepsilon(\tilde{\beta})\tilde{z}\rangle - \frac{1}{2}\langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}), \tilde{\beta}\rangle, \end{aligned} \tag{15}$$

where we used the second condition. Thus, in order to establish the majorization property iii., we need to prove the inequality

$$\begin{aligned} 0 \leqslant Q_\varepsilon(\tilde{z}, \tilde{\beta}) - \tilde{f}_\varepsilon(\tilde{z}) &= \tilde{f}_\varepsilon(\tilde{\beta}) - \tilde{f}_\varepsilon(\tilde{z}) - \frac{1}{2}\langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}), \tilde{\beta}\rangle + \frac{1}{2}\langle \tilde{z}, W_\varepsilon(\tilde{\beta})\tilde{z}\rangle \\ = j_\xi(\|\tilde{X}\tilde{\beta}\|_2) - j_\xi(\|\tilde{X}\tilde{z}\|_2) &- \frac{\|\tilde{X}\tilde{\beta}\|_2^2 - \|\tilde{X}\tilde{z}\|_2^2}{2\max\{\|\tilde{X}\tilde{\beta}\|_2, \xi\}} + \lambda \sum_{i=1}^{p}\left[j_\delta(\beta_i) - j_\delta(z_i) - \frac{|\beta_i|^2 - |z_i|^2}{2\max\{|\beta_i|, \delta\}}\right]. \end{aligned}$$

As all summands have a similar structure, we can prove that for $M \in \mathbb{R}^{p\times q}$, $\gamma > 0$ and for all $v, u \in \mathbb{R}^q$, it holds that

$$j_\gamma(\|Mv\|_2) - j_\gamma(\|Mu\|_2) - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\max\{\|Mv\|_2, \gamma\}} \geqslant 0. \tag{16}$$

Then, note that the theorem will be established by applying Equation (16), first with $M = \tilde{X}$ and $\gamma = \xi$ and finally with $M = E^{i,i}$, $i = 1, \dots, p$ and $\gamma = \delta$, where $E^{i,i}$ is a matrix with a single non-zero element $E^{i,i}_{i,i} = 1$.

In order to prove Equation (16), we consider four different values that the left-hand side may take depending on $\|Mv\|_2$, $\|Mu\|_2$ and γ. Let us consider each of them separately.
Case 1: $\|Mv\|_2 < \gamma$, $\|Mu\|_2 < \gamma$. Then, the left-hand side is given by

$$\frac{\|Mv\|_2^2}{2\gamma} + \frac{\gamma}{2} - \frac{\|Mu\|_2^2}{2\gamma} - \frac{\gamma}{2} - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\gamma} = 0.$$

Case 2: $\|Mv\|_2 < \gamma$, $\|Mu\|_2 \geqslant \gamma$. Using the arithmetic-geometric mean inequality, we get

$$\frac{\|Mv\|_2^2}{2\gamma} + \frac{\gamma}{2} - \|Mu\|_2 - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\gamma} = \frac{\|Mu\|_2^2}{2\gamma} + \frac{\gamma}{2} - \|Mu\|_2 \geqslant 0.$$

Case 3: $\|Mv\|_2 \geqslant \gamma$, $\|Mu\|_2 \geqslant \gamma$. Likewise, as in the previous case,

$$\|Mv\|_2 - \|Mu\|_2 - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\|Mv\|_2} = \frac{\|Mv\|_2}{2} + \frac{\|Mu\|_2^2}{2\|Mv\|_2} - \|Mu\|_2 \geqslant 0.$$

Case 4: $\|Mv\|_2 \geqslant \gamma$, $\|Mu\|_2 < \gamma$. We have

$$\|Mv\|_2 - \frac{\|Mu\|_2^2}{2\gamma} - \frac{\gamma}{2} - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\|Mv\|_2} = \frac{1}{2}\left[\|Mv\|_2 - \gamma + \|Mu\|_2^2\left[\frac{1}{\|Mv\|_2} - \frac{1}{\gamma}\right]\right].$$

Since $\|Mv\|_2 \geqslant \gamma$, the second term is negative, and we can further decrease it by applying $\|Mu\|_2 < \gamma$, which gives

$$\begin{aligned}\|Mv\|_2 - \frac{\|Mu\|_2^2}{2\gamma} - \frac{\gamma}{2} - \frac{\|Mv\|_2^2 - \|Mu\|_2^2}{2\|Mv\|_2} &\geqslant \frac{1}{2}\left[\|Mv\|_2 - \gamma + \gamma^2\left[\frac{1}{\|Mv\|_2} - \frac{1}{\gamma}\right]\right] \\ &= \frac{1}{2}\left[\|Mv\|_2 + \frac{\gamma}{\|Mv\|_2} - 2\gamma\right] \geqslant 0,\end{aligned}$$

where the last inequality is the arithmetic-geometric mean inequality again. □

Appendix B. Global Convergence With Smoothing Parameter Decay

In this section, we present a proof of Theorem 4. The first step is to quantify the function value decay for a single iteration.

Lemma 11 (General function value decay rate) *Fix $\xi, \delta > 0$ and let $\beta \in \mathbb{R}^p$. If the iterate β^k of Algorithm 1 satisfies $\langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}^k), \tilde{\beta} - \tilde{\beta}^k \rangle \leqslant 0$ and $\beta \neq \beta^k$, then we have*

$$f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta^k) \leqslant Q_\varepsilon(\tilde{\beta}^{k+1}, \tilde{\beta}^k) - f_\varepsilon(\beta^k) \leqslant -\frac{|\langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}^k), \tilde{\beta} - \tilde{\beta}^k \rangle|^2}{2\langle W_\varepsilon(\tilde{\beta}^k)(\tilde{\beta} - \tilde{\beta}^k), \tilde{\beta} - \tilde{\beta}^k \rangle}.$$

Proof: By construction, $f_\varepsilon(\beta^{k+1}) = Q_\varepsilon(\tilde{\beta}^{k+1}, \tilde{\beta}^k) \leqslant Q_\varepsilon(\tilde{z}, \tilde{\beta}^{k+1})$ for any $\tilde{z} \in \mathbb{R}^{p+1}$ such that $\tilde{z}_0 = 1$. Consider $v^k = \tilde{\beta} - \tilde{\beta}^k$ and let us evaluate the difference $Q_\varepsilon(\tilde{\beta}^k + tv^k, \tilde{\beta}^k) - \tilde{f}_\varepsilon(\tilde{\beta}^k)$ for some $t > 0$. The idea is that for a properly chosen $t > 0$, the difference $Q_\varepsilon(\tilde{\beta}^k + tv^k, \tilde{\beta}^k) - \tilde{f}_\varepsilon(\tilde{\beta}^k)$ will be negative, which will imply that $\tilde{f}_\varepsilon(\tilde{\beta}^{k+1}) < \tilde{f}_\varepsilon(\tilde{\beta}^k)$. Expanding $Q_\varepsilon(\tilde{\beta}^k + tv^k, \tilde{\beta}^k)$ yields

$$Q_\varepsilon(\tilde{\beta}^k + tv^k, \tilde{\beta}^k) - \tilde{f}_\varepsilon(\tilde{\beta}^k) = t\langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}^k), v^k \rangle + \frac{t^2}{2}\langle W_\varepsilon(\tilde{\beta}^k)v^k, v^k \rangle = bt + at^2. \tag{17}$$

This is a quadratic polynomial with a positive leading coefficient. Indeed,

$$a = \frac{1}{2}\langle W_\varepsilon(\tilde{\beta}^k)v^k, v^k \rangle = \frac{1}{2}\frac{\|\tilde{X}v^k\|_2^2}{\max\{\|\tilde{X}\tilde{\beta}^k\|_2, \xi\}} + \frac{\lambda}{2}\sum_{j=1}^{p}\frac{|v_j^k|^2}{\max\{|\tilde{\beta}_j^k|, \delta\}} \geqslant 0.$$

The equality is possible if and only if all summands are equal to zero. This means that $v^k = 0$, i.e., $\beta = \beta^k$ which contradicts our assumption. Hence, the quadratic polynomial $Q_\varepsilon(\tilde{\beta}^k + tv^k, \tilde{\beta}^k) - \tilde{f}_\varepsilon(\beta^k)$ attains its minimum at $t = -b/2a$. Therefore, we have that

$$\begin{aligned} f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta^k) &\leqslant Q_\varepsilon(\tilde{\beta}^{k+1}, \tilde{\beta}^k) - f_\varepsilon(\beta^k) = \min_{z \ s.t.\ z_0=1} Q_\varepsilon(\tilde{z}, \tilde{\beta}^k) - f_\varepsilon(\beta^k) \\ &\leqslant Q_\varepsilon(\tilde{\beta}^k + tv^k, \tilde{\beta}^k) - f_\varepsilon(\beta^k) \leqslant -\frac{b^2}{4a} = -\frac{|\langle \nabla \tilde{f}_\varepsilon(\beta^k), v^k\rangle|^2}{2\langle W_\varepsilon(\tilde{\beta}^k)v^k, v^k\rangle}. \end{aligned}$$

□

Remark 12 *The essence of Theorem 11 shares similarities with the sufficient decrease lemma (Beck, 2017, Lemma 10.4), which is commonly employed to establish convergence properties of the proximal gradient descent method. However, the latter introduces the gradient mapping G_L as a generalization of the traditional gradient concept, which enables the establishment of convergence guarantees tailored specifically for the proximal gradient descent algorithm, while here, for the sake of completeness, we established it directly for the square-root LASSO objective.*

With the help of Theorem 11, we can establish the first result regarding the sublinear convergence of IRLS. In particular, we will show that the sequence $\{\beta^k\}$ converges sublinearly to the minimizer of the regularized problem, here denoted by β^*_ε.

Theorem 13 *Let $\xi, \delta > 0$ and $k \geqslant 2$ and denote by β^*_ε the minimizer of f_ε. Then the iterates β^k of Algorithm 1 satisfy the inequality*

$$f_\varepsilon(\beta^k) - f_\varepsilon(\beta^*_\varepsilon) \leqslant \max\left\{2^{-\frac{k-1}{2}}(f_\varepsilon(\beta^0) - f_\varepsilon(\beta^*_\varepsilon)), \frac{8(\xi^{-1}\|X\|^2 + \lambda\delta^{-1})(f_\varepsilon(\beta^0) + f_0(\beta^*_\varepsilon))^2}{\lambda^2(k-1)}\right\}.$$

Proof: We apply Theorem 11 with $\beta = \beta^*_\varepsilon$. Let $v^k = \tilde{\beta}^*_\varepsilon - \tilde{\beta}^k$. Note that convexity of $\tilde{f}_\varepsilon$ gives

$$f_\varepsilon(\beta^*_\varepsilon) = \tilde{f}_\varepsilon(\tilde{\beta}^*_\varepsilon) \geqslant \tilde{f}_\varepsilon(\tilde{\beta}^k) + \langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}^k), v^k\rangle = f_\varepsilon(\beta^k) + \langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}^k), v^k\rangle,$$

or, equivalently,

$$\langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}^k), v^k\rangle \leqslant f_\varepsilon(\beta^*_\varepsilon) - f_\varepsilon(\beta^k) \leqslant 0 \text{ and } |\langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}^k), v^k\rangle| \geqslant f_\varepsilon(\beta^k) - f_\varepsilon(\beta^*_\varepsilon). \tag{18}$$

Hence, Theorem 11 yields

$$f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta^k) \leqslant -\frac{|\langle \nabla \tilde{f}_\varepsilon(\tilde{\beta}^k), v^k\rangle|^2}{2\langle W_\varepsilon(\tilde{\beta}^k)v^k, v^k\rangle} \leqslant -\frac{(f_\varepsilon(\beta^k) - f_\varepsilon(\beta^*_\varepsilon))^2}{2\langle W_\varepsilon(\tilde{\beta}^k)v^k, v^k\rangle}. \tag{19}$$

Let us show that the denominator is bounded from above. By construction, we have

$$\begin{aligned} \langle W(\tilde{\beta}^k)v^k, v^k\rangle &= \frac{\|\tilde{X}v^k\|_2^2}{\max\{\|\tilde{X}\tilde{\beta}^k\|_2, \xi\}} + \lambda\sum_{j=1}^p \frac{|v_j^k|^2}{\max\{|\tilde{\beta}_j^k|^2, \delta\}} \\ &\leqslant \frac{\|X(\beta^*_\varepsilon - \beta^k)\|_2^2}{\xi} + \frac{\lambda}{\delta}\sum_{j=1}^p |(\beta^*_\varepsilon - \beta^k)_j|^2 \leqslant (\xi^{-1}\|X\|^2 + \lambda\delta^{-1})\|\beta^*_\varepsilon - \beta^k\|_2^2. \end{aligned}$$

Consequently, we only need to bound $\|\beta_\varepsilon^* - \beta^k\|_2$. This can be done by contradiction. Assume that $\|\beta_\varepsilon^* - \beta^k\|_2 > \lambda^{-1}(f_0(\beta^k) + f_0(\beta_\varepsilon^*))$. Then, by the reverse triangle inequality, we have

$$\begin{aligned} f_0(\beta^k) &= \|X\beta^k - b\|_2 + \lambda\|\beta^k\|_1 \\ &\geqslant \|X(\beta_\varepsilon^* - \beta^k)\|_2 - \|X\beta_\varepsilon^* + b\|_2 + \lambda\|\beta_\varepsilon^* - \beta^k\|_1 - \lambda\|\beta_\varepsilon^*\|_1. \end{aligned}$$

The first term is nonnegative. The second and the fourth terms together are equal to $-f_0(\beta_\varepsilon^*)$. The third term can be bounded from below by $\lambda\|\beta_\varepsilon^* - \beta^k\|_2$ by monotonicity of the ℓ_p-norms. Hence, using the assumption $\|\beta_\varepsilon^* - \beta^k\|_2 > \lambda^{-1}(f_0(\beta^k) + f_0(\beta_\varepsilon^*))$, we get

$$f_0(\beta^k) \geqslant 0 + \lambda\|\beta_\varepsilon^* - \beta^k\|_2 - f_0(\beta_\varepsilon^*) > \lambda\lambda^{-1}(f_0(\beta^k) + f_0(\beta_\varepsilon^*)) - f_0(\beta_\varepsilon^*) = f_0(\beta^k),$$

which is a contradiction. Therefore,

$$\|\beta_\varepsilon^* - \beta^k\|_2 \leqslant \lambda^{-1}(f_0(\beta^k) + f_0(\beta_\varepsilon^*)) \leqslant \lambda^{-1}(f_\varepsilon(\beta^k) + f_0(\beta_\varepsilon^*)) \leqslant \lambda^{-1}(f_\varepsilon(\beta^0) + f_0(\beta_\varepsilon^*)),$$

where we used that $f_\varepsilon(\beta^k) \geqslant f_\varepsilon(\beta^{k+1}) \geqslant f_0(\beta^{k+1})$ for all $k \geqslant 0$. Now, we substitute the obtained bound in (19), which leads to

$$\begin{aligned} [f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta_\varepsilon^*)] - [f_\varepsilon(\beta^k) - f_\varepsilon(\beta_\varepsilon^*)] &\leqslant -\frac{\lambda^2(f_\varepsilon(\beta^k) - f_\varepsilon(\beta_\varepsilon^*))^2}{2(\xi^{-1}\|X\|^2 + \lambda\delta^{-1})(f_\varepsilon(\beta^0) + f_0(\beta_\varepsilon^*))^2} \\ &\leqslant -\frac{\lambda^2(f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta_\varepsilon^*))^2}{2(\xi^{-1}\|X\|^2 + \lambda\delta^{-1})(f_\varepsilon(\beta^0) + f_0(\beta_\varepsilon^*))^2} \end{aligned} \tag{20}$$

The result of the theorem follows by applying (Beck, 2015, Lemma 3.8) for the sequence $\{f_\varepsilon(\beta^k) - f_\varepsilon(\beta^*)\}_{k\geqslant 0}$. □

Remark 14 *Theorem 13 is similar in its nature to (Beck, 2015, Theorem 4.2). The difference is that in (Beck, 2015, Theorem 4.2), the proof is given for an alternating minimization strategy. This would correspond to the regularized Scaled LASSO objective function, while here, we establish it directly for the regularized square-root LASSO formulation without an alternating procedure.*

We can now establish a new linear convergence analysis when the objective gap is away from zero.

Theorem 15 *Let $\xi, \delta > 0$ and denote by β_ε^* the minimizer of f_ε. If $f_\varepsilon(\beta^k) - f_\varepsilon(\beta_\varepsilon^*) \geqslant \gamma$ for some constant $\gamma > 0$, then the iterates β^k of Algorithm 1 satisfy the inequality*

$$f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta_\varepsilon^*) \leqslant \Big(f_\varepsilon(\beta^k) - f_\varepsilon(\beta_\varepsilon^*)\Big)\left[1 - \frac{\gamma}{2\max\{\frac{1}{\xi}, \frac{1}{\lambda\delta}\}f_0^2(\beta_\varepsilon^*) + 4\gamma}\right].$$

Proof: We start by noticing that it would be possible to obtain a linear decay rate by substituting the assumption $f_\varepsilon(\beta^k) - f_\varepsilon(\beta_\varepsilon^*) \geqslant \gamma$ into (20). However, such a rate would depend on $f_0(\beta^0)$, which can potentially be large. Thus, we take a step back to (19) and bound the denominator $\langle W_\varepsilon(\tilde\beta^k)v^k, v^k\rangle$ with $v^k = \tilde\beta_\varepsilon^* - \tilde\beta^k$ differently. More precisely, we first decompose it into two parts and connect it with the first derivative $\langle \nabla\tilde f_\varepsilon(\tilde\beta^k), v^k\rangle$ by using that $W_\varepsilon(\tilde\beta)\tilde\beta = \nabla\tilde f_\varepsilon(\tilde\beta)$ and that $W_\varepsilon(\tilde\beta)$ is a self-adjoint and positive semidefinite matrix,

$$\begin{aligned}\langle W_\varepsilon(\tilde\beta^k)v^k, v^k\rangle &= \langle W_\varepsilon(\tilde\beta^k)v^k, \tilde\beta^*_\varepsilon\rangle - \langle W_\varepsilon(\tilde\beta^k)v^k, \tilde\beta^k\rangle \\ &= \langle W_\varepsilon(\tilde\beta^k)(\tilde\beta^*_\varepsilon - \tilde\beta^k), \tilde\beta^*_\varepsilon\rangle - \langle \nabla \tilde f_\varepsilon(\tilde\beta^k), v^k\rangle \\ &= \langle W_\varepsilon(\tilde\beta^k)\tilde\beta^*_\varepsilon, \tilde\beta^*_\varepsilon\rangle - \langle W_\varepsilon(\tilde\beta^k)\tilde\beta^k, \tilde\beta^*_\varepsilon \pm \tilde\beta^k\rangle - \langle \nabla \tilde f_\varepsilon(\tilde\beta^k), v^k\rangle \\ &= \langle W_\varepsilon(\tilde\beta^k)\tilde\beta^*_\varepsilon, \tilde\beta^*_\varepsilon\rangle - \langle W_\varepsilon(\tilde\beta^k)\tilde\beta^k, \tilde\beta^k\rangle - 2\langle \nabla \tilde f_\varepsilon(\tilde\beta^k_\varepsilon), v^k\rangle \\ &\leqslant \langle W_\varepsilon(\tilde\beta^k)\tilde\beta^*_\varepsilon, \tilde\beta^*_\varepsilon\rangle + 2|\langle \nabla \tilde f_\varepsilon(\tilde\beta^k), v^k\rangle|.\end{aligned} \tag{21}$$

Next, the term $\langle W_\varepsilon(\tilde\beta^k)\tilde\beta^*_\varepsilon, \tilde\beta^*_\varepsilon\rangle$ is bounded from above as

$$\begin{aligned}\langle W_\varepsilon(\tilde\beta^k)\tilde\beta^*_\varepsilon, \tilde\beta^*_\varepsilon\rangle &= \frac{\|\tilde X\tilde\beta^*_\varepsilon\|_2^2}{\max\{\|\tilde X\tilde\beta^k\|_2, \xi\}} + \lambda\sum_{j=1}^p \frac{|(\beta^*_\varepsilon)_j|^2}{\max\{|\beta^k_j|, \delta\}} \\ &\leqslant \frac{\|\tilde X\tilde\beta^*_\varepsilon\|_2^2}{\xi} + \frac{\lambda^2}{\lambda}\sum_{j=1}^p \frac{|(\beta^*_\varepsilon)_j|^2}{\delta} \\ &\leqslant \max\left\{\frac{1}{\xi}, \frac{1}{\lambda\delta}\right\}\left[\|X\beta^*_\varepsilon - y\|_2^2 + \lambda^2\sum_{j=1}^p |(\beta^*_\varepsilon)_j|^2\right] \\ &\leqslant \max\left\{\frac{1}{\xi}, \frac{1}{\lambda\delta}\right\}\left[\|X\beta^*_\varepsilon - y\|_2 + \lambda\sum_{j=1}^p |(\beta^*_\varepsilon)_j|\right]^2 = \max\left\{\frac{1}{\xi}, \frac{1}{\lambda\delta}\right\} f_0^2(\beta^*_\varepsilon).\end{aligned} \tag{22}$$

Now, turning to the nominator in Equation (19), the bound in (18) gives

$$|\langle \nabla \tilde f_\varepsilon(\tilde\beta^k), v^k\rangle|^2 \geqslant |\langle \nabla \tilde f_\varepsilon(\tilde\beta^k), v^k\rangle|[f_\varepsilon(\beta^k) - f_0(\beta^*_\varepsilon)]. \tag{23}$$

By plugging Equation (21), Equation (22) and Equation (23) into Equation (19), we obtain

$$f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta^*_\varepsilon) \leqslant \Big[f_\varepsilon(\beta^k) - f_\varepsilon(\beta^*_\varepsilon)\Big]\left[1 - \frac{|\langle \nabla \tilde f_\varepsilon(\tilde\beta^k), v^k\rangle|}{2\max\{\frac{1}{\xi}, \frac{1}{\lambda\delta}\}f_0^2(\beta^*_\varepsilon) + 4|\langle \nabla \tilde f_\varepsilon(\tilde\beta^k), v^k\rangle|}\right],$$

The second term of the right-hand side has the form $1 - \frac{t}{a+4t} = \frac{3}{4} + \frac{a}{4a+16t}$, where $t = |\langle \nabla \tilde f_\varepsilon(\tilde\beta^k), v^k\rangle| \geqslant \gamma$. The function $\frac{a}{4a+16t}$ is decreasing and, thus, attains its maximum at $t = \gamma$. This gives

$$f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta^*) \leqslant \Big[f_\varepsilon(\beta^{k+1}) - f_\varepsilon(\beta^*)\Big]\left[1 - \frac{\gamma}{2\max\{\frac{1}{\xi}, \frac{1}{\lambda\delta}\}f_0^2(\beta^*) + 4\gamma}\right], \tag{24}$$

□

In contrast to the previous theorem, in the case with decreasing ε_k, we require a more intricate analysis since the condition $\langle \nabla \tilde f_{\varepsilon_k}(\tilde\beta^k), \tilde\beta^*_0 - \tilde\beta^k\rangle \leqslant 0$ used in Theorem 11 may no longer hold. Yet, in the following, we obtain a similar result, where the bound on the objective gap varies with smoothing parameters.

Theorem 16 *Let $\{\xi_k\}_{k\geqslant 0}$ and $\{\delta_k\}_{k\geqslant 0}$ be two non-increasing sequences and set β_0^* as the minimizer of f_0. If $f_\varepsilon(\beta^k) - f_\varepsilon(\beta_\varepsilon^*) \geqslant \gamma(\lambda p\delta_k + \xi_k)$ for some constant $\gamma > 1$, then the iterates β^k of Algorithm 1 satisfy the inequality*

$$f_{\varepsilon_k}(\beta^{k+1}) - f_0(\beta_0^*) \leqslant \Big(f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)\Big)\left[1 - \frac{(\gamma-1)^2(\lambda p\delta_k + \xi_k)}{2\gamma\max\{\frac{1}{\xi_k}, \frac{1}{\lambda\delta_k}\}f_0^2(\beta_0^*) + 4(\gamma-1)^2(\lambda p\delta_k + \xi_k)}\right].$$

Proof: The proof follows a similar structure to that of Theorems 4 and 5 discussed earlier. More specifically, we employ Theorem 11 and establish an upper bound for the denominator, analogous to the approach utilized in Theorem 5. However, Theorem 11 relies on Equation (18), which is no longer true for β_0^*. As a consequence, we derive an alternative bound to address this issue. Similarly to the previous case, the convexity of $\tilde{f}_\varepsilon$ yields

$$f_{\varepsilon_k}(\beta_0^*) = \tilde{f}_{\varepsilon_k}(\tilde{\beta}_0^*) \geqslant \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k) + \langle\nabla\tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{\beta}_0^* - \tilde{\beta}^k\rangle = f_{\varepsilon_k}(\beta^k) + \langle\nabla\tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{\beta}_0^* - \tilde{\beta}^k\rangle$$

Then, by $f_0(\beta) \leqslant f_\varepsilon(\beta) \leqslant f_0(\beta) + \xi + \lambda p\delta$ and $f_{\varepsilon_1}(\beta) \leqslant f_{\varepsilon_2}(\beta)$ whenever $0 \leqslant \xi_1 \leqslant \xi_2$, $0 \leqslant \delta_1 \leqslant \delta_2$ and the assumption $f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) \geqslant \gamma(\lambda p\delta_k + \xi_k)$, we obtain

$$\begin{aligned} -\langle\nabla\tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{\beta}_0^* - \tilde{\beta}^k\rangle &\geqslant f_{\varepsilon_k}(\beta^k) - f_{\varepsilon_k}(\beta_0^*) \geqslant f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) - (\lambda p\delta_k + \xi_k) \\ &\geqslant (\gamma-1)(\lambda p\delta_k + \xi_k) \geqslant 0, \end{aligned} \tag{25}$$

and

$$-\langle\nabla\tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{\beta}_0^* - \tilde{\beta}^k\rangle \geqslant f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) - (\lambda p\delta_k + \xi_k) \geqslant (1 - 1/\gamma)[f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)].$$

The use of these bounds instead of (18) leads to the desired result. □

Without additional assumptions on the decay of smoothing parameters, we can only guarantee that the steps $\beta^{k+1} - \beta^k$ converge to zero, similarly to the standard results for gradient descent.

Theorem 17 *Let $\{\xi_k\}_{k\geqslant 0}$ and $\{\delta_k\}_{k\geqslant 0}$ be two non-increasing sequences. Then, the iterates β^k generated by Algorithm 1 satisfy*

$$\lim_{k\to\infty}\|\beta^{k+1} - \beta^k\|_2 = 0 \text{ and } \min_{k=0,\ldots,K-1}\|\beta^{k+1} - \beta^k\|_2^2 \leqslant \frac{2\max\{\lambda\delta_0, f_{\varepsilon_0}(\beta^0)\}}{\lambda^2 K}[f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*)].$$

Proof: We start by quantifying the difference between $f_{\varepsilon_{k+1}}(\beta^{k+1})$ and $f_{\varepsilon_k}(\beta^k)$, i.e., by proving that

$$f_{\varepsilon_{k+1}}(\beta^{k+1}) - f_{\varepsilon_k}(\beta^k) \leqslant Q_{\varepsilon_k}(\tilde{\beta}^{k+1}, \tilde{\beta}^k) - f_{\varepsilon_k}(\beta^k) = -\tfrac{1}{2}\langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k\rangle.$$

If $\beta^{k+1} = \beta^k$, the first inequality is trivial. Otherwise, we apply Theorem 11 with $\beta = \beta^{k+1}$ instead of $\beta = \beta_0^*$ as it was done in the previous proofs. Note that by the convexity of $\tilde{f}_{\varepsilon_k}$ and definition of β^{k+1}, the assumption of Theorem 11 is satisfied,

$$\langle\nabla\tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \beta^{k+1} - \beta^k\rangle \leqslant f_{\varepsilon_k}(\beta^{k+1}) - f_{\varepsilon_k}(\beta^k) \leqslant 0. \tag{26}$$

Thus, we get

$$Q_{\varepsilon_k}(\tilde{\beta}^{k+1}, \tilde{\beta}^k) - f_{\varepsilon_k}(\beta^k) \leqslant -\frac{|\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle|^2}{2\langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle}.$$

Substituting the definition of $Q_{\varepsilon_k}(\tilde{\beta}^{k+1}, \tilde{\beta}^k)$ and combining it with (26) gives

$$\begin{aligned} -|\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle| + \frac{1}{2}\langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle& \\ \leqslant -\frac{|\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle|^2}{2\langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle}&. \end{aligned}$$

Let us denote

$$a := |\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle| \geqslant 0 \quad \text{and} \quad b := \langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle > 0.$$

Then, the inequality above is equivalent to $-2ab+b^2 \leqslant -a^2$ and $(a-b)^2 \leqslant 0$, which is only possible if $a = b$. Another way of seeing that this holds is to look at the KKT conditions of the problem $\min_{z \in \mathbb{R}^{p+1},\ z_0=1} Q_{\varepsilon_k}(\tilde{z}, \tilde{\beta}^k)$. In fact, it holds that $\langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), v \rangle = -\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), v \rangle$ for all $v \in \mathbb{R}^{p+1}$ such that $v_0 = 0$. Substituting the obtained equality into $Q_{\varepsilon_k}(\tilde{\beta}^{k+1}, \tilde{\beta}^k)$ gives

$$Q_{\varepsilon_k}(\tilde{\beta}^{k+1}, \tilde{\beta}^k) = \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k) - a + \tfrac{b}{2} = \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k) - \tfrac{1}{2}\langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle.$$

Together with the majorization property from Theorem 2, it yields the inequality stated at the beginning of the proof, namely,

$$\tilde{f}_{\varepsilon_{k+1}}(\tilde{\beta}^{k+1}) \leqslant \tilde{f}_{\varepsilon_k}(\tilde{\beta}^{k+1}) \leqslant Q_{\varepsilon_k}(\tilde{\beta}^{k+1}, \tilde{\beta}^k) = \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k) - \tfrac{1}{2}\langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle. \tag{27}$$

Now, we bound the quadratic term from below in terms of the squared distance $\|\beta^{k+1} - \beta^k\|_2^2$. Using the definition of $W_{\varepsilon_k}(\tilde{\beta}^k)$, we get

$$\begin{aligned} \langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle &= \frac{\|\tilde{X}(\tilde{\beta}^{k+1} - \tilde{\beta}^k)\|_2^2}{\max\{\|\tilde{X}\tilde{\beta}^k\|_2^2, \xi_k\}} + \lambda \sum_{j=1}^{p} \frac{|(\tilde{\beta}^{k+1} - \tilde{\beta}^k)_j|^2}{\max\{|\tilde{\beta}^k_j|, \delta_k\}} \\ &\geqslant 0 + \lambda \sum_{j=1}^{p} \frac{|(\beta^{k+1} - \beta^k)_j|^2}{\max\{|\beta^k_j|, \delta_k\}} \geqslant \frac{\lambda \|\beta^{k+1} - \beta^k\|_2^2}{\max\{\|\beta^k\|_\infty, \delta_k\}}. \end{aligned}$$

Furthermore, by construction, we have $\xi_k \leqslant \xi_0$ and $\delta_k \leqslant \delta_0$ so that

$$\lambda\|\beta^k\|_\infty \leqslant \lambda\|\beta^k\|_1 \leqslant \|X\beta^k - y\|_2 + \lambda\|\beta^k\|_1 = f_0(\beta^k) \leqslant f_{\varepsilon_k}(\beta^k) \leqslant f_{\varepsilon_0}(\beta^0).$$

Consequently, the quadratic term satisfies

$$\langle W_{\varepsilon_k}(\tilde{\beta}^k)(\tilde{\beta}^{k+1} - \tilde{\beta}^k), \tilde{\beta}^{k+1} - \tilde{\beta}^k \rangle \geqslant \frac{\lambda\|\beta^{k+1} - \beta^k\|_2^2}{\max\{\lambda^{-1} f_{\varepsilon_0}(\beta^0), \delta_0\}} = \frac{\lambda^2\|\beta^{k+1} - \beta^k\|_2^2}{\max\{f_{\varepsilon_0}(\beta^0), \lambda\delta_0\}}.$$

Returning to Equation (27), we obtain

$$\frac{\lambda^2\|\beta^{k+1} - \beta^k\|_2^2}{2\max\{f_{\varepsilon_0}(\beta^0), \lambda\delta_0\}} \leqslant \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k) - \tilde{f}_{\varepsilon_{k+1}}(\tilde{\beta}^{k+1}) = f_{\varepsilon_k}(\beta^k) - f_{\varepsilon_{k+1}}(\beta^{k+1})$$

Summing up for $k = 0, \dots, K-1$, for some $K \in \mathbb{N}$, leads to

$$\sum_{k=0}^{K-1} \frac{\lambda^2 \|\beta^{k+1} - \beta^k\|_2^2}{2\max\{f_{\varepsilon_0}(\beta^0), \lambda\delta_0\}} \leqslant \sum_{k=0}^{K-1} [f_{\varepsilon_k}(\beta^k) - f_{\varepsilon_{k+1}}(\beta^{k+1})] = f_{\varepsilon_0}(\beta^0) - f_{\varepsilon_k}(\beta^K).$$

By taking $\frac{\lambda^2}{2\max\{f_{\varepsilon_0}(\beta^0),\lambda\delta_0\}}$ to the right-hand side, we observe that the partial sum of the series is bounded from above by

$$\sum_{k=0}^{K-1} \|\beta^{k+1}-\beta^k\|_2^2 \leqslant \frac{2\max\{f_{\varepsilon_0}(\beta^0), \lambda\delta_0\}}{\lambda^2}[f_{\varepsilon_0}(\beta^0)-f_{\varepsilon_k}(\beta^K)] \leqslant \frac{2\max\{f_{\varepsilon_0}(\beta^0), \lambda\delta_0\}}{\lambda^2}[f_{\varepsilon_0}(\beta^0)-f_0(\beta_0^*)].$$

This bound is independent of K and, thus, the series $\sum_{k=0}^{\infty} \|\beta^{k+1} - \beta^k\|_2^2$ is convergent. As a result its summands $\|\beta^{k+1} - \beta^k\|_2^2$ converge to zero as $k \to \infty$. Finally, we bound the minimum of the first K summands by their mean,

$$\min_{k=0,\dots,K-1} \|\beta^{k+1} - \beta^k\|_2^2 \leqslant \frac{1}{K}\sum_{k=0}^{K-1} \|\beta^{k+1} - \beta^k\|_2^2 \leqslant \frac{2\max\{f_{\varepsilon_0}(\beta^0), \lambda\delta_0\}}{\lambda^2 K}[f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*)].$$

□

Appendix C. Discussion on the smoothing parameter decay

An important element of Theorem 16 is the condition $f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) \geqslant \gamma(\lambda p\delta_k + \xi_k)$, which plays a crucial role in establishing convergence. While Theorem 5 and Theorem 16 share similarities, the latter exhibits a convergence rate that approaches one as δ_k and ξ_k tend to zero. However, depending on the rate at which the smoothing parameters approach zero, the bound presented in Theorem 16 may not always yield a meaningful convergence result. In this section, we establish a connection between the decay of smoothing parameters and the convergence rate, providing further insights into the analysis. We start by proving one lemma that illustrates the importance of the regularization parameter in the convergence rate.

Lemma 18 *Let $K \geqslant 1$, $\gamma > 1$ and $0 < \nu < 1$. Assume that $\xi_k = \lambda\delta_k$ and define*

$$c := \gamma\lambda(p+1) \quad \textit{and} \quad d := \frac{2\gamma f_0^2(\beta_0^*)}{\lambda^2(\gamma-1)^2(p+1)} \tag{28}$$

Then, the iterate β^{K+1} of Algorithm 1 admits

$$f_{\varepsilon_{K+1}}(\beta^{K+1}) - f_0(\beta_0^*) \leqslant \max\left\{ c\delta_{\lfloor K^\nu \rfloor}, (f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*)) \prod_{k=\lfloor K^\nu \rfloor}^{K} \left[1 - \frac{1}{d\delta_k^{-2} + 4}\right] \right\},$$

where $\delta_{\lfloor K^\nu \rfloor}$ is the regularization parameter δ_k at the iteration $k = \lfloor K^\nu \rfloor$ for a certain $0 < \nu < 1$.

Proof: The proof differentiates between two possible cases depending on how many times the inequality $f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) < c\delta_k$ is satisfied. Firstly, assume that it is satisfied for at least $\lfloor K^\nu \rfloor$ indices k and denote the largest one of them by k_0, i.e., $k_0 \geqslant \lfloor K^\nu \rfloor$. Then, by the inequality $0 \leqslant f_{\varepsilon_{k+1}}(\beta^{k+1}) \leqslant f_{\varepsilon_k}(\beta^{k+1}) \leqslant f_{\varepsilon_k}(\beta^k)$, we have

$$f_{\varepsilon_{K+1}}(\beta^{K+1}) \leqslant f_{\varepsilon_k}(\beta^K) \leqslant \ldots \leqslant f_{\varepsilon_{k_0}}(\beta^{k_0}) \leqslant f_{\varepsilon_{\lfloor K^\nu \rfloor}}(\beta^{\lfloor K^\nu \rfloor}) < f_0(\beta_0^*) + c\delta_{\lfloor K^\nu \rfloor}.$$

Now, let us assume that the opposite holds, i.e., that there are less than $\lfloor K^\nu \rfloor$ indices k for which $f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) < c\delta_k$. Then, there are at least $K - \lfloor K^\nu \rfloor + 1$ indices k such that the opposite inequality $f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) \geqslant c\delta_k$ holds. Let us denote all these indices by a set $\mathcal{K}$. If $k \notin \mathcal{K}$, we can use the bound $f_{\varepsilon_{k+1}}(\beta^{k+1}) - f_0(\beta_0^*) \leqslant f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)$ that holds due to the monotonicity of δ_k. Otherwise, by Theorem 16, we have

$$f_{\varepsilon_{k+1}}(\beta^{k+1}) - f_0(\beta_0^*) \leqslant f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) \left[1 - \frac{1}{d\delta_k^{-2} + 4}\right].$$

Combining these two bounds yields

$$f_{\varepsilon_{k+1}}(\beta^{K+1}) - f_0(\beta_0^*) \leqslant (f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*)) \prod_{k \in \mathcal{K}} \left[1 - \frac{1}{d\delta_k^{-2} + 4}\right]$$

By construction, $\delta_{k+1} \leqslant \delta_k$. Thus, the product on the right-hand side is the largest when the set $\mathcal{K}$ is $\{\lfloor K^\nu \rfloor, \ldots, K\}$, which gives

$$f_{\varepsilon_{k+1}}(\beta^{K+1}) - f_0(\beta_0^*) \leqslant (f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*)) \prod_{k=\lfloor K^\nu \rfloor}^{K} \left[1 - \frac{1}{d\delta_k^{-2} + 4}\right].$$

□

We first note that assumption $\xi_k = \lambda\delta_k$ is only used to simplify the formulas. Theorem 18 highlights the impact of δ_k. If δ_k decays slowly, the product quickly becomes small, and the first term dominates. On the other hand, if δ_k decays fast, the product may converge to a nonzero value. By (Little et al., 2022, Theorem 2.2.2), the infinite product $\prod_{k \geqslant 0} \left[1 - \frac{1}{d\delta_k^{-2}+4}\right]$ diverges to zero [3] if and only if the series $\sum_{k \geqslant 0} [d\delta_k^{-2} + 4]^{-1}$ diverges. The latter, in turn, is equivalent to the divergence of the series $\sum_{k \geqslant 0} \delta_k^2$.

For instance, consider a sequence $\delta_k = \delta_0(1 + k)^{-\theta}$ with starting value $\delta_0 > 0$ and decay parameter $\theta > 0$. Consequently, if $\theta > 1/2$, the product does not diverge to zero. Yet, Theorem 18 only provides an upper bound for the rate and does not imply that $f_0(\beta_0^*)$ is not the limit of $f_{\varepsilon_k}(\beta^k)$. When $\theta \leqslant 1/2$, the product diverges to zero. However, even if the product vanishes quickly, the right-hand side is proportional to δ_k, which decays sublinearly. In general, we are able to establish the following sublinear convergence rate.

[3] We follow the standard denomination from the theory of infinite products that treats zero as a special case since the product *diverges to zero* if and only if the series $\sum_{n=1}^{\infty} \log(a_n)$ diverges to $-\infty$.

Theorem 19 *Consider the sequences $\xi_k = \lambda\delta_k$ and $\delta_k = \delta_0(1+k)^{-\theta}$ with starting value $\delta_0 > 0$ and decay parameter and $0 < \theta \leqslant 1/2$. For $K \geqslant 1$ and $\gamma > 1$, the sequence β^k generated by Algorithm 1 admits the following decay*

$$f_{\varepsilon_{K+1}}(\beta^{K+1}) - f_0(\beta_0^*) \leqslant \max\left\{c\delta_0, e^{\delta_0^2/d}(6 + 4d^{-1}\delta_0^2)^{\delta_0^2/d}(f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*))\right\} K^{-\frac{\theta\delta_0^2}{d\theta+\delta_0^2}}.$$

where constants c and d are defined in Equation (28).

Proof: Let $q = d^{-1}\delta_0^2$ and $0 < \nu < 1$, whose precise value of which will be determined later. Then, an application of Theorem 18 gives

$$f_{\varepsilon_{K+1}}(\beta^{K+1}) - f_0(\beta_0^*) \leqslant \max\left\{ c\delta_0(\lfloor K^\nu\rfloor + 1)^{-\theta}, (f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*)) \prod_{k=\lfloor K^\nu\rfloor}^{K}\left[1 - \frac{1}{q^{-1}(k+1)^{2\theta} + 4}\right]\right\}.$$

The first term is already of the desired form since

$$c\delta_0(\lfloor K^\nu\rfloor + 1)^{-\theta} \leqslant c\delta_0(K^\nu - 1 + 1)^{-\theta} = c\delta_0 K^{-\nu\theta}.$$

Hence, we look at the second term and estimate the product. Note that this product is increasing in θ, and we can bound it by

$$\begin{aligned}\prod_{k=\lfloor K^\nu\rfloor}^{K}\left[1 - \frac{1}{q^{-1}(k+1)^{2\theta} + 4}\right] &\leqslant \prod_{k=\lfloor K^\nu\rfloor}^{K}\left[1 - \frac{1}{q^{-1}(k+1) + 4}\right] \\ = \prod_{k=\lfloor K^\nu\rfloor}^{K}\left[1 - \frac{q}{k+1+4q}\right] &\leqslant \prod_{k=\lfloor K^\nu\rfloor}^{K}\left[1 - \frac{q}{k+1+4\lceil q\rceil}\right] = \prod_{k=\lfloor K^\nu\rfloor+1+4\lceil q\rceil}^{K+1+4\lceil q\rceil}\left[1 - \frac{q}{k}\right].\end{aligned}$$

Let $N = K + 4\lceil q\rceil$, $P = \lfloor K^\nu\rfloor + 4\lceil q\rceil$, $t = \lfloor q\rfloor$ and $r = q - t$. Then, the product can be expressed in terms of gamma function Γ,

$$\begin{aligned}\prod_{k=P+1}^{N+1}\left[1 - \frac{q}{k}\right] = \prod_{k=P+1}^{N+1}\frac{k-q}{k} &= \frac{(N+1-q)\cdot\ldots\cdot(P+1-q)}{(N+1)\cdot\ldots\cdot(P+1)}\cdot\frac{\Gamma(P+1-q)}{\Gamma(P+1-q)}\cdot\frac{P!}{P!} \\ &= \frac{\Gamma(N+2-q)\,P!}{(N+1)!\,\Gamma(P+1-q)} = \frac{\Gamma(N-t+1+1-r)\,P!}{(N+1)!\,\Gamma(P-t+1-r)}.\end{aligned}$$

Note that all factors here are strictly positive since

$$(N+1-q) \geqslant \ldots \geqslant (P+1-q) = \lfloor K^\nu\rfloor + 4\lceil q\rceil + 1 - q \geqslant \lfloor K^\nu\rfloor + 3\lceil q\rceil + 1 \geqslant 2 > 0.$$

The next step is to bound the Gamma functions using Gautschi's double inequality, e.g., see Gautschi (1959); Qi (2010). This leads to

$$\Gamma(j+1)(j+1)^{-(1-\alpha)} \leqslant \Gamma(j+\alpha) \leqslant \Gamma(j+1)j^{-(1-\alpha)}, \quad j \in \mathbb{N},\ 0 \leqslant \alpha \leqslant 1.$$

Its application for $j = N - t + 1$ and $j = P - t$ with $\alpha = 1 - r$ gives

$$\begin{aligned}\frac{\Gamma(N-t+1+1-r)\; P!}{(N+1)!\;\Gamma(P-t+1-r)} &\leqslant \frac{\Gamma(N-t+2)\;(N-t+1)^{-r}\;P!}{(N+1)!\Gamma(P-t+1)\;(P-t+1)^{-r}}\\ &= \frac{(N-t+1)!\;P!\;(P-t+1)^{r}}{(N+1)!\;(P-t)!\;(N-t+1)^{r}}.\end{aligned}$$

Recall that for the binomial coefficient $\binom{j}{k} := \frac{j!}{k!(j-k)!}$, the bound $\frac{j^k}{k^k} \leqslant \binom{j}{k} \leqslant \frac{e^k j^k}{k^k}$ holds. In our case, this leads to

$$\frac{(N-t+1)!\;P!\;(P-t+1)^{r}}{(N+1)!\;(P-t)!\;(N-t+1)^{r}} = \frac{\binom{P}{t}\;(P-t+1)^{r}}{\binom{N+1}{t}\;(N-t+1)^{r}} \leqslant \frac{e^t P^t (P-t+1)^r}{(N+1)^t (N-t+1)^r}.$$

To simplify the resulting fraction, we note that $e^t \leqslant e^{r+t}$, $P^t \leqslant (P+1)^t$, $(P-t+1)^r \leqslant (P+1)^r$ and that

$$N+1 \geqslant N+1-t = K + 4\lceil q\rceil + 1 - \lfloor q\rfloor \geqslant K + 3q + 1 \geqslant K.$$

Consequently, these estimates yield

$$\begin{aligned}\frac{e^t P^t (P-t+1)^r}{(N+1)^t (N-t+1)^r} &\leqslant \frac{e^{t+r}(P+1)^{t+r}}{K^{t+r}} = \frac{e^q (P+1)^q}{K^q} = \frac{e^q(\lfloor K^\nu\rfloor + 4\lceil q\rceil + 1)^q}{K^q}\\ &\leqslant e^q(2+4\lceil q\rceil)^q K^{\nu q} K^{-q} \leqslant e^q(6+4q)^q K^{-(1-\nu)q}.\end{aligned}$$

Combining everything together, we arrive at

$$f_{\varepsilon_{k+1}}(\beta^{K+1}) - f_0(\beta_0^*) \leqslant \max\left\{c\delta_0 K^{-\nu\theta}, e^q(6+4q)^q(f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*))K^{-(1-\nu)q}\right\}.$$

The last step is to select ν such that the powers $\nu\theta$ and $(1-\nu)q$ coincide. This gives $\nu = q/(\theta + q)$, which implies that $0 < \nu < 1$ and

$$K^{-(1-\nu)q} = K^{-\nu\theta} = K^{-\theta q/(\theta+q)}.$$

Substituting $q = d^{-1}\delta_0^2$ concludes the proof. □

Theorem 19 derives a sublinear convergence rate, which is slightly worse than $K^{-\theta}$, which is the decay rate of δ_k. As the power $\theta\delta_0^2/(d\theta + \delta_0^2)$ is increasing as a function of δ_0, $K^{-\theta q/(\theta+q)}$ will converge asymptotically to $K^{-\theta}$ as $\delta_0 \to +\infty$. However, in this case, the constant $e^{\delta_0^2/d}(6 + 4d^{-1}\delta_0^2)^{\delta_0^2/d}$ blows up at a much faster pace. In any case, it is important to understand the rate given by the previous theorem. Indeed, the maximum of the exponent $-\theta q/(\theta + q)$ is given by

$$\max_{q,\theta} \frac{\theta q}{\theta + q} = \max_{q,\theta} \frac{(\theta+q)q}{\theta+q} - \frac{q^2}{\theta+q} = \max_{q,\theta} q - \frac{q^2}{\theta+q}. \tag{29}$$

Since $\theta \in (0, 1/2]$. the maximum above, as a function of θ, is attained when $\theta = 1/2$. Hence, this yields

$$\max_q q - \frac{q^2}{1/2+q} = \max_q \frac{1}{2}\frac{q}{q+1/2}. \tag{30}$$

When $q \to \infty$, the maximum of the expression above is $1/2$. Therefore, $K^{-1/2}$ is, in principle, the best possible rate that one can obtain by using this technique.

Now, we consider a construction of δ_k that, to a certain extent, optimizes the constant involved in the previous theorem. We choose a sequence δ_k that makes the rate established in Theorem 19 independently from $d = \frac{2\gamma f_0^2(\beta_0^*)}{\lambda^2(\gamma-1)^2(p+1)}$.

Corollary 20 *Consider the sequences* $\xi_k = \lambda\delta_k$ *and*

$$\delta_k = \frac{2f_{\varepsilon_{k-1}}(\beta^k)}{\lambda\sqrt{p+1}(1+k)^\theta}, \qquad k \geqslant 0,$$

where, for convenience, $\delta_{-1} > 0$ *and* $0 < \theta \leqslant 1/2$. *For* $K \geqslant 1$, *the sequence* β^k *generated by Algorithm 1 admits*

$$f_{\varepsilon_{k+1}}(\beta^{K+1}) - f_0(\beta_0^*) \leqslant \max\left\{4\sqrt{p+1}f_{\varepsilon_0}(\beta^0), 10e(f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*))\right\} K^{-\frac{\theta}{\theta+1}}.$$

Proof: The proof is similar to the proof of Theorem 19 with only minor changes. An application of Theorem 18 with $\gamma = 2$ and any $0 < \nu < 1$ gives

$$f_{\varepsilon_{k+1}}(\beta^{K+1}) - f_0(\beta_0^*) \leqslant \max\Bigg\{\frac{2cf_{\varepsilon_{\lfloor K^\nu\rfloor-1}}(\beta^{\lfloor K^\nu\rfloor})}{\lambda\sqrt{2p+2}}(\lfloor K^\nu\rfloor+1)^{-\theta},$$
$$(f_{\varepsilon_0}(\beta^0) - f_0(\beta_0^*))\prod_{k=\lfloor K^\nu\rfloor}^{K}\left[1 - \frac{1}{2^{-2}df_{\varepsilon_{k-1}}^{-2}(\beta^k)\lambda^2(2p+2)(k+1)^{2\theta}+4}\right]\Bigg\},$$

where c and d are defined in (28). The first term is again bounded by

$$\frac{2cf_{\varepsilon_{\lfloor K^\nu\rfloor-1}}(\beta^{\lfloor K^\nu\rfloor})}{\lambda\sqrt{p+1}}(\lfloor K^\nu\rfloor+1)^{-\theta} \leqslant 4\sqrt{p+1}f_{\varepsilon_0}(\beta^0)(K^\nu-1+1)^{-\theta} = 4\sqrt{p+1}f_{\varepsilon_0}(\beta^0)K^{-\nu\theta}.$$

For the second term, we observe that

$$2^{-2}df_{\varepsilon_{k-1}}^{-2}(\beta^k)\lambda^2(2p+2) = \frac{f_0^2(\beta_0^*)}{f_{\varepsilon_{k-1}}^2(\beta^k)} \leqslant 1, \quad \text{for all } k \geqslant 0.$$

Using this bound for the product, the rest of the proof repeats the steps from Theorem 19 with $q = 1$ and $\nu = 1/(1+\theta)$. □

Remark 21 *The proof above can be carried out identically for the monotone sequence* $\delta_k = \frac{2\min_{s=0,\dots,k} f_0(\beta^s)}{\lambda\sqrt{p+1}(1+k)^\theta}$ *instead of* $\delta_k = \frac{2f_{\varepsilon_{k-1}}(\beta^k)}{\lambda\sqrt{p+1}(1+k)^\theta}$ *which yields Theorem 6.*

Without any additional assumptions, deriving a global linear convergence rate result for IRLS, aimed at minimizing the loss $f_\varepsilon(\beta)$, appears to be unattainable, both in cases with a fixed ε and with a decaying ε, as we expect that similar lower bounds as the one by Chizat (2022) can be obtained. However, it is possible to establish such convergence rate for a fixed ε by choosing, instead of the Huber function $j_\alpha(x)$, the smoothing function $\sqrt{|x|^2+\alpha^2}$, which leads to the objective

$$\hat{f}_\varepsilon(\beta) := \sqrt{\|X\beta - y\|_2^2 + \xi^2} + \lambda\sum_{j=1}^{p}\sqrt{|\beta_j|^2+\delta^2}.$$

In this context, we outline a concise explanation for establishing the linear convergence rate by utilizing the KL property. See Attouch et al. (2010); Yu et al. (2022).

Definition 22 (Kurdyka-Łojasiewicz property) *We say that a proper closed function $h : \mathbb{X} \to \mathbb{R} \cup \{\infty\}$ satisfies the Kurdyka-Łojasiewicz (KL) property at $\hat{\beta} \in \operatorname{dom} \partial h$ if there are $a \in (0, \infty]$, a neighborhood V of $\hat{\beta}$ and a continuous concave function $\varphi : [0, a) \to [0, \infty)$ with $\varphi(0) = 0$ such that*

1. *φ is continuously differentiable on $(0, a)$ with $\varphi' > 0$ on $(0, a)$;*
2. *For any $\beta \in V$ with $h(\hat{\beta}) < h(\beta) < h(\hat{\beta}) + a$, it holds that*

$$\varphi'(h(\beta) - h(\hat{\beta}))\operatorname{dist}(0, \partial h(\beta)) \geqslant 1. \tag{31}$$

If h satisfies the KL property at $\hat{\beta} \in \operatorname{dom} \partial h$ and the $\varphi(s)$ in (31) can be chosen as $\bar{c}\, s^{1-\alpha}$ for some $\bar{c} > 0$ and $\alpha \in [0, 1)$, then we say that h satisfies the KL property at $\hat{\beta}$ with exponent α.

A proper closed function h satisfying the KL property at every point in $\operatorname{dom} \partial h$ is said to be a KL function, and a proper closed function h satisfying the KL property with exponent $\alpha \in [0, 1)$ at every point in $\operatorname{dom} \partial h$ is said to be a KL function with exponent α.

Firstly, note that the results of Theorem 5 remain true for IRLS applied to $\hat{f}_\varepsilon$. Next, we establish a local linear convergence rate of IRLS based on (Bolte and Pauwels, 2016, Proposition 4). For that, we need three conditions to be satisfied:

1. $\hat{f}_\varepsilon(\beta^{k+1}) - \hat{f}_\varepsilon(\beta^k) \leqslant -C_1\|\beta^{k+1} - \beta^k\|_2$ for some constant $C_1 > 0$;
2. $\|\nabla \hat{f}_\varepsilon(\beta^k)\|_2 \leqslant C_2\|\beta^{k+1} - \beta^k\|_2$ for some constant $C_2 > 0$;
3. $\hat{f}_\varepsilon$ satisfies KL property with exponent $1/2$ or less; see Lemma 2.2 of Yu et al. (2022).

The first property follows similarly to Theorem 17. The second property follows from the $1/\alpha$-smoothness of $\sqrt{|x|^2 + \alpha^2}$ (see (Beck, 2017, Example 10.44)) combined with Karush-Kuhn-Tucker conditions for the constrained problem described in Equation (9). Lastly, unlike the scaled Huber function $j_\alpha(x)$, the function $\sqrt{|x|^2 + \alpha^2}$ can be represented via linear matrix inequalities Yu et al. (2022), and, by (Yu et al., 2022, Theorem 4.3), the loss $\hat{f}_\varepsilon$ has KL exponent $1/2$. Thus, by (Beck and Shtern, 2017, Lemma 2.5), the KL constant $\bar{c} > 0$ can be estimated and, therefore, by following an argument similar to the one developed in Bolte et al. (2017), it is possible to show that there exists a neighborhood of β^*_ε, estimated via $\bar{c} > 0$, in which IRLS admits a linear convergence rate. Then, by selecting ε in Theorem 5 appropriately, the IRLS algorithm designed to minimize the function $\hat{f}_\varepsilon$ admits a linear convergence rate to this neighborhood, which implies that IRLS has a global linear convergence rate to a solution of the smoothed square-root LASSO problem. However, this solution potentially differs from the solution of the true square-root LASSO objective function. Moreover, this proof will not remain true for a decaying sequence ε_k as the smoothness of $\hat{f}_\varepsilon$ deteriorates as ε_k vanishes and, therefore, the second property no longer holds. We leave the characterization of the set of smoothed versions of the square-root LASSO for future investigation, for which the technique above can be applied.

Appendix D. Proofs for Section 6

Before we start the convergence analysis, we state a few facts connected to the NSP. The first one is an equivalent formulation, which is more suitable for our analysis.

Lemma 23 *(Foucart and Rauhut, 2013, Lemma 4.20) The matrix $X \in \mathbb{R}^{n\times p}$ satisfies the robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ if and only if for any set $S \subset [p]$ of cardinality $|S| \leqslant s$ we have*

$$\|z-\beta\|_1 \leqslant \frac{1+\rho}{1-\rho}(\|z\|_1 - \|\beta\|_1 + 2\|\beta_{S^c}\|_1) + \frac{2\tau}{1-\rho}\|X(z-\beta)\|_2 \tag{32}$$

for all vectors $\beta, z \in \mathbb{R}^p$.

The first benefit of NSP, when used in the convergence analysis, is that it is possible to track the distance to the ground truth signal β_* in terms of the function value gap.

Lemma 24 (Error bound in terms of function value gap) *Let $X \in \mathbb{R}^{n\times p}$ admit the robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ of order s. If $\lambda \leqslant \frac{(1+\rho)}{2\tau}$, then for all $z \in \mathbb{R}^n$ we have*

$$\|z-\beta_*\|_1 \leqslant \frac{2(1+\rho)}{(1-\rho)\lambda}\left[\lambda\sigma_s(\beta_*)_{\ell_1} + \|e\|_2 + \frac{1}{2}(f_0(z) - f_0(\beta_*))\right],$$

where β_* is the ground truth signal that gives origin to the data $y = X\beta_* + e$.

Proof: By Theorem 23 and the choice of λ, we get

$$\begin{aligned}\|z-\beta_*\|_1 &\leqslant \frac{1+\rho}{1-\rho}(\|z\|_1 - \|\beta_*\|_1 + 2\sigma_s(\beta_*)_{\ell_1}) + \frac{2\tau}{1-\rho}\|X(z-\beta_*)\|_2 \\ &\leqslant \frac{1+\rho}{(1-\rho)\lambda}\left[\lambda(\|z\|_1 - \|\beta_*\|_1 + 2\sigma_s(\beta_*)_{\ell_1}) + \|X(z-\beta_*)\|_2\right]\end{aligned}$$

Since β_* is the true signal, we have $X\beta_* = y - e$ and $\|e\|_2 = \|X\beta_* - y\|_2$. Hence, triangle inequality gives

$$\begin{aligned}\|z-\beta_*\|_1 &\leqslant \frac{1+\rho}{(1-\rho)\lambda}\left[2\lambda\sigma_s(\beta_*)_{\ell_1} + \lambda\|z\|_1 + \|Xz-y\|_2 + \|e\|_2 - \lambda\|\beta_*\|_1\right] \\ &= \frac{1+\rho}{(1-\rho)\lambda}\left[2\lambda\sigma_s(\beta_*)_{\ell_1} + 2\|e\|_2 + f_0(z) - f_0(\beta_*)\right].\end{aligned}$$

□

For the iterates of IRLS, a consequence of Theorem 24 is the following statement:

Corollary 25 *Under assumptions of Theorem 24 for all $\xi, \delta \geqslant 0$, the iterates β^k generated by Algorithm 1 admit*

$$\|\beta^k-\beta_*\|_1 \leqslant \frac{2(1+\rho)}{(1-\rho)\lambda}\left[\lambda\sigma_s(\beta_*)_{\ell_1} + \|e\|_2 + \frac{1}{2}(f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*))\right].$$

Proof: It follows Theorem 24 combined with inequalities $f_0(\beta_*) \geqslant f_0(\beta_0^*)$ and $f_0(\beta^k) \leqslant f_{\varepsilon_k}(\beta^k)$. □

As a consequence, the minimization of the objective function (3) implies that the distance between the iterates and the ground truth is being minimized. Furthermore, the bounds established in Section 4 and Section 5 can be combined with Theorem 25 to derive the convergence results in terms of the distance to the ground truth. A similar inequality can be derived for the minimizer of the square-root LASSO, Equation (3).

Corollary 26 *Under assumptions of Theorem 24 solution β_0^* of square-root LASSO admits*

$$\|\beta_0^* - \beta_*\|_1 \leqslant \frac{2(1+\rho)}{(1-\rho)\lambda}\left[\lambda\sigma_s(\beta_*)_{\ell_1} + \|e\|_2\right]$$

and

$$\|X\beta_0^* - y\|_2 \leqslant \left[1 + \frac{2(1+\rho)}{(1-\rho)}\right]\|e\|_2 + \frac{2(1+\rho)\lambda}{(1-\rho)}\sigma_s(\beta_*)_{\ell_1}.$$

Proof: The first inequality follows from Theorem 24 with $z = \beta_0^*$. For the second inequality, we use the optimality of β_0^*,

$$\|X\beta_0^* - y\|_2 + \lambda\|\beta_0^*\|_1 \leqslant \|X\beta_* - y\|_2 + \lambda\|\beta_*\|_1 = \|e\|_2 + \lambda\|\beta_*\|_1. \tag{33}$$

By bringing $\|\beta_0^*\|_1$ to the right-hand side and applying the first inequality, we get

$$\begin{aligned}\|X\beta_0^* - y\|_2 &\leqslant \|e\|_2 + \lambda(\|\beta_*\|_1 - \|\beta_0^*\|_1) \leqslant \|e\|_2 + \lambda\|\beta_* - \beta_0^*\|_1\\ &\leqslant \|e\|_2 + \frac{2(1+\rho)}{1-\rho}\left[\lambda\sigma_s(\beta_*)_{\ell_1} + \|e\|_2\right]\\ &= \left[1 + \frac{2(1+\rho)}{1-\rho}\right]\|e\|_2 + \frac{2(1+\rho)\lambda}{1-\rho}\sigma_s(\beta_*)_{\ell_1}.\end{aligned}$$

□ The first bound was proven in (Petersen and Jung, 2021, Theorem 3.1). We re-derive it for the sake of completeness[4]. It implies that if the noise is absent and β is sparse, square-root LASSO recovers β uniquely. The second bound is rather a technical result, which will be useful later in this section. We can now proceed to formally state the main theorem of this work.

Theorem 27 *Let X satisfies NSP with constants $0 < \rho < \frac{1}{4}$ and $\tau > 0$ and assume that $y = X\beta_* + e$. Consider the sequence $\delta_k = \min\left\{\delta_{k-1}, \frac{\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}}{\lambda(p+1)}\right\}$ and $\xi_k = \lambda\delta_k$ for $k \geqslant 0$ and $\delta_{-1} = +\infty$. Then, for $k \leqslant \hat{k} := \min\left\{k \in \mathbb{N} : f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) > 3\lambda(p+1)\delta_k/4\right\}$ it holds that the following is true for the iterates β^k of Algorithm 1 with $\lambda \leqslant \rho/\tau$:*

$$f_{\varepsilon_{k+1}}(\beta^{k+1}) - f_0(\beta_0^*) \leqslant \left[1 - \frac{(1-\rho)^4}{96(1+\rho)^2(2+\rho)^2(p+1)}\right]\left[f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)\right].$$

Moreover, for $0 \leqslant f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) \leqslant 3\lambda(p+1)\delta_k/4$, it holds that

$$\|\beta^k - \beta_*\|_1 \leqslant \frac{2(1+\rho)}{(1-\rho)\lambda}\left[1 + \frac{3(1+\rho)}{2-8\rho}\right]\left[\lambda\sigma_s(\beta_*)_{\ell_1} + \|e\|_2\right].$$

The proof consists of two complementary parts. In the first part, we will establish that outside a certain region, i.e., when $f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) > C\delta_k$ for a certain $C > 0$, we obtain a linear decay on the function value. Then, in the second part, we will prove that when the basin of attraction is reached, the iterates β^k are already close enough to the ground truth of the sparse linear regression problem.

The proof of Theorem 9 is based on Theorem 11 with $\beta = \beta_0^*$, $v^k = \beta_0^* - \beta^k$ and $\tilde{v}^k = (0, v^k) = \tilde{\beta}_0^* - \tilde{\beta}^k$. Let us denote by S the support of the k largest entries of β^k in absolute value. The first part consists of three main steps:

[4] The original theorem was stated with the condition $\lambda \geqslant \frac{2}{1+\rho}\tau$. Here, for our convenience, we performed the change of variable $\lambda \mapsto \frac{1}{\lambda}$ and state it under the assumption $\lambda \leqslant \frac{(1+\rho)}{2\tau}$.

1. Bound the first order term $-|\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{\beta}_0^* - \tilde{\beta}^k \rangle|$ from below.
2. Bound the second order term $\langle W_\varepsilon(\tilde{\beta}^k)\tilde{v}^k, \tilde{v}^k \rangle$ from above.
3. Show that the function value gap $f_{\varepsilon_k}(\beta^{k+1}) - f_{\varepsilon_k}(\beta^k)$ is bounded by δ_k.
4. To finish, by using a suitable choice of δ_k together with the three bounds above, the convergence rate will be finally obtained.

Proof: The proof will be divided into two parts. First, assume that $f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) > 3\lambda(p+1)\delta_k/4$ holds. In this case, the idea of the proof is based on Theorem 5 and Theorem 16.
Part I: Bounding the linear term: By assuming that $\xi_k = \lambda\delta_k$, the first-order term can be rewritten as

$$\langle \nabla \tilde{f}_{\varepsilon_k}(\beta^k), \tilde{v}^k \rangle = \frac{\langle \tilde{v}^k, \tilde{X}^T \tilde{X} \tilde{\beta}^k \rangle}{\max\{\|\tilde{X}\tilde{\beta}^k\|_2, \lambda\delta_k\}} + \lambda \sum_{i=1}^{p} \frac{v_i^k \beta_i^k}{\max\{|\beta_i^k|, \delta_k\}} = \lambda \sum_{i=0}^{p} \frac{\langle M_i \tilde{v}^k, M_i \tilde{\beta}^k \rangle}{\max\{\|M_i\tilde{\beta}^k\|_2, \delta_k\}},$$

with $M_0 = \lambda^{-1}\tilde{X}$ and $M_i = E^{i,i}$ for $i = 1, \ldots, p$, where $E^{i,i}$ is a matrix with a single non-zero entry $E_{i,i}^{i,i} = 1$. For a single summand, we have,

$$\begin{aligned}
\frac{\langle M_i \tilde{v}^k, M_i \tilde{\beta}^k \rangle}{\max\{\|M_i\tilde{\beta}^k\|_2, \delta_k\}} &= \frac{\langle M_i \tilde{\beta}_0^*, M_i \tilde{\beta}^k \rangle}{\max\{\|M_i\tilde{\beta}^k\|_2, \delta_k\}} - \frac{\|M_i\tilde{\beta}^k\|_2^2}{\max\{\|M_i\tilde{\beta}^k\|_2, \delta_k\}} \\
&\leqslant \frac{\|M_i \tilde{\beta}_0^*\|_2 \|M_i \tilde{\beta}^k\|_2}{\max\{\|M_i\tilde{\beta}^k\|_2, \delta_k\}} - \frac{\|M_i\tilde{\beta}^k\|_2^2}{\max\{\|M_i\tilde{\beta}^k\|_2, \delta_k\}} \\
&\leqslant \|M_i \tilde{\beta}_0^*\|_2 - \frac{\|M_i\tilde{\beta}^k\|_2^2}{\max\{\|M_i\tilde{\beta}^k\|_2, \delta_k\}}.
\end{aligned}$$

If $\|M_i\tilde{\beta}^k\|_2 \geqslant \delta_k$, then

$$\frac{\langle M_i \tilde{v}^k, M_i \tilde{\beta}^k \rangle}{\max\{\|M_i\tilde{\beta}^k\|_2, \delta_k\}} \leqslant \|M_i \tilde{\beta}_0^*\|_2 - \|M_i\tilde{\beta}^k\|_2 = \|M_i \tilde{\beta}_0^*\|_2 - j_{\delta_k}(\|M_i\tilde{\beta}^k\|_2).$$

Otherwise, if $\|M_i\tilde{\beta}^k\|_2 < \delta_k$, we have

$$\begin{aligned}
\frac{\langle M_i \tilde{v}^k, M_i \tilde{\beta}^k \rangle}{\max\{\|M_i\tilde{\beta}^k\|_2, \delta_k\}} &\leqslant \|M_{\tilde{\tilde{i}}} \beta_0^*\|_2 - \frac{\|M_i\tilde{\beta}^k\|_2^2}{\delta_k} \\
&= \|M_i \tilde{\beta}_0^*\|_2 - \frac{\|M_i\tilde{\beta}^k\|_2^2}{2\delta_k} - \frac{\delta_k}{2} + \frac{\delta_k}{2} = \|M_i \tilde{\beta}_0^*\|_2 - j_{\delta_k}(\|M_i\tilde{\beta}^k\|_2) + \frac{\delta_k}{2}.
\end{aligned}$$

Thus, in any case, the latter bound applies since $\frac{\delta_k}{2} > 0$. Hence, the first-order term can be bounded by

$$\begin{aligned}
\langle \nabla \tilde{f}_{\varepsilon_k}(\beta^k), \tilde{v}^k \rangle &\leqslant \lambda \sum_{i=0}^{p} \left[\|M_i \tilde{\beta}_0^*\|_2 - j_{\delta_k}(\|M_i\tilde{\beta}^k\|_2) + \tfrac{\delta_k}{2} \right] \\
&= \|\tilde{X}\tilde{\beta}_0^*\|_2 + \lambda\|\beta_0^*\|_1 - j_{\delta_{k,0}}(\|\tilde{X}\tilde{\beta}^k\|_2) + \lambda \sum_{i=1}^{p} j_{\delta_k}(\beta_j^k) + \tfrac{1}{2}\lambda(p+1)\delta_k \\
&= f_0(\beta_0^*) - f_{\varepsilon_k}(\beta^k) + \tfrac{1}{2}\lambda(p+1)\delta_k.
\end{aligned}$$

Now, by using the hypothesis $f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) > \frac{3}{4}\lambda(p+1)\delta_k$, it follows that

$$\langle \nabla \tilde{f}_{\varepsilon_k}(\beta^k), \tilde{v}^k \rangle \leqslant f_0(\beta_0^*) - f_{\varepsilon_k}(\beta^k) + \tfrac{1}{2}\lambda(p+1)\delta_k \leqslant -\tfrac{1}{4}\lambda(p+1)\delta_k \leqslant 0,$$

and

$$\begin{aligned}\langle \nabla \tilde{f}_{\varepsilon_k}(\beta^k), \tilde{v}^k \rangle &\leqslant f_0(\beta_0^*) - f_{\varepsilon_k}(\beta^k) + \frac{1}{2}\lambda(p+1)\delta_k \\ &\leqslant f_0(\beta_0^*) - f_{\varepsilon_k}(\beta^k) + \frac{2}{3}[f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)] \leqslant \tfrac{1}{3}[f_0(\beta_0^*) - f_{\varepsilon_k}(\beta^k)] \leqslant 0.\end{aligned}$$

Hence, for $\beta^k \neq \beta_0^*$ we apply Theorem 11, which gives

$$f_{\varepsilon_k}(\beta^{k+1}) - f_{\varepsilon_k}(\beta^k) \leqslant -\frac{|\langle \nabla \tilde{f}_{\varepsilon_k}(\tilde{\beta}^k), \tilde{v}^k \rangle|^2}{2\langle W_{\varepsilon_k}(\tilde{\beta}^k)\tilde{v}^k, \tilde{v}^k \rangle} \leqslant -\frac{\lambda(p+1)\delta_k(f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*))}{24\langle W_{\varepsilon_k}(\tilde{\beta}^k)v^k, v^k \rangle}. \tag{34}$$

Now, we need to bound the denominator $\langle W_{\varepsilon_k}(\tilde{\beta}^k)v^k, v^k \rangle$.

Part II: Bounding the quadratic term: From the definition of W_ε, we obtain

$$\begin{aligned}\langle W_\varepsilon(\tilde{\beta}^k)\tilde{v}^k, \tilde{v}^k \rangle &= \frac{\|\tilde{X}\tilde{v}^k\|_2^2}{\max\{\|\tilde{X}\tilde{\beta}^k\|_2, \lambda\delta_k\}} + \lambda \sum_{j=1}^{p} \frac{|v_j^k|^2}{\max\{|\beta_j^k|, \delta_k\}} \\ &\leqslant \frac{\|Xv^k\|_2^2}{\lambda\delta_k} + \frac{\lambda^2}{\lambda\delta_k}\|v^k\|_2^2 = \frac{1}{\lambda\delta_k}\left[\|Xv^k\|_2^2 + \lambda^2\|v^k\|_2^2\right] \\ &\leqslant \frac{1}{\lambda\delta_k}\left[\|Xv^k\|_2 + \lambda\|v^k\|_2\right]^2 \leqslant \frac{1}{\lambda\delta_k}\left[\|X(\beta_0^* - \beta^k)\|_2 + \lambda\|\beta_0^* - \beta^k\|_1\right]^2.\end{aligned}$$

Next, we bound the term $\|X(\beta_0^* - \beta^k)\|_2 + \lambda\|\beta_0^* - \beta^k\|_1$ by using the NSP. Theorem 23 yields

$$\begin{aligned}&\|X(\beta_0^* - \beta^k)\|_2 + \lambda\|\beta_0^* - \beta^k\|_1 \\ &\leqslant \frac{\lambda(1+\rho)}{1-\rho}\left[\|\beta_0^*\|_1 - \|\beta^k\|_1 + 2\sigma_s(\beta^k)_{\ell_1}\right] + \left[1 + \frac{2\tau\lambda}{1-\rho}\right]\|X(\beta_0^* - \beta^k)\|_2 \\ &\leqslant \frac{\lambda(1+\rho)}{1-\rho}\left[\|\beta_0^*\|_1 - \|\beta^k\|_1 + 2\sigma_s(\beta^k)_{\ell_1}\right] + \left[1 + \frac{2\tau\lambda}{1-\rho}\right](\|X\beta_0^* - y\|_2 + \|X\beta^k - y\|_2)\end{aligned}$$

Since β_0^* is the minimizer of f_0, we have

$$\|X\beta_0^* - y\|_2 + \lambda\|\beta_0^*\|_1 \leqslant \|X\beta^k - y\|_2 + \lambda\|\beta^k\|_1,$$

which is equivalent to

$$\lambda(\|\beta_0^*\|_1 - \|\beta^k\|_1) \leqslant \|X\beta^k - y\|_2 - \|X\beta_0^* - y\|_2.$$

Thus, we get

$$\begin{aligned}\|X(\beta_0^* - \beta^k)\|_2 + \lambda\|\beta_0^* - \beta^k\|_1 &\leqslant \frac{2\lambda(1+\rho)}{1-\rho}\sigma_s(\beta^k)_{\ell_1} \\ &+ \left[1 + \frac{2\tau\lambda}{1-\rho} + \frac{1+\rho}{1-\rho}\right]\|X\beta^k - y\|_2 + \left[1 + \frac{2\tau\lambda}{1-\rho} - \frac{1+\rho}{1-\rho}\right]\|X\beta_0^* - y\|_2.\end{aligned}$$

By assumption $\lambda \leqslant \rho/\tau$, which implies that

$$\left[1 + \frac{2\tau\lambda}{1-\rho} - \frac{1+\rho}{1-\rho}\right] \leqslant 0 \quad \text{and} \quad \left[1 + \frac{2\tau\lambda}{1-\rho} + \frac{1+\rho}{1-\rho}\right] \leqslant \frac{2(1+\rho)}{1-\rho}.$$

Therefore, we obtain

$$\|X(\beta_0^* - \beta^k)\|_2 + \lambda\|\beta_0^* - \beta^k\|_1 \leqslant \frac{2(1+\rho)}{1-\rho}\left[\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}\right],$$

and

$$\langle W_\varepsilon(\tilde{\beta}^k)\tilde{v}^k, \tilde{v}^k\rangle \leqslant \frac{4(1+\rho)^2}{\lambda\delta_k(1-\rho)^2}\left[\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}\right]^2. \tag{35}$$

The remaining major step of this part is to bound $\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}$ in terms of δ_k. Recall that, by hypothesis, $\delta_k = \min\left\{\delta_{k-1}, \frac{\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}}{\lambda(p+1)}\right\}$. If the minimum is attained by the second term, the bound is trivial since

$$\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1} = \lambda(p+1)\delta_k \leqslant \frac{2+\rho}{(1-\rho)}\lambda(p+1)\delta_k.$$

Otherwise, there exists index $j < k$, such that

$$\delta_k = \delta_j = \frac{\|X\beta^j - y\|_2 + \lambda\sigma_s(\beta^j)_{\ell_1}}{\lambda(p+1)}.$$

By using that $f_0(\beta) \leqslant f_\varepsilon(\beta) \leqslant f_0(\beta) + \xi + \lambda p\delta$ and $f_{\varepsilon_1}(\beta) \leqslant f_{\varepsilon_2}(\beta)$ whenever $0 \leqslant \xi_1 \leqslant \xi_2$, $0 \leqslant \delta_1 \leqslant \delta_2$ and by the construction of the iterates, we have

$$f_0(\beta^k) \leqslant f_{\varepsilon_k}(\beta^k) \leqslant f_{\varepsilon_j}(\beta^j) \leqslant f_0(\beta^j) + \lambda(p+1)\delta_j.$$

Expanding both the right- and left-most parts leads to

$$\|X\beta^k - y\|_2 + \lambda\|\beta^k\|_1 \leqslant 2\|X\beta^j - y\|_2 + \lambda\|\beta^j\|_1 + \lambda\sigma_s(\beta^j)_{\ell_1}$$

Let us denote by S_j the set of indices corresponding to the best-s term approximation of β^j. That is, we have $\|\beta^j_{S_j^c}\|_1 = \sigma_s(\beta^j)_{\ell_1}$ and $\|\beta^k_{S_j^c}\|_1 \geqslant \sigma_s(\beta^k)_{\ell_1}$. Thus, by splitting the norms $\|\beta^t\|_1 = \|\beta^t_{S_j}\|_1 + \|\beta^t_{S_j^c}\|_1$, for $t = j, k$, we arrive at

$$\|X\beta^k - y\|_2 + \lambda\|\beta^k_{S_j}\|_1 + \lambda\|\beta^k_{S_j^c}\|_1 \leqslant 2\|X\beta^j - y\|_2 + \lambda\|\beta^j_{S_j}\|_1 + 2\lambda\sigma_s(\beta^j)_{\ell_1}.$$

Hence, bringing $\lambda\|\beta^k_{S_j}\|_1$ to the right-hand side, yields

$$\begin{aligned}\|X\beta^k - y\|_2 + \lambda\|\beta^k_{S_j^c}\|_1 &\leqslant 2\left[\|X\beta^j - y\|_2 + \lambda\sigma_s(\beta^j)_{\ell_1}\right] + \lambda[\|\beta^j_{S_j}\|_1 - \|\beta^k_{S_j}\|_1] \\ &\leqslant 2\left[\|X\beta^j - y\|_2 + \lambda\sigma_s(\beta^j)_{\ell_1}\right] + \lambda\|(\beta^j - \beta^k)_{S_j}\|_1. \end{aligned} \tag{36}$$

Moreover, the definition of NSP gives

$$\begin{aligned}\|(\beta^j - \beta^k)_{S_j}\|_1 &\leqslant \rho\|(\beta^j - \beta^k)_{S_j^c}\|_1 + \tau\|X(\beta^j - \beta^k)\|_2 \\ &\leqslant \rho\|\beta^j_{S_j^c}\|_1 + \rho\|\beta^k_{S_j^c}\|_1 + \tau\|X\beta^j - y\|_2 + \tau\|X\beta^k - y\|_2 \\ &\leqslant \rho\sigma_s(\beta^j)_{\ell_1} + \rho\|\beta^k_{S_j^c}\|_1 + \tfrac{\rho}{\lambda}\|X\beta^j - y\|_2 + \tfrac{\rho}{\lambda}\|X\beta^k - y\|_2.\end{aligned}$$

Incorporating this bound in Equation (36) leads to

$$\|X\beta^k - y\|_2 + \lambda\|\beta^k_{S_j^c}\|_1 \leqslant (2+\rho)\left[\|X\beta^j - y\|_2 + \lambda\sigma_s(\beta^j)_{\ell_1}\right] + \rho\|X\beta^k - y\|_2 + \lambda\rho\|\beta^k_{S_j^c}\|_1,$$

which, in turn, is equivalent to

$$\|X\beta^k - y\|_2 + \lambda\|\beta^k_{S_j^c}\|_1 \leqslant \frac{2+\rho}{1-\rho}\left[\|X\beta^j - y\|_2 + \lambda\sigma_s(\beta^j)_{\ell_1}\right].$$

Since $\|\beta^k_{S_j^c}\|_1 \geqslant \sigma_s(\beta^k)_{\ell_1}$, we obtain

$$\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1} \leqslant \frac{2+\rho}{1-\rho}\left[\|X\beta^j - y\|_2 + \lambda\sigma_s(\beta^j)_{\ell_1}\right] = \frac{2+\rho}{(1-\rho)}\lambda(p+1)\delta_k.$$

Returning to the bound for the quadratic term (35), this gives

$$\langle W_\varepsilon(\tilde{\beta}^k)\tilde{v}^k, \tilde{v}^k\rangle \leqslant \frac{4(1+\rho)^2(2+\rho)^2\lambda(p+1)^2\delta_k}{(1-\rho)^4}. \tag{37}$$

Adding the pieces together: Now, the bound for the quadratic term can be combined with the inequality (34). This finally gives

$$f_{\varepsilon_k}(\beta^{k+1}) - f_{\varepsilon_k}(\beta^k) \leqslant -\frac{\lambda(p+1)\delta_k(1-\rho)^4}{96(1+\rho)^2(2+\rho)^2\lambda(p+1)^2\delta_k}\left[f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)\right].$$

As for the last step, we add and subtract $f_0(\beta_0^*)$ and rearrange all the terms, which gives

$$\begin{aligned} f_{\varepsilon_{k+1}}(\beta^{k+1}) - f_0(\beta_0^*) &\leqslant f_{\varepsilon_k}(\beta^{k+1}) - f_0(\beta_0^*) \\ &\leqslant \left[1 - \frac{(1-\rho)^4}{96(1+\rho)^2(2+\rho)^2(p+1)}\right]\left[f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)\right]. \end{aligned}$$

In the proximity of global minimum: Next, we assume that $0 \leqslant f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) \leqslant 3\lambda(p+1)\delta_k/4$ holds. Since $\lambda \leqslant \rho/\tau = 2\rho/2\tau \leqslant (1+\rho)/2\tau$, by Theorem 25, we have

$$\begin{aligned} \|\beta^k - \beta_*\|_1 &\leqslant \frac{2(1+\rho)}{(1-\rho)\lambda}\left[\lambda\sigma_s(\beta_*)_{\ell_1} + \|e\|_2 + \frac{1}{2}(f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*))\right] \\ &\leqslant \frac{2(1+\rho)}{(1-\rho)\lambda}\left[\lambda\sigma_s(\beta_*)_{\ell_1} + \|e\|_2 + \tfrac{3}{8}\lambda(p+1)\delta_k\right]. \end{aligned} \tag{38}$$

To finish the proof, we will establish a bound of the form

$$\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1} \leqslant c_1[\|e\|_2 + \lambda\sigma_s(\beta_*)_{\ell_1}],$$

for a given $c_1 \geqslant 0$. Since β_0^* is the minimizer of f_0, we have

$$f_0(\beta^k) \leqslant f_{\varepsilon_k}(\beta^k) \leqslant f_0(\beta_0^*) + \tfrac{3}{4}\lambda(p+1)\delta_k \leqslant f_0(\beta_*) + \tfrac{3}{4}[\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}].$$

Expanding f_0 and rearranging terms gives

$$(1 - \tfrac{3}{4})\|X\beta^k - y\|_2 + \lambda\|\beta^k\|_1 - \tfrac{3\lambda}{4}\sigma_s(\beta^k)_{\ell_1} \leqslant \|X\beta_* - y\|_2 + \lambda\|\beta_*\|_1 = \|e\|_2 + \lambda\|\beta_*\|_1.$$

In a similar way to what was done above, let us denote by S the set of indices corresponding to the best-s term approximation of β_*. That is, we have $\|(\beta_*)_{S^c}\|_1 = \sigma_s(\beta_*)_{\ell_1}$ and $\|\beta^k_{S^c}\|_1 \geqslant \sigma_s(\beta^k)_{\ell_1}$. Thus, by splitting the norms $\|v\|_1 = \|v_S\|_1 + \|v_{S^c}\|_1$ for $v = \beta^k$ and $v = \beta_*$, we arrive at

$$(1 - \tfrac{3}{4})\|X\beta^k - y\|_2 + \lambda\|\beta^k_S\|_1 + \lambda\|\beta^k_{S^c}\|_1 - \tfrac{3\lambda}{4}\sigma_s(\beta^k)_{\ell_1} \leqslant \|e\|_2 + \lambda\|(\beta_*)_S\|_1 + \lambda\sigma_s(\beta_*)_{\ell_1}.$$

This, together with reverse triangle inequality, yields

$$\begin{aligned}(1 - \tfrac{3}{4})\|X\beta^k - y\|_2 + \lambda\|\beta^k_{S^c}\|_1 - \tfrac{3\lambda}{4}\sigma_s(\beta^k)_{\ell_1} &\leqslant \|e\|_2 + \lambda[\|\beta_S\|_1 - \|\beta^k_S\|_1] + \lambda\sigma_s(\beta_*)_{\ell_1} \\ &\leqslant \|e\|_2 + \lambda[\|(\beta_* - \beta^k)_S\|_1] + \lambda\sigma_s(\beta_*)_{\ell_1}\end{aligned}$$

By the definition of the NSP, (8), we have

$$\begin{aligned}\|(\beta_* - \beta^k)_S\|_1 &\leqslant \rho\|(\beta_* - \beta^k)_{S^c}\|_1 + \tau\|X(\beta_* - \beta^k)\|_2 \\ &\leqslant \rho\|(\beta_*)_{S^c}\|_1 + \rho\|\beta^k_{S^c}\|_1 + \tau\|X\beta_* - y\|_2 + \tau\|X\beta^k - y\|_2 \\ &\leqslant \rho\sigma_s(\beta_*)_{\ell_1} + \rho\|\beta^k_{S^c}\|_1 + \tfrac{\rho}{\lambda}\|e\|_2 + \tfrac{\rho}{\lambda}\|X\beta^k - y\|_2.\end{aligned}$$

Thus,

$$(1 - \tfrac{3}{4} - \rho)\|X\beta^k - y\|_2 + (1 - \rho)\lambda\|\beta^k_{S^c}\|_1 - \tfrac{3\lambda}{4}\sigma_s(\beta^k)_{\ell_1} \leqslant (1 + \rho)[\|e\|_2 + \lambda\sigma_s(\beta_*)_{\ell_1}]$$

Since, by assumption, $1 - \frac{3}{4} - \rho > 0$ and $\|\beta^k_{S^c}\|_1 \geqslant \sigma_s(\beta^k)_{\ell_1}$, we get

$$\begin{aligned}\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1} &\leqslant \|X\beta^k - y\|_2 + \frac{1 - \rho}{1 - \frac{3}{4} - \rho}\lambda\|\beta^k_{S^c}\|_1 - \frac{\frac{3\lambda}{4}}{1 - \frac{3}{4} - \rho}\sigma_s(\beta^k)_{\ell_1} \\ &\leqslant \frac{1 + \rho}{1 - \frac{3}{4} - \rho}[\|e\|_2 + \lambda\sigma_s(\beta_*)_{\ell_1}].\end{aligned}$$

Thus, $\lambda(p + 1)\delta_k$ is bounded from above as

$$\lambda(p + 1)\delta_k \leqslant [\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}] \leqslant \frac{(1 + \rho)}{1 - \frac{3}{4} - \rho}[\|e\|_2 + \lambda\sigma_s(\beta_*)_{\ell_1}].$$

By applying this bound to (38), we finally conclude that

$$\|\beta^k - \beta_*\|_1 \leqslant \frac{2(1 + \rho)}{(1 - \rho)\lambda}\left[1 + \frac{3(1 + \rho)}{2 - 8\rho}\right][\lambda\sigma_s(\beta_*)_{\ell_1} + \|e\|_2]\,.$$

□

This concludes linear convergence rate proof.

Remark 28 *The null space constant $\rho < 1/4$ and the constant $\frac{3}{4}$ in $f_{\varepsilon_k}(\beta^k) - f_0(\beta^*_0) > 3\lambda(p + 1)\delta_k/4$ are not optimized. In fact, inspired by Aravkin et al. (2019), one could choose the smoothing parameter δ_k as $\delta_k = \min\left\{\delta_{k-1}, c\frac{\|X\beta^k - y\|_2 + \lambda\sigma_s(\beta^k)_{\ell_1}}{\lambda(p+1)}\right\}$ for a certain constant $0 < c < 2$. This constant would appear in the definition of the null space constant in the form $0 < \rho < 1 - \frac{3c}{4}$*

and in the hypothesis for the "basis of attraction" that would be given by $f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*) > \min\{1, c^{-1}\}\lambda(p+1)\delta_k$. *Hence, in this case, the convergence results read as*

$$f_{\varepsilon_{k+1}}(\beta^{k+1}) - f_0(\beta_0^*) \leqslant \left[1 - \frac{c^2(1-\rho)^4}{96(1+\rho)^2(1+c+\rho)^2(p+1)}\right]\left[f_{\varepsilon_k}(\beta^k) - f_0(\beta_0^*)\right],$$

and

$$||\beta^k - \beta_*||_1 \leqslant \frac{2(1+\rho)}{(1-\rho)\lambda}\left[1 + \frac{3c(1+\rho)}{8-6c-8\rho}\right][\lambda\sigma_s(\beta_*)_{\ell_1} + ||\eta||_2].$$

As shown in Theorem 26, the oracle inequality for the square-root LASSO is given by

$$\|\beta_0^* - \beta_*\|_1 \leqslant \frac{2(1+\rho)}{(1-\rho)\lambda}[\lambda\sigma_s(\beta_*)_{\ell_1} + \|e\|_2]$$

The second part of this theorem shows that IRLS achieves, by only assuming the compatibility condition, the same oracle inequality as the one above up to a constant factor. Finally, we derive Theorem 9 as a corollary to Theorem 27.

Proof: Since X fulfills the null space property with constants $\rho \leqslant 1/6$ and $\tau \leqslant 6/7$, by (13), the same property with larger constants $\rho = 1/6$ and $\tau = 7/6$ applies. By plugging those values into Theorem 27, we derive the condition $\lambda \leqslant 1/7$, and the result follows. □

Appendix E. Iteratively Reweighted Least Squares (IRLS) implementation

In order to describe the IRLS implementation, we first include the proof of Lemma 3, which shows that the minimization step stated in Equation (9) is equivalent to solving least squares.

Lemma 29 *The iterate* β^{k+1} *defined in* (9) *is the minimizer of the unconstrained least squares problem*

$$\min_{z\in\mathbb{R}^p} \frac{\|Xz - y\|_2^2}{\max\{\|X\beta^k - y\|_2, \xi\}} + \lambda\sum_{j=1}^p \frac{|z_j|^2}{\max\{|\beta_j^k|, \delta\}}.$$

Proof: **of Theorem 3** In view of (15), the minimizer of $Q_\varepsilon(\tilde{z}, \tilde{\beta}^k)$ is also the minimizer of

$$\begin{aligned}\langle \tilde{z}, W_\varepsilon(\tilde{\beta})\tilde{z}\rangle &= \frac{\|\tilde{X}\tilde{z}\|_2^2}{\max\{\|\tilde{X}\tilde{\beta}^k\|_2, \xi\}} + \lambda\sum_{j=1}^p \frac{|\tilde{z}_j|^2}{\max\{|\tilde{\beta}_j^k|, \delta\}} \\ &= \frac{\|Xz - \tilde{z}_0 y\|_2^2}{\max\{\|X\beta^k - \tilde{\beta}_0^k y\|_2, \xi\}} + \lambda\sum_{j=1}^p \frac{|z_j|^2}{\max\{|\beta_j^k|, \delta\}}.\end{aligned}$$

Reversing the change of variables from $\tilde{z}$ to z with the equalities $\tilde{z}_0 = \tilde{\beta}_0^k = 1$ gives the desired unconstrained least squares problem,

$$\begin{aligned}\underset{\tilde{z}\in\mathbb{R}^{p+1},\ \tilde{z}_0=1}{\arg\min}\ Q_\varepsilon(\tilde{z}, \tilde{\beta}^k) &= \underset{\tilde{z}\in\mathbb{R}^{p+1},\ \tilde{z}_0=1}{\arg\min} \frac{\|Xz - \tilde{z}_0 y\|_2^2}{\max\{\|X\beta^k - \tilde{\beta}_0^k y\|_2, \xi\}} + \lambda\sum_{j=1}^p \frac{|z_j|^2}{\max\{|\beta_j^k|, \delta\}} \\ &= \underset{z\in\mathbb{R}^p}{\arg\min} \frac{\|Xz - y\|_2^2}{\max\{\|X\beta^k - y\|_2, \xi\}} + \lambda\sum_{j=1}^p \frac{|z_j|^2}{\max\{|\beta_j^k|, \delta\}}.\end{aligned}$$

□

The resulting least squares problem can be further rewritten in the following way. Let us define

$$s_k := \Big(\max\{\|X\beta^k - y\|_2, \xi_k\}\Big)^{-1} \quad \text{and} \quad \sigma_j^k := \Big(\max\{|\beta_j^k|, \delta_k\}\Big)^{-1}.$$

Furthermore, let Σ_k be a diagonal matrix formed by vector σ^k. Then, the least squares problem in Theorem 3 can be rewritten as

$$\arg\min_{\beta} \left\| \begin{bmatrix} \sqrt{s_k} X \\ \sqrt{\lambda}\Sigma_k^{1/2} \end{bmatrix} \beta - \begin{bmatrix} \sqrt{s_k} y \\ 0 \end{bmatrix} \right\|_2^2$$

The corresponding solution β^{k+1} to the above problem is given by

$$\beta^{k+1} = (s_k X^T X + \lambda\Sigma_k)^{-1}\sqrt{s_k} X^T y = \sqrt{s_k}\lambda(\lambda^{-1} s_k X^T X + \Sigma_k)^{-1} X^T y, \tag{39}$$

where the inverse matrix is well-defined as Σ_k is positive definite. Standard solvers for equation of the form (39), e.g., LSQR solver for least squares developed by Paige and Saunders (1982), which is algebraically equivalent to the conjugate gradient method on the normal equations, require multiple multiplications of vectors with the matrix $\lambda^{-1} s_k X^T X + \Sigma_k$. This has a computational cost of $\mathcal{O}(np)$ operations, which can be costly for large p. Since square-root LASSO finds a sparse solution β_0^*, the number of nonzero entries in β^k is expected to be much smaller than p. In the following, we show that the computational cost can also be reduced to the number of active indices $J_k := \{j \in [p] : |\beta_j^k| > \delta_k\}$. For this, we employ the Sherman-Morrison-Woodbury formula.

Lemma 30 (Sherman and Morrison (1950)) *Let $B \in \mathbb{R}^{p\times p}$, $C \in \mathbb{R}^{m\times m}$ and $E, F \in \mathbb{R}^{p\times m}$. If B and C are invertible, then*

$$(ECF^T + B)^{-1} = B^{-1} - B^{-1}E(C^{-1} + F^T B^{-1} E)^{-1} F^T B^{-1}.$$

Set m as $\operatorname{rank}(X)$ and let $X^T X = U\Lambda U^T$ be the eigendecomposition of $X^T X$ with orthogonal $U \in \mathbb{R}^{p\times m}$ and diagonal $\Lambda \in \mathbb{R}^{m\times m}$ with positive diagonal entries. Then, we apply the Sherman-Morrison-Woodbury formula with $C = \lambda^{-1} s_k \Lambda$, $B = \Sigma_k$ and $E = F = U$. This yields

$$\beta^{k+1} = \sqrt{s_k}\lambda\left(\Sigma_k^{-1} - \Sigma_k^{-1} U(\lambda s_k^{-1}\Lambda^{-1} + U^T\Sigma_k^{-1}U)^{-1} U^T \Sigma_k^{-1}\right) X^T y, \tag{40}$$

Let us focus on $M := \lambda s_k^{-1}\Lambda^{-1} + U^T\Sigma_k^{-1}U$. By the definition of Σ_k, we get

$$(\Sigma_k^{-1})_{j,j} = \max\{|\beta_j^k|, \delta_k\} = \delta_k + \max\{|\beta_j^k| - \delta_k, 0\} =: \delta_k (I_p)_{j,j} + (D_k)_{j,j}, \quad j \in [p],$$

where I_p is the $p\times p$ identity matrix. Note that $D_{j,j} = 0$ whenever $j \notin J_k$. This leads to

$$\begin{aligned} M_k &= \lambda s_k^{-1}\Lambda^{-1} + U^T\Sigma_k^{-1}U = \lambda s_k^{-1}\Lambda^{-1} + \delta_k U^T I_p U + U^T D_k U \\ &= \lambda s_k^{-1}\Lambda^{-1} + \delta_k I_m + U_{J_k}^T (D_k)_{J_k} U_{J_k}, \end{aligned}$$

where $U_{J_k} \in \mathbb{R}^{|J_k|\times m}$ denotes the matrix formed by rows of U with indices in J_k and $(D_k)_{J_k} \in \mathbb{R}^{|J_k|\times|J_k|}$ is a diagonal matrix with nonzero entries of D_k.

Rewriting (40) in the form

$$U^T X^T y - s_k^{-1/2} \lambda^{-1} U^T \Sigma_k \beta^{k+1} = M_k^{-1} U^T \Sigma_k^{-1} X^T y.$$

gives again a system of the form (39), however, this time multiplication with M requires only $\mathcal{O}(m|J_k|)$ operations. As $m = \mathcal{O}(n)$ and $|J_k| = \mathcal{O}(s)$, the computational costs reduce significantly.

Thus, we summarize a fast IRLS implementation in Algorithm 2. We set $\xi_k = \lambda\delta_k$ in the numerical experiments. For the least squares, we again used the LSQR solver developed in Paige and Saunders (1982), which stopped after 100 iterations or if the relative residual is below 10^{-8}.

Algorithm 2 Fast IRLS Implementation

```
Input: Parameters $\lambda > 0$, $\xi_0, \delta_0 > 0$, initial guess $\beta^0$.
Precompute eigendecomposition of $X^T X = U^T \Lambda U$.
for $k = 1, 2 \ldots$ do
    Set $z_0 = U^T X^T y - s_k^{-1/2} \lambda^{-1} U^T \Sigma_k \beta^k$
    Find a solution $z$ of the linear system $M_k z = U^T \Sigma_k^{-1} X^T y$ with a least squares solver given an initial guess $z_0$.
    Compute $\beta^{k+1} = \sqrt{s_k} \lambda \Sigma_k^{-1} \left( X^T y - U z \right)$
    Construct $\xi_{k+1}$ and $\delta_{k+1}$ using the strategy of choice.
    Check stopping criteria
end for
return Sequence $\{\beta^k\}_{k \geqslant 0}$.
```

Appendix F. On the optimization procedures

In the following, we provide a more detailed description of the other optimization algorithms used for numerical trials in Section 7 and in the last part of this appendix. In particular, we consider the concominant LASSO by Ndiaye et al. (2017), proximal gradient and proximal Newton methods applied to the square-root LASSO objective function by Li et al. (2020) and the Information-Theoretic Exact Method (ITEM) by Taylor and Drori (2023), which in turn is equivalent to the Optimized Gradient Method developed by Kim and Fessler (2016) when the function to be minimized is convex but not strongly convex, as in this work and, finally, Frank-Wolfe algorithm adapter for square-root LASSO objective via epigraphic lifting Frank and Wolfe (1956); Harchaoui et al. (2012). All experiments are conducted on a laptop with Intel(R) Core(TM) i7-8550U CPU and 16 GB RAM.

F.1. Oracle

In this work, by the *oracle algorithm* in our experiments, we refer to a least squares solver, which is a priori restricted on the support of the ground truth. This means that if β is supported on S, and we denote by X_S the matrix formed by columns in S, then we find $\arg\min_{\beta \in \mathbb{R}^s} \|X_S \beta - y\|_2^2$ with a stable and efficient least-squares solver such as the LSQR developed by Paige and Saunders (1982).

F.2. Information-Theoretic Exact Method (ITEM)

The Information-Theoretic Exact Method (ITEM) was introduced by Taylor and Drori (2023). It is an optimal gradient method in the sense that it attains the lower bound on the oracle complexity for smooth strongly convex minimization problems, i.e., it is an accelerated optimal first-order method with the best possible constants to minimize the Lipschitz smooth (strongly) convex functions. As (3) does not satisfy the assumptions required by the paper, we instead minimize

$$\hat{f}_\varepsilon(\beta) := \sqrt{\|X\beta - y\|_2^2 + \xi^2} + \lambda \sum_{j=1}^{p} \sqrt{|\beta_j|^2 + \delta^2},$$

with $\delta = 10^{-4}$ and $\xi = \sqrt{\lambda}\delta$. The corresponding smoothness constant is then $L = \lambda\delta^{-1} + \xi^{-1}$. As $\hat{f}_\varepsilon$ is not strongly convex, the strong convexity constant μ is set to zero in Algorithm 3. With all the parameters, Algorithm 3 is performed until β^{k+1} satisfies the desired stopping criteria.

Algorithm 3 Information-Theoretic Exact Method (ITEM)

```
**Input:** Function $\hat{f}_\varepsilon$ with $0 \leqslant \mu \leqslant L < \infty$, initial guess $\beta^0$.
Initialize $z^{-1} = \beta^0 = \beta^0$, $\gamma_0 = 0$ and $q = \mu/L$.
**for** $k = 0, 1, 2 \dots$ **do**
    Set $\gamma_{k+1} = \frac{(1+q)\gamma_k + 2\left(1+\sqrt{(1+\gamma_k)(1+q\gamma_k)}\right)}{(1-q)^2}$
    Set $\alpha_k = \gamma_k/(1-q)\gamma_{k+1}$ and $\eta_k = \frac{1}{2}\frac{(1-q)^2\gamma_{k+1} - (1+q)\gamma_k}{1+q+q\gamma_k}$
    $z^k = (1-\alpha_k)\beta^k + \alpha_k\beta^k$
    $\beta^{k+1} = z^k - \frac{1}{L}\nabla\hat{f}_\varepsilon(z^k)$
    $\beta^{k+1} = (1-q\eta_k)z^k + q\eta_k\beta^k - \frac{\eta_k}{L}\nabla\hat{f}_\varepsilon(z^k)$
**end for**
**return** Sequence $\{\beta^k\}_{k\geqslant 0}$.
```

F.3. Proximal gradient descent

We consider an objective with smoothed data fidelity term,

$$\sqrt{\|X\beta - y\|_2^2 + \xi^2} + \lambda\|\beta\|_1 =: \check{f}_\xi(\beta) + \lambda\|\beta\|_1. \tag{41}$$

Given a previous iterate β^k, the proximal gradient method devised by Li et al. (2020) constructs β^{k+1} by minimizing the quadratic approximation of the above loss at point β_k,

$$Q_\lambda(\beta, \beta^k) = \check{f}_\xi(\beta^k) + \nabla\check{f}_\xi(\beta^k)^T(\beta - \beta^k) + \frac{L_k}{2}\|\beta - \beta^k\|_2^2 + \lambda\|\beta\|_1,$$

where L^k denotes the step size chosen via backtracking line search. The closed-form solution is

$$\beta^{k+1} = \mathcal{S}_{\lambda/L_k}(\beta^k - \nabla\check{f}_\xi(\beta)/L_k), \tag{42}$$

with $\mathcal{S}$ denoting the soft thresholding operator. The complete algorithm is described in Algorithm 4.

The performance of the proximal gradient depends on the smoothness constant $L = \xi^{-1}$ of $\check{f}$ and, consequently, ξ. If it is large, the function is smooth, but its global minimum is far from the solution of square-root LASSO. On the other hand, small ξ results in a large smoothness constant and small gradient steps. To reduce this dependency on ξ, it is initially set to 10^2, and once the algorithm converges, we decrease ξ by multiplying it with $10^{-1/4}$. Note that this is similar to the pathwise technique used in Li et al. (2020) and Ndiaye et al. (2017) for λ. In our experiments, the parameter λ remains constant in order not to change the solution of the initial square-root LASSO problem. It was also applied to ITEM but did not impact performance much.

Algorithm 4 Proximal gradient

```
Input: Parameters $\lambda > 0$, $\xi \geqslant 0$, $L_{\max} > 0$, initial guess $\beta^0$.
Initialize $L_0 = \tilde{L}_0 = L_{\max}$, $\beta = \beta^0$.
for $k = 0, 1, 2 \ldots$ do
    repeat
        Construct $\beta^{k+1}$ via (42) with $\tilde{L}_k$
        if $\check{f}_\xi(\beta^{k+1}) + \lambda\|\beta^{k+1}\|_1 < Q_\lambda(\beta^{k+1}, \beta^k)$ then
            $\tilde{L}_k = \tilde{L}_k/2$
        end if
    until $\check{f}_\xi(\beta^{k+1}) + \lambda\|\beta^{k+1}\|_1 \geqslant Q_\lambda(\beta^{k+1}, \beta^k)$
    $L_k = \min\{2\tilde{L}_k, L_{\max}\}$, $\tilde{L}_{k+1} = L_k$
    Construct $\beta^{k+1}$ via (42) with $L_k$ and check stopping criteria
end for
return Sequence $\{\beta^k\}_{k\geqslant 0}$.
```

Algorithm 5 Decreasing smoothing parameter

```
Input: Initial $\xi \geqslant 0$, initial guess $\beta^0$.
for $k = 0, 1, 2 \ldots$ do
    Set $L_{\max} = 10\xi_k^{-1}$ (if required)
    Construct $\beta^{k+1}$ by running the algorithm of the choice with $\xi_k$ and $\beta^{k-1}$ as initialization.
    Set $\xi_{k+1} = 10^{-1/4}\xi_k$ and check the stopping criteria.
end for
return Sequence $\{\beta^k\}_{k\geqslant 0}$.
```

In our experiments, for Algorithm 4 combined with Algorithm 5, we chose $\xi_0 = \sqrt{\lambda}$.

F.4. Proximal Newton method

Compared to the proximal gradient, the proximal Newton algorithm, also proposed by Li et al. (2020), includes second-order derivatives into quadratic approximation,

$$Q_\lambda^{New}(\beta, \beta^k) = \check{f}_\xi(\beta^k) + \nabla\check{f}_\xi(\beta^k)^T(\beta - \beta^k) + \tfrac{1}{2}(\beta - \beta^k)^T\nabla^2\check{f}_\xi(\beta^k)(\beta - \beta^k) + \lambda\|\beta\|_1.$$

Then, the new iterate of the proximal Newton algorithm is constructed as

$$\beta^{k+1/2} = \arg\min_\beta Q_\lambda^{New}(\beta, \beta^k), \quad \beta^{k+1} = \beta^k + \nu_k(\beta^{k+1/2} - \beta^k),$$

where the minimum is found via coordinate descent with safe screening, see Zhao et al. (2018), and ν_k is selected by backtracking line search. The procedure is summarized in Algorithm 6 and repeated until the desired stopping criterion is reached.

Algorithm 6 Proximal Newton

```
**Input:** Parameters $\lambda > 0$, $\xi_0 \geqslant 0$, initial guess $\beta^0$.
Initialize $\mu = 0.9$, $\alpha = 0.25$.
for $k = 0, 1, 2 \dots$ do
    Construct $\beta^{k+1/2} = \arg\min_\beta Q_\lambda^{New}(\beta, \beta^k)$ via coordinate descent.
    $\Delta^k = \beta^{k+1/2} - \beta^k$
    $\gamma_k = \nabla \check{f}_\xi(\beta^k)^T \Delta^k + \lambda(\|\beta^{k+1/2}\|_1 - \|\beta^k\|_1)$
    $\nu_k = 1$
    repeat
        $\nu_k = \mu\nu_k$
        $\beta^{k+1} = \beta^k + \nu_k \Delta^k$
    until $\check{f}_\xi(\beta^{k+1}) + \lambda\|\beta^{k+1}\|_1 \leqslant \check{f}_\xi(\beta^k) + \lambda\|\beta^k\|_1 + \alpha\nu_k\gamma_k$
end for
return Sequence $\{\beta^k\}_{k\geqslant 0}$.
```

Similarly, we combined Algorithm 6 with Algorithm 5. In the case $\lambda = \frac{1}{100}$ initial smoothing was set to $\sqrt{10\lambda}$ and for $\lambda = \frac{1}{7}$ to $\sqrt{\lambda/10}$.

F.5. Concomitant LASSO

The concomitant LASSO, developed by Ndiaye et al. (2017), considers an optimization problem

$$\operatorname*{arg\,min}_{\beta\in\mathbb{R}^p, \sigma\in\mathbb{R}} \frac{1}{2\sigma}\|X\beta - y\|_2^2 + \frac{\sigma}{2} + \lambda\|\beta\|_1 + \iota_{[\xi,+\infty)}(\sigma). \tag{43}$$

Here, $\xi \geqslant 0$ and $\iota_{[\xi,+\infty)}$ denotes the characteristic function of the set $[\xi, +\infty)$ which is $+\infty$ if $\sigma \notin [\xi, +\infty)$ and zero otherwise. The loss function described in (43) can be split into a convex $1/\xi$-smooth function $\frac{1}{2\sigma}\|X\beta - y\|_2^2 + \frac{\sigma}{2}$ and a separable $\lambda\|\beta\|_1 + \iota_{[\xi,+\infty)}(\sigma)$, and, therefore, it is optimized via a coordinate descent algorithm, where σ is updated after each sub-iteration on each coordinate.

The authors describe that they achieve computational efficiency by implementing safe screening rules that rule out unnecessary variables at each iteration as described in Algorithm 7. See (Ndiaye et al., 2017, Section 3) for more details.

Furthermore, Algorithm 7 is applied together with Algorithm 5. In our experiments, the initial values are $\xi_0 = \sqrt{10\lambda}$ for $\lambda = \frac{1}{100}$ and $\xi_0 = \sqrt{\lambda/10}$ for $\lambda = \frac{1}{7}$.

F.6. Frank-Wolfe

The Frank-Wolfe method is a fast first-order algorithm for minimizinga convex function over a convex set Frank and Wolfe (1956); Jaggi (2013); Lacoste-Julien and Jaggi (2015); Cherfaoui et al. (2019). In this section, we apply it for minimization of (41) using an epigraphical lifting technique similar to Harchaoui et al. (2012); Denoyelle et al. (2019); Jarret et al. (2022). First, we observe that the minimizer β^* of (41) admits

$$\lambda\|\beta^*\|_1 \leqslant \check{f}(\beta^*) + \lambda\|\beta^*\|_1 = \arg\min \check{f}(\beta) + \lambda\|\beta\|_1 \leqslant \check{f}(0) + \lambda \cdot 0 = \sqrt{\|y\|_2^2 + \xi^2}.$$

Algorithm 7 Concomitant LASSO

```
**Input:** Parameters $\lambda > 0$, $\xi > 0$, initial guess $\beta^0$.
Initialize $r = X\beta^0 - y$, $q = \|r\|_2$, $\sigma = \max\{\xi, q\}$.
Precompute the squared norms $n_j = \|X_j\|_2^2$ for $j = 1, \ldots, p$.
**for** $k = 1, 2 \ldots$ **do**
    $\beta^k = \beta^{k-1}$
    **if** $k \mod 5 = 1$ **then**
        Perform the safe screening to obtain index set $\mathcal{I} \subseteq [p]$.
    **end if**
    **for** $j \in \mathcal{I}$ **do**
        Set $X_j$ as a $j$-th column of $X$.
        $\Delta = X_j^T r$
        $\beta_j^k = \mathcal{S}_{\sigma\lambda/n_j}(\beta_j^k - \Delta)/n_j$
        $r = r + (\beta_j^k - \beta_j^{k-1})X_j$
        $q = \sqrt{q^2 + 2(\beta_j^k - \beta_j^{k-1})\Delta + (\beta_j^k - \beta_j^{k-1})^2 n_j}$
        $\sigma = \max\{\xi, q\}$
    **end for**
    Check stopping criteria
**end for**
**return** Sequence $\{\beta^k\}_{k \geqslant 0}$.
```

Consequently, the search space can restricted to ℓ_1-ball $\{\, \beta \in \mathbb{R}^p \;:\; \|\beta\|_1 \leqslant \rho\}$ with $\rho := \sqrt{\|y\|_2^2 + \xi^2}/\lambda$.

Then, the minimization of (41) can be rewritten as

$$\operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \check{f}(\beta) + \lambda\|\beta\|_1 = \operatorname*{arg\,min}_{(\beta,t) \in C} \check{f}(\beta) + \lambda t,$$

where

$$C := \{(\beta, t) \in \mathbb{R}^p \times \mathbb{R}_{\geqslant 0} \;:\; \|\beta\|_1 \leqslant t \leqslant \rho\}\,.$$

This step is known as the epigraphic lifting. As C is nonempty, closed, and convex, and $\check{f}$ is a proper, coercive, continuous, convex function, the Frank-Wolfe algorithm can be applied. Given an initial guess β^0 and $t^0 = \|\beta^0\|_1$, it constructs the iterates $\{(\beta^k, t^k)\}_{k \geqslant 0}$ in a descend-like fashion. More precisely, the descent direction is constructed by minimization of the first-order Taylor approximation of $\check{f}(\beta) + \lambda t$,

$$(z^k, r^k) \in \operatorname*{arg\,min}_{(\beta,t) \in C} \; \check{f}(\beta^k) + \langle \nabla \check{f}(\beta^k), \beta \rangle + \lambda t$$

As the above functional is linear in (β, t) and C is convex, the minimum is achieved in one of the extreme points of C, namely, $\{(0,0), (\pm\rho e_j, \rho),\ j \in [p]\}$ with $\{e_j\}_{j=1}^p$ denoting the standard basis. Consequently, with $q \in \arg\max_j |[\nabla \check{f}(\beta^k)]_j|$ the minimizer is given by

$$(z^k, r^k) = \begin{cases} (-\rho \operatorname{sgn}([\nabla \check{f}(\beta^k)]_q), \rho), & \text{if } \|\nabla \check{f}(\beta^k)\|_\infty > \lambda, \\ (0,0), & \text{otherwise.} \end{cases} \tag{44}$$

With (z^k, r^k) computed, the next iterates (β^{k+1}, t^{k+1}) are selected on the interval connecting (z^k, r^k) and (β^k, t^k) such that $\check{f}(\beta^{k+1}) + \lambda t^{k+1}$ is minimal. That is, we search for the minimizer of

$$\min_{\gamma \in [0,1]} \check{f}(\beta^k + \gamma(z^k - \beta^k)) + \lambda(t^k + \gamma(r^k - t^k)). \tag{45}$$

The Lagrange functional for the above problem is

$$\mathcal{L}(\gamma, u, \ell) = \check{f}(\beta^k + \gamma(z^k - \beta^k)) + \gamma\lambda(r^k - t^k) + u(\gamma - 1) - \ell\gamma,$$

with dual variables $u, \ell \geqslant 0$, and the KKT conditions read as

$$\begin{aligned} (z^k - \beta^k)^T \nabla \check{f}(\beta^k + \gamma(z^k - \beta^k)) + \lambda(r^k - t^k) + u - \ell = 0 \\ 0 \leqslant \gamma \leqslant 1, \quad \ell\gamma = 0, \quad u(\gamma - 1) = 0. \end{aligned}$$

In the following, we narrow down the candidates for γ to at most 4 values. The first two are the extreme points $\gamma = 0$ and $\gamma = 1$. When $0 < \gamma < 1$, both dual variables u and ℓ are zero. Using that

$$\nabla \check{f}(\beta) = \frac{X^T(X\beta - y)}{\sqrt{\|X\beta - y\|_2^2 + \xi^2}}$$

and notation $v := X(z^k - \beta^k)$, $w := X\beta^k - y$, and $\alpha := \lambda(t^k - r^k)$ we rewrite the remaining KKT condition as

$$\frac{v^T(w + \gamma v)}{\sqrt{\|w + \gamma v\|_2^2 + \xi^2}} = \alpha.$$

If there exists γ satisfying the above equality, it also satisfies

$$\left(\langle w, v\rangle + \gamma\|v\|_2^2\right)^2 = \alpha^2\left(\|w + \gamma v\|_2^2 + \xi^2\right).$$

Expanding the squares and rearranging the terms leads to the following quadratic equation

$$\gamma^2\|v\|_2^2(\alpha^2 - \|v\|_2^2) + 2\gamma\langle w, v\rangle(\alpha^2 - \|v\|_2^2) + \alpha^2(\|w\|_2^2 + \xi^2) - \langle w, v\rangle^2 = 0. \tag{46}$$

Note, the if either $v = 0$ or $\|v\|_2^2 = \alpha^2$, the both coefficients in front of γ^2 and γ are zero. Therefore, the above equation has no solutions unless $\alpha = \|v\|_2 = 0$, in which case (45) is constant with respect to γ.

In all other cases, if the roots γ_+ and γ_- of (46) are in the open interval $(0, 1)$, they may be the extremum points of (45). Hence, we obtain the minimizer of the objective of (45) by evaluating the objective at $0, 1$, and, if relevant, γ_+ and γ_-. In summary, we obtain the following version of the Frank-Wolfe algorithm.

For our numerical trials, we used $\xi = 10^{-4}\sqrt{\lambda}$. Unlike other algorithms, where initialization is random, for the Frank-Wolfe algorithm, we used $\beta^0 = 0$, which leads to better performance.

Algorithm 8 Frank-Wolfe for square-root LASSO

```
**Input:** Parameters $\lambda > 0$, $\xi \geqslant 0$, initial guess $\beta^0$.
Set $t^0 = \|\beta^0\|_1$ and $\rho = \sqrt{\|y\|_2^2 + \xi^2}/\lambda$
for $k = 0, 1, 2 \ldots$ do
    Construct $(z^k, r^k)$ as in (44).
    Compute roots $\gamma_+$ and $\gamma_-$ of (46).
    Set up $\Gamma = \{0, 1, \gamma_+, \gamma_-\} \cap [0, 1]$.
    Evaluate the loss function in (45) for $\gamma \in \Gamma$ and set $\gamma_*$ to be its minimizer.
    Update $\beta^{k+1} = (1 - \gamma_*)\beta^k + \gamma_* z^k$ and $t^{k+1} = (1 - \gamma_*)t^k + \gamma_* r^k$.
end for
return Sequence $\{\beta^k\}_{k\geqslant 0}$.
```