**Parameter instability regimes in sparse proximal denoising programs**

Aaron Berk, Yaniv Plan, Özgür Yilmaz

# PARAMETER INSTABILITY REGIMES IN SPARSE PROXIMAL DENOISING PROGRAMS

*Aaron Berk*[†] *Yaniv Plan*[†] *Özgür Yilmaz*[†]

[†] Department of Mathematics, The University of British Columbia

## ABSTRACT

Compressed sensing theory explains why LASSO programs recover structured high-dimensional signals with minimax order-optimal error. Yet, the optimal choice of the program's governing parameter is often unknown in practice. It is still unclear how variation of the governing parameter impacts recovery error in compressed sensing, which is otherwise provably stable and robust. We establish a novel notion of instability in LASSO programs when the measurement matrix is identity. This is the proximal denoising setup. We prove asymptotic cusp-like behaviour of the risk as a function of the parameter choice, and illustrate the theory with numerical simulations. For example, a 0.1% underestimate of a LASSO parameter can increase the error significantly; and a 50% underestimate can cause the error to increase by a factor of $10^9$. We hope that revealing parameter instability regimes of LASSO programs helps to inform a practitioner's choice.

***Index Terms*—** Compressed sensing, Sparse proximal denoising, Parameter instability, Convex optimization, Lasso

## 1. INTRODUCTION

Compressed sensing (CS) is a provably stable and robust [1] technique for simultaneous data acquisition and dimension reduction. Take the sparse linear model $y = Ax_0$ where $x_0 \in \mathbb{R}^N$ is $s$-sparse. The now classical CS result [1, 2, 3, 4, 5, 6] shows if $A$ is suitably random and has $m \geq Cs\log(N/s)$ rows, then one may efficiently recover $x_0$ from $(y, A)$. Numerical implementations of CS are commonly tied to one of three convex $\ell_1$ programs: constrained LASSO, unconstrained LASSO, and quadratically constrained basis pursuit [7]. The advent of suitable fast and scalable algorithms has made the associated family of convex $\ell_1$ minimization problems extremely useful in practice [7, 8, 9, 10].

Proximal denoising (PD) simplifies its CS counterpart, as its measurement matrix is identity. PD uses convex optimization to recover a structured signal corrupted by additive noise. We define three convex programs for PD: constrained proximal denoising, basis pursuit proximal denoising, and unconstrained proximal denoising. For greatest relevance to CS, we assume that $x_0$ is $s$-sparse, having no more than $s$ non-zero entries, and that $y = x_0 + \eta z$, where $z \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\eta > 0$. For $\tau, \sigma, \lambda > 0$, respectively,

$$\hat{x}(\tau) := \underset{x\in\mathbb{R}^N}{\arg\min} \left\{ \|y - x\|_2^2 : \|x\|_1 \leq \tau \right\} \qquad (\text{LS}_\tau^*)$$

$$\tilde{x}(\sigma) := \underset{x\in\mathbb{R}^N}{\arg\min} \left\{ \|x\|_1 : \|y - x\|_2^2 \leq \sigma^2 \right\} \qquad (\text{BP}_\sigma^*)$$

$$x^\sharp(\lambda) := \underset{x\in\mathbb{R}^N}{\arg\min} \left\{ \frac{1}{2}\|y - x\|_2^2 + \lambda\|x\|_1 \right\}. \qquad (\text{QP}_\lambda^*)$$

Minimax order-optimal recovery results for CS and PD programs rely on specific choices of the program's governing parameter (*i.e.,* "using an oracle") [1]. However, the optimal choice of the parameter for these programs is generally unknown in practice. Consequently, it is desirable that the error of the solution exhibit stability with respect to variation of the parameter about its optimal setting. If the optimal choice of parameter yields order-optimal recovery error, then one may hope that a "nearly" optimal choice of parameter admits "nearly" order-optimal recovery error, too (*e.g.,* if the error is no more than a multiplicative constant worse than the optimal one). For example, if $R(\alpha)$ is the mean-squared error of a convex program, with parameter $\alpha > 0$, and $\alpha^* > 0$ is the optimal parameter choice, then one may hope for smooth dependence on $\alpha$, such as

$$R(\alpha) \lesssim \max\left\{ \frac{\alpha^*}{\alpha}, \frac{\alpha}{\alpha^*} \right\}^2 R(\alpha^*).$$

Unfortunately, such a hope cannot be guaranteed in general. We prove the existence of regimes in which PD programs exhibit *parameter instability* — small changes in parameter values can lead to blow-up in risk. We suggest how this behaviour provides intuition in CS for the existence of LASSO parameter instability regimes. Furthermore, we provide an explanation of how PD may perform well in practical settings, despite the existence of parameter instability regimes. This serves to disambiguate these seemingly contradictory results from those of the immense body of work in CS. For a covering of related work, see §2.1. Our numerical results are discussed in §3.

A. Berk is partially supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Canada Graduate Scholarship (CGSD3-489677). Y. Plan is partially supported by an NSERC Discovery Grant (22R23068) and PIMS CRG 33: High-Dimensional Data Analysis. O. Yilmaz was funded in part by an NSERC Discovery Grant (22R82411), an NSERC Accelerator Award (22R68054) and PIMS CRG 33: HDDA.

## 2. MAIN RESULTS

By "risk", we mean the noise-normalized expected squared error of an estimator. For $\hat{x}(\tau), x^\sharp(\lambda)$ and $\tilde{x}(\sigma)$ the risks are

$$\hat{R}(\tau; x_0, N, \eta) = \frac{1}{\eta^2}\mathbb{E}\|\hat{x}(\tau) - x_0\|_2^2,$$
$$R^\sharp(\lambda; x_0, N, \eta) = \frac{1}{\eta^2}\mathbb{E}\|x^\sharp(\eta\lambda) - x_0\|_2^2,$$
$$\tilde{R}(\sigma; x_0, N, \eta) = \frac{1}{\eta^2}\mathbb{E}\|\tilde{x}(\sigma) - x_0\|_2^2.$$

Denote $\Sigma_s^N := \{x \in \mathbb{R}^N : \|x\|_0 \le s\}$ where $\|x\|_0$ gives the number of non-zero entries of $x$, and define the following optimally tuned worst-case risk for $(\mathrm{LS}_\tau^*)$:

$$R^*(s, N, \eta) := \sup_{x_0 \in \Sigma_s^N} \hat{R}(\|x_0\|_1; x_0, N, \eta)$$
$$= \max_{x_0 \in \Sigma_s^N;\, \|x_0\|_1 = 1} \lim_{\eta \to 0} \hat{R}(1; x_0, N, \eta).$$

The second equality is proved in [11]. We use $R^*(s, N)$ as a benchmark, noting it is order-optimal in Proposition 3.

In (1), we show that $(\mathrm{LS}_\tau^*)$ exhibits an asymptotic singularity in the limiting low-noise regime. Namely, $\hat{R}(\tau; x_0, N, \eta)$ blows up for $\tau \neq \|x_0\|_1$. Intuitively, $(\mathrm{LS}_\tau^*)$ is sensitive to $\tau$ when $\eta$ is small, suggesting limited empirical applicability in the low-noise regime when $\|x_0\|_1$ is unknown.

In (2), we show that $(\mathrm{QP}_\lambda^*)$ exhibits an asymptotic phase transition. The worst-case risk over $x_0 \in \Sigma_s^N$ is minimized for parameter choice $\lambda^* = O(\sqrt{\log(N/s)})$ [12]. While $\lambda^*$ has no closed form expression, it satisfies $\lambda^*/\sqrt{2\log(N)} \xrightarrow{N\to\infty} 1$ for $s$ fixed [11]. Thus, we consider the normalized parameter $\mu = \lambda/\sqrt{2\log(N)}$. The risk $R^\sharp(\lambda; x_0, N, \eta)$ is minimax order-optimal when $\mu > 1$ and suboptimal for $\mu < 1$.

Lastly, we show in (3) that $(\mathrm{BP}_\sigma^*)$ is poorly behaved for all $\sigma > 0$ when $x_0$ is very sparse. Namely, $\tilde{R}(\sigma; x_0, N, \eta)$ is asymptotically suboptimal for *any* $\sigma > 0$ when $s/N$ is sufficiently small.

**Theorem 1** (PD Asymptotic Instability)**.** *Where* $\tau^* = 1$*, and* $\lambda(\mu, N) := \mu\sqrt{2\log N}$*,*

$$\lim_{N\to\infty} \max_{\substack{x_0 \in \Sigma_s^N \\ \|x_0\|_1 = 1}} \lim_{\eta \to 0} \frac{\hat{R}(\tau; x_0, N, \eta)}{R^*(s, N, \eta)} = \begin{cases} \infty & \tau < \tau^* \\ 1 & \tau = \tau^* \\ \infty & \tau > \tau^* \end{cases} \quad (1)$$

$$\lim_{N\to\infty} \sup_{x_0 \in \Sigma_s^N} \frac{R^\sharp(\lambda(\mu, N); x_0, N, \eta)}{R^*(s, N, \eta)} = \begin{cases} O(\mu^2) & \mu \ge 1 \\ \infty & \mu < 1 \end{cases} \quad (2)$$

$$\lim_{N\to\infty} \sup_{x_0 \in \Sigma_s^N} \inf_{\sigma > 0} \frac{\tilde{R}(\sigma; x_0, N, \eta)}{R^*(s, N, \eta)} = \infty \quad (3)$$

The proof of (1) computes an approximating sequence $\hat{R}(\tau; x_0, N, \eta_j)$ for $\eta_j \to 0$. The proof of (2) obtains the limits directly from a tractable closed form expression. The proof of (3) proceeds by an involved geometric argument using a novel projection lemma, Lemma 4, and recent results on *local Gaussian mean width* of convex polytopes [13]. Full proofs of the results in this section may be found in an arXiv manuscript [11]. Next, we add two clarifications. First, the three PD programs are equivalent in a sense.

**Proposition 2.** *For* $s \ge 1$*, fix* $x_0 \in \mathbb{R}^N$ *and* $\lambda > 0$*. Where* $x^\sharp(\lambda)$ *solves* $(\mathrm{QP}_\lambda^*)$*, define* $\tau := \|x^\sharp(\lambda)\|_1$ *and* $\sigma := \|y - x^\sharp(\lambda)\|_2$*. Then* $x^\sharp(\lambda)$ *solves* $(\mathrm{LS}_\tau^*)$ *and* $(\mathrm{BP}_\sigma^*)$*.*

However, $\tau$ and $\sigma$ have *stochastic* dependence on $z$, and this mapping may not be smooth. Thus, parameter stability of one program is not implied by that of another.

Second, $R^*(s, N, \eta)$ is computable up to constants. The proof follows by [12] and standard bounds in [1].

**Proposition 3.** *Let* $s \ge 1, N \ge 2$ *be integers, let* $\eta > 0$ *and suppose* $y = x_0 + \eta z$ *for* $z \in \mathbb{R}^N$ *with* $z_i \overset{iid}{\sim} \mathcal{N}(0, 1)$*. Let* $M^*(s, N) := \inf_{x_*} \sup_{x_0 \in \Sigma_s^N} \eta^{-2}\|x_* - x_0\|_2^2$ *be the minimax risk over arbitrary estimators* $x_* = x_*(y)$*. There is* $c, C_1, C_2 > 0$ *such that for* $N \ge N_0 = N_0(s)$*, with* $N_0 \ge 2$ *sufficiently large,*

$$cs\log(N/s) \le M^*(s, N) \le \inf_{\lambda > 0} \sup_{x_0 \in \Sigma_s^N} R^\sharp(\lambda; x_0, N, \eta)$$
$$\le C_1 R^*(s, N, \eta) \le C_2 s\log(N/s).$$

Thus, instead of $R^*(s, N, \eta)$, in Theorem 1 we could have normalized by any of the expressions above, because the three are asymptotically equivalent up to constants. In contrast, a consequence of Proposition 3 using (3) is:

$$\sup_{x_0 \in \Sigma_s^N} \inf_{\sigma > 0} \tilde{R}(\sigma; x_0, N, \eta) \ge \inf_{\sigma > 0} \sup_{x_0 \in \Sigma_s^N} \tilde{R}(\sigma; x_0, N, \eta)$$
$$\gg R^*(s, N, \eta)$$

Importantly, removing dependence of the parameters on the noise destroys the equivalence attained in Proposition 2.

The next result is a projection lemma used in the proof of (3), but we believe it is interesting in its own right. To our knowledge it is novel. Let $\mathrm{P}_C(x) := \arg\min_{y \in C} \|x - y\|_2$ for $\emptyset \neq C$ closed. Given $z \in \mathbb{R}^N$, the one-parameter family $z_t := \mathrm{P}_{tK}(z)$ admits the ordering $\|\mathrm{P}_{tK}(z)\|_2 \le \|\mathrm{P}_{uK}(z)\|_2$ for $0 < t \le u < \infty$ when $0 \in K \subseteq \mathbb{R}^N$ is closed and convex (*cf.* Figure 1a). Consequently, the efficacy with which a PD program recovers the 0 vector may be controlled by a program from the same class.

$K$ need be neither symmetric nor origin-centered, but must contain the origin in the current phrasing. It must be convex; a pictorial counterexample in Figure 1b depicts why.

**Lemma 4** (Projection lemma)**.** *Let* $0 \in K \subseteq \mathbb{R}^n$ *be closed and convex, and fix* $\lambda \ge 1$*. For* $z \in \mathbb{R}^n$*,*

$$\|\mathrm{P}_K(z)\|_2 \le \|\mathrm{P}_{\lambda K}(z)\|_2.$$

*Remark* 1. The proof examines the derivative of the function $f(t) := \|u_t\|_2^2$, where $u_t := t\mathrm{P}_{\lambda K}(z) + (1 - t)\mathrm{P}_K(z)$, and yields a growth rate of this derivative at $t = 0$:

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} f(t)\Big|_{t=0} = \langle z_1, z_\lambda - z_1\rangle \ge \frac{\|z_\lambda - z_1\|_2^2}{\lambda - 1}.$$

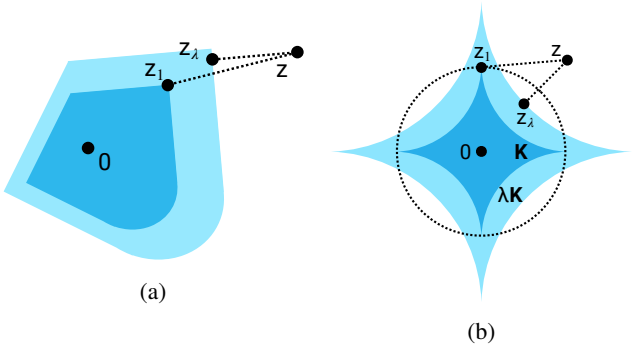


**Fig. 1**: (a) A representation of the projection lemma. Projecting $z$ onto the outer and inner set gives $z_\lambda$ and $z_1$, respectively; evidently, $\|z_1\|_2 \leq \|z_\lambda\|_2$. (b) A counterexample using scaled $\ell_p$ balls for some $0 < p < 1$ to suggest why $K$ must be convex in general. Here, $z$ is projected inwards onto $\lambda K$, but towards a distal vertex when projected onto $K$.

### 2.1. Related work

PD is a simple model that elucidates crucial properties of models in general [14]. As a central model for denoising, it lays the groundwork for CS, deconvolution and inpainting problems [15]. A fundamental signal recovery phase transition in CS is predicted by geometric properties of PD [16], because the minimax risk for PD is equal to the statistical dimension of the signal class [12]. This quantity is a generalized version of $R^*(s, N, \eta)$ introduced above.

A sensitivity to constraint set perturbation is quantified in [12], including an expression for right-sided stability of unconstrained PD. Essentially, PD programs are proximal operators, a powerful tool in convex and non-convex optimization [17, 18, 19, 20, 21]. Thus is PD interesting in its own right, as argued in [12].

Equivalence of the above programs is illuminated from several perspectives [7, 12, 19]. PD risk is considered with more general convex constraints [22]. A connection has been made between the risk of Unconstrained LASSO and $R^\sharp(\lambda; x_0, N, \eta)$ [23, 24].

## 3. NUMERICAL RESULTS

Let $\mathfrak{P} \in \{(\mathrm{LS}^*_\tau), (\mathrm{QP}^*_\lambda), (\mathrm{BP}^*_\sigma)\}$ be a PD program with solution $x^*(\varrho)$ where $\varrho \in \{\tau, \lambda, \sigma\}$ is the associated parameter. Given a signal $x_0$ and noise $\eta z$, denote by $\mathcal{L}(\varrho; x_0, N, \eta z)$ the loss associated to $\mathfrak{P}$ and define $\varrho^* = \varrho(x_0, \eta) > 0$ to be the value of $\varrho$ yielding best risk (*i.e.,* where $\mathbb{E}_z \mathcal{L}(\varrho; x_0, N, \eta z)$ is minimal). We say the normalized parameter $\rho$ for the problem $\mathfrak{P}$ is given by $\rho := \varrho/\varrho^*$ and note that $\rho = 1$ is a population estimate of the argmin of $\mathcal{L}(\varrho; x_0, N, \eta\hat{z})$; by the law of large numbers, this risk estimates well an average of such losses over many realizations $\hat{z}$. Finally, define the auxiliary function $L(\rho; x_0, N, \eta\hat{z}) := \mathcal{L}(\rho\varrho^*; x_0, N, \eta\hat{z})$.

The plots in Figures 2a, 2c and 2d visualize

$$\bar{L}(\rho_i; x_0, N, \eta, k) := \frac{1}{k}\sum_{j=1}^{k} L(\rho_i; x_0, N, \eta\hat{z}_{ij})$$

for each program, evaluated on a grid $\{\rho_i\}_{i=1}^n$ of size $n$ and plotted on a log-log scale, where $L(\rho; x_0, N, \eta\hat{z}) = \eta^{-2}\|x^*(\varrho) - x_0\|_2^2$. Here, each of the $nk$ realizations of the noise is distributed according to $\hat{z}_{ij} \sim \mathcal{N}(0, 1)$, the noise level given by $\eta$ and the signal by $x_0$ where $x_0 = N\sum_{i=1}^s e_i$ with $e_i$ being the $i$th standard basis vector. The grid $\{\rho_i\}_{i=1}^n$ was logarithmically spaced and centered about $\rho_{(n+1)/2} = 1$ with $n$ always odd. The solutions to each PD problem were obtained using standard available methods in Python: `sklearn`'s `minimize_scalar` function from the `optimize` module was used for solving $(\mathrm{LS}^*_\tau)$ and $(\mathrm{BP}^*_\sigma)$ [25], while the solution to $(\mathrm{QP}^*_\lambda)$ was obtained *via* soft-thresholding. Finally, the optimal values $\tau^*, \lambda^*$ and $\sigma^*$ were either determined analytically (*e.g.,* $\tau^* = \|x_0\|_1$), or estimated on a dense grid about an approximately optimal value for that parameter. Initial guesses for $\sigma^*$ and $\lambda^*$ were $\eta\sqrt{N}$ and $\sqrt{2\log(N/s)}$ respectively.

Parameter settings for Figure 2a demonstrating parameter instability of $(\mathrm{LS}^*_\tau)$ in the low-noise regime were given by $(N, s, k, n) = (10^3, 20, 150, 301)$. Pronounced parameter instability was observed for $\eta = 10^{-3}$. At this noise level, the entries of $x_0$ are well-separated from the noise: $N/\eta \sim 10^6, 10^9$. Notably, the $(\mathrm{LS}^*_\tau)$ parameter instability manifests in very low dimensions relative to practical problem sizes. Moreover, the cusp-like curve for $(\mathrm{LS}^*_\tau)$ risk supports the asymptotic singularity described by (1).

The analytic expression for $R^\sharp(\lambda; s, N)$ is plotted in Figure 2b for $\lambda \in \{1 - 10^{-2}, 1 - 10^{-3}, 2\}$ [11, Prop 15]. It is evident from the reference lines $y \sim N^{2/5}$ and $y \sim \sqrt{N}$ that $R^\sharp(u\bar{\lambda}; s, N)$ scales like a power law of $N$ for $u < 1$, while $R^\sharp(2\bar{\lambda}; s, N)$ appears to have approximately order-optimal growth, as predicted by (2).

Parameter settings for Figure 2c demonstrating suboptimality of $(\mathrm{BP}^*_\sigma)$ in the very sparse regime were given by $(N, s, \eta, k, n) = (10^7, 1, 1, 10, 237)$. We limited the number of realizations and grid points because the problem size was computationally prohibitive. The minimal average loss observed on the plot was significantly larger than the respective minimal average losses of $(\mathrm{LS}^*_\tau)$ and $(\mathrm{QP}^*_\lambda)$ by a factor of 82.2, supporting the theory. We also noticed a cusp-like behaviour, which would be an interesting object of further study.

To clarify the relationship of constants appearing in the proof of (3), we provide two examples of minimal $N_0$ values guaranteeing parameter instability behaviour of $(\mathrm{BP}^*_\sigma)$ for given parameter choices. The theory does not claim these values to be optimal, nor do we claim that the constants are tuned. We computed $N_0$ analytically from the proof of (3) for particular parameter choices. Thus, the theory guarantees

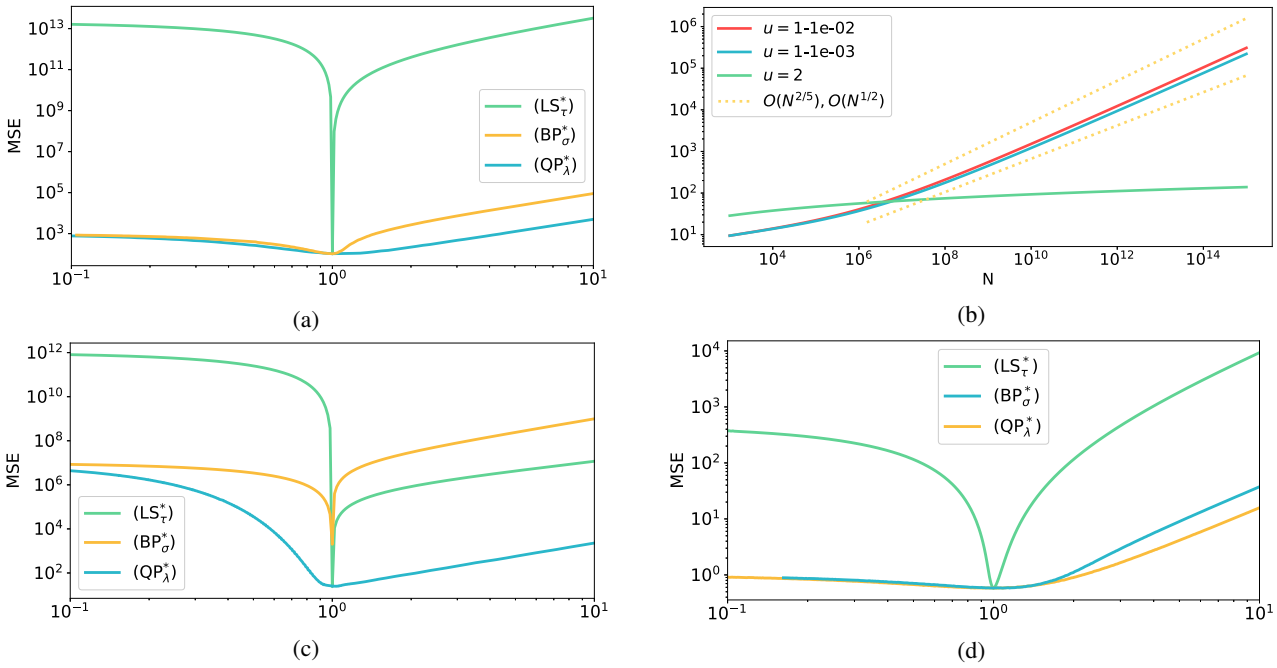


**Fig. 2**: **(a, c, d)** Average losses plotted on a log-log scale with respect to the normalized parameter. **(a)** Low-noise regime parameter instability of $(\mathrm{LS}^*_\tau)$ **(b)** Low-noise regime parameter instability of $(\mathrm{QP}^*_\lambda)$, $R^\sharp(u\lambda^*; s, N)$ computed analytically as a function of $N$ [11, Prop 15] **(c)** Very sparse regime parameter instability of $(\mathrm{BP}^*_\sigma)$ **(d)** Parameter *stability* regime.

parameter instability for all $N \geq N_0$ when [11]

$$\begin{cases} N_0 \approx 1.5\text{e}6 & (a_1, C_1, C_2, L) \approx (1.45, 5, 4, 3.78) \\ N_0 \approx 4.9\text{e}5 & (a_1, C_1, C_2, L) \approx (1.58, 4.04, 4, 3.62). \end{cases}$$

These numbers appear pessimistic, given that $N_0$ is large, while $(C_2, C_1) \approx (4, 5)$ implies the instability arises on the event $\{\|z\|_2^2 - N \in (4\sqrt{N}, 5\sqrt{N})\}$, which occurs with relatively minute (but constant) probability. Thus, it may not be all that surprising that $(\mathrm{BP}^*_\sigma)$ suboptimality is difficult to ascertain empirically from a small number of realizations in only moderately large dimension when $\sigma \approx \sigma^*$. The parameters $a_1$ and $L$ are artifacts of the proof fully described in [11].

Parameter settings for Figure 2d demonstrating a regime in which the three programs exhibit better parameter stability were $(N, s, \eta, k, n) = (10^4, 2500, 233.0, 25, 401)$. As the noise is large, this setting lies (mostly) outside the regime in which $(\mathrm{LS}^*_\tau)$ and $(\mathrm{QP}^*_\lambda)$ exhibit parameter instability. The signal is not very sparse, since $s/N = .25$. Thus, this setting lies outside the regime of $(\mathrm{BP}^*_\sigma)$ parameter instability. Accordingly, smooth risk curves are seen for $(\mathrm{BP}^*_\sigma)$ and $(\mathrm{QP}^*_\lambda)$. While $(\mathrm{QP}^*_\lambda)$ and $(\mathrm{BP}^*_\sigma)$ appear relatively gradual, $(\mathrm{LS}^*_\tau)$ appears at least to avoid a cusp-like point about $\tau/\tau^* = 1$.

## 4. CONCLUSIONS

We have illustrated regimes in which each program is unstable. We hope this informs practitioners about which program to use. Future works include extending this to the CS set-up and to more general atomic norms, some of which are in preparation by the authors.

## 5. REFERENCES


[1] Simon Foucart and Holger Rauhut, *A Mathematical Introduction to Compressive Sensing*, Number 1 in 3. Birkhäuser Basel, 2013.

[2] Emmanuel J Candès, Justin Romberg, and Terence Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[3] Emmanuel J Candes, Justin K Romberg, and Terence Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[4] Emmanuel J Candes and Terence Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[5] David L Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[6] Mark A. Davenport, Marco F. Duarte, Yonina C. Eldar, and Gitta Kutyniok, *Introduction to Compressed Sensing*, p. 164, Cambridge University Press, 2012.

[7] Ewout Van Den Berg and Michael P Friedlander, "Probing the pareto frontier for basis pursuit solutions," *SIAM Journal on Scientific Computing*, vol. 31, no. 2, pp. 890–912, 2008.

[8] Michael P Friedlander, Ives Macedo, and Ting Kei Pong, "Gauge optimization and duality," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 1999–2022, 2014.

[9] Mee Young Park and Trevor Hastie, "$L_1$-regularization path algorithm for generalized linear models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 4, pp. 659–677, 2007.

[10] Jerome Friedman, Trevor Hastie, and Rob Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1, 2010.

[11] Aaron Berk, Yaniv Plan, and Ozgur Yilmaz, "Parameter instability regimes for sparse proximal denoising programs," *arXiv preprint arXiv:1810.11968*, 2018.

[12] Samet Oymak and Babak Hassibi, "Sharp MSE bounds for proximal denoising," *Foundations of Computational Mathematics*, vol. 16, no. 4, pp. 965–1029, 2016.

[13] Pierre C Bellec, "Localized Gaussian width of $M$-convex hulls with applications to LASSO and convex aggregation," *arXiv preprint arXiv:1705.10696*, 2017.

[14] Michael Elad, "Sparse and redundant representation modeling — What next?," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 922–928, 2012.

[15] Michael Elad, Mario AT Figueiredo, and Yi Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, 2010.

[16] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[17] Jérôme Bolte, Patrick L Combettes, and J-C Pesquet, "Alternating proximal algorithm for blind image recovery," in *17th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2010, pp. 1673–1676.

[18] Patrick L Combettes and Jean-Christophe Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pp. 185–212. Springer, 2011.

[19] Dimitri P Bertsekas, Angelia Nedi, Asuman E Ozdaglar, et al., *Convex Analysis and Optimization*, Athena Scientific, 2003.

[20] Jonathan Eckstein and Dimitri P Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1-3, pp. 293–318, 1992.

[21] R Tyrrell Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM journal on control and optimization*, vol. 14, no. 5, pp. 877–898, 1976.

[22] Sourav Chatterjee et al., "A new perspective on least squares under convex constraint," *The Annals of Statistics*, vol. 42, no. 6, pp. 2340–2381, 2014.

[23] Mohsen Bayati and Andrea Montanari, "The LASSO risk for Gaussian matrices," *IEEE Transactions on Information Theory*, vol. 58, no. 4, pp. 1997–2017, 2012.

[24] Mohsen Bayati and Andrea Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764–785, 2011.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.