Trimming the ℓ_1 Regularizer: Statistical Analysis, Optimization, and Applications to Deep Learning

Jihun Yun [1] **Peng Zheng** [2] **Eunho Yang** [1 3] **Aurélie C. Lozano** [4] **Aleksandr Aravkin** [2]

Abstract

We study high-dimensional estimators with the trimmed ℓ_1 penalty, which leaves the h largest parameter entries penalty-free. While optimization techniques for this nonconvex penalty have been studied, the statistical properties have not yet been analyzed. We present the first statistical analyses for M-estimation, and characterize support recovery, ℓ_∞ and ℓ_2 error of the trimmed ℓ_1 estimates as a function of the trimming parameter h. Our results show different regimes based on how h compares to the true support size. Our second contribution is a new algorithm for the trimmed regularization problem, which has the same theoretical convergence rate as difference of convex (DC) algorithms, but in practice is faster and finds lower objective values. Empirical evaluation of ℓ_1 trimming for sparse linear regression and graphical model estimation indicate that trimmed ℓ_1 can outperform vanilla ℓ_1 and non-convex alternatives. Our last contribution is to show that the trimmed penalty is beneficial beyond M-estimation, and yields promising results for two deep learning tasks: input structures recovery and network sparsification.

1. Introduction

We consider high-dimensional estimation problems, where the number of variables p can be much larger that the number of observations n. In this regime, consistent estimation can be achieved by imposing low-dimensional structural constraints on the estimation parameters. *Sparsity* is a prototypical structural constraint, where at most a small set of parameters can be non-zero. A key class of sparsity-constrained estimators is based on regularized M-estimators using *convex* penalties, with the ℓ_1 penalty by far the most common. In the context of linear regression, the Lasso estimator (Tibshirani, 1996) solves an ℓ_1 regularized or constrained least squares problem, and has strong statistical guarantees, including prediction error consistency (van de Geer & Buhlmann, 2009), consistency of the parameter estimates in some norm (van de Geer & Buhlmann, 2009; Meinshausen & Yu, 2009; Candes & Tao, 2007), and variable selection consistency (Meinshausen & Bühlmann, 2006; Wainwright, 2009a; Zhao & Yu, 2006). In the context of sparse Gaussian graphical model (GMRF) estimation, the graphical Lasso estimator minimizes the Gaussian negative log-likelihood regularized by the ℓ_1 norm of the off-diagonal entries of the concentration (Yuan & Lin, 2007; Friedman et al., 2007; Bannerjee et al., 2008). Strong statistical guarantees for this estimator have been established (see Ravikumar et al. (2011) and references therein).

Recently, there has been significant interest in *non-convex* penalties to alleviate the bias incurred by convex approaches, including SCAD and MCP penalties (Fan & Li, 2001; Breheny & Huang, 2011; Zhang et al., 2010; Zhang & Zhang, 2012). In particular, Zhang & Zhang (2012) established consistency for the global optima of least-squares problems with certain non-convex penalties. Loh & Wainwright (2015) showed that under some regularity conditions on the penalty, any stationary point of the objective function will lie within statistical precision of the underlying parameter vector and thus provide ℓ_2- and ℓ_1- error bounds for any stationary point. Loh & Wainwright (2017) proved that for a class of *amenable* non-convex regularizers with vanishing derivative away from the origin (including SCAD and MCP), any stationary point is able to recover the parameter support without requiring the typical incoherence conditions needed for convex penalties. All of these analyses apply to non-convex penalties that are *coordinate-wise separable*.

Our starting point is a family of M-estimators with trimmed ℓ_1 regularization, which leaves the largest h parameters unpenalized. This non-convex family includes the Trimmed

[1]School of Computing, KAIST, Daejeon, South Korea [2]Department of Applied Mathematics, University of Washington, Seattle, USA [3]AITRICS, Seoul, South Korea [4]IBM T.J. Watson Research Center, Yorktown Heights, NY, USA. Correspondence to: Jihun Yun <arcprime@kaist.ac.kr>, Peng Zheng <zhengp@uw.edu>, Eunho Yang <eunhoy@kaist.ac.kr>, Aurélie C. Lozano <aclozano@us.ibm.com>, Aleksandr Aravkin <saravkin@uw.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Lasso (Gotoh et al., 2017; Bertsimas et al., 2017) as a special case. Unlike SCAD and MCP, trimmed regularization exactly solves constrained best subset selection for large enough values of the regularization parameter, and offers more direct control of sparsity via the parameter h. While Trimmed Lasso has been studied from an optimization perspective and with respect to its connections to existing penalties, it has *not* been analyzed from a statistical standpoint.

Contributions:

- We present the *first* statistical analysis of M-estimators with trimmed regularization, *including* Trimmed Lasso. Existing results for non-convex regularizers (Loh & Wainwright, 2015; 2017) cannot be applied as trimmed regularization is neither coordinate-wise decomposable nor "ameanable". We provide support recovery guarantees, ℓ_∞ and ℓ_2 estimation error bounds for general M-estimators, and derive specialized corollaries for linear regression and graphical model estimation. Our results show different regimes based on how the trimming parameter h compares to the true support size.
- To optimize the trimmed regularized problem we develop and analyze a new algorithm, which performs better than difference of convex (DC) functions optimization (Khamaru & Wainwright, 2018).
- Experiments on sparse linear regression and graphical model estimation show ℓ_1 trimming is competitive with other non-convex penalties and vanilla ℓ_1 when h is selected by cross-validation, and has consistent benefits for a wide range of values for h.
- Moving beyond M-estimation, we apply trimmed regularization to two deep learning tasks: (i) recovering input structures of deep models and (ii) network pruning (a.k.a. sparsification, compression). Our experiments on input structure recovery are motivated by Oymak (2018), who quantify complexity of sparsity encouraging regularizers by introducing the covering dimension, and demonstrates the benefits of regularization for learning over-parameterized networks. We show trimmed regularization achieves superior sparsity pattern recovery compared to competing approaches. For network pruning, we illustrate the benefits of trimmed ℓ_1 over vanilla ℓ_1 on MNIST classification using the LeNet-300-100 architecture. Next, motivated by recently developed pruning methods based on variational Bayesian approaches (Dai et al., 2018; Louizos et al., 2018), we propose Bayesian neural networks with trimmed ℓ_1 regularization. In our experiments, these achieve superior results compared to competing approaches with respect to both error and sparsity level. Our work therefore indicates broad relevance of trimmed regularization in multiple problem classes.

2. Trimmed Regularization

Trimming has been typically applied to the *loss* function $\mathcal{L}$ of M-estimators. We can handle outliers by trimming *observations* with large residuals in terms of $\mathcal{L}$: given a collection of n samples, $\mathcal{D} = \{Z_1, \ldots, Z_n\}$, we solve

$$\underset{\boldsymbol{\theta}\in\Omega, \boldsymbol{w}\in\{0,1\}^n}{\text{minimize}} \sum_{i=1}^n w_i \mathcal{L}(\boldsymbol{\theta}; Z_i) \quad \text{s.t.} \sum_{i=1}^n w_i = n - h,$$

where Ω denotes the parameter space (e.g., $\mathbb{R}^p$ for linear regression). This amounts to trimming h outliers as we learn $\boldsymbol{\theta}$ (see Yang et al. (2018) and references therein).

In contrast, we consider here a family of M-estimators with trimmed *regularization* for general high-dimensional problems. We trim entries of $\boldsymbol{\theta}$ that incur the largest penalty using the following program:

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\Omega,\, \boldsymbol{w}\in[0,1]^p}{\text{minimize}} \ & \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) + \lambda_n \sum_{j=1}^p w_j |\theta_j| \\ & \text{s.t. } \mathbf{1}^\top \boldsymbol{w} \geq p - h\,. \end{aligned} \tag{1}$$

Defining the order statistics of the parameter $|\theta_{(1)}| > |\theta_{(2)}| > \ldots > |\theta_{(p)}|$, we can partially minimize over $\boldsymbol{w}$ (setting w_i to 0 or 1 based on the size of $|\theta_i|$), and rewrite the reduced version of problem (1) in $\boldsymbol{\theta}$ alone:

$$\underset{\boldsymbol{\theta}\in\Omega}{\text{minimize}} \ \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) + \lambda_n \mathcal{R}(\boldsymbol{\theta}; h) \tag{2}$$

where the regularizer $\mathcal{R}(\boldsymbol{\theta}; h)$ is the smallest $p - h$ absolute sum of $\boldsymbol{\theta}$: $\sum_{j=h+1}^p |\theta_{(j)}|$. The constrained version of (2) is equivalent to minimizing a loss subject to a sparsity penalty (Gotoh et al., 2017): $\text{minimize}_{\boldsymbol{\theta}\in\Omega} \mathcal{L}(\boldsymbol{\theta}; \mathcal{D})$ s. t. $\|\boldsymbol{\theta}\|_0 \leq h$. For statistical analysis, we focus on the reduced problem (2). When optimizing, we exploit the structure of (1), treating weights $\boldsymbol{w}$ as auxiliary optimization variables, and derive a new fast algorithm with a custom analysis that does not use DC structure.

We focus on two key examples: sparse linear models and sparse graphical models. We also present empirical results for trimmed regularization of deep learning tasks to show that the ideas and methods generalize well to these areas.

Example 1: Sparse linear models. In high-dimensional linear regression, we observe n pairs of a real-valued target $y_i \in \mathbb{R}$ and its covariates $\boldsymbol{x}_i \in \mathbb{R}^p$ in a linear relationship:

$$\boldsymbol{y} = X\boldsymbol{\theta}^* + \boldsymbol{\epsilon}. \tag{3}$$

Here, $\boldsymbol{y} \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a vector of n independent observation errors. The goal is to estimate the k-sparse vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$. According to (2), we use the least squares loss function with trimmed ℓ_1 regularizer (instead of the standard ℓ_1 norm in Lasso (Tibshirani, 1996)):

$$\underset{\boldsymbol{\theta}\in\mathbb{R}^p}{\text{minimize}} \frac{1}{n}\big\|X\boldsymbol{\theta} - \boldsymbol{y}\big\|_2^2 + \lambda_n \mathcal{R}(\boldsymbol{\theta}; h). \tag{4}$$

Example 2: Sparse graphical models. GGMs form a powerful class of statistical models for representing distributions over a set of variables (Lauritzen, 1996), using undirected graphs to encode conditional independence conditions among variables. In the high-dimensional setting, graph sparsity constraints are particularly pertinent for estimating GGMs. The most widely used estimator, the graphical Lasso minimizes the negative Gaussian log-likelihood regularized by the ℓ_1 norm of the entries (or the off-diagonal entries) of the precision matrix (see Yuan & Lin (2007); Friedman et al. (2007); Bannerjee et al. (2008)). In our framework, we replace ℓ_1 norm with its trimmed version: $\text{minimize}_{\Theta \in \mathcal{S}^p_{++}} \text{ trace}\big(\widehat{\Sigma}\Theta\big) - \log\det\big(\Theta\big) + \lambda_n \mathcal{R}(\Theta_{\text{off}}; h)$ where $\mathcal{S}^p_{++}$ denotes the convex cone of symmetric and strictly positive definite matrices, $\mathcal{R}(\Theta_{\text{off}}; h)$ does the smallest $p(p-1) - h$ absolute sum of off-diagonals.

Relationship with SLOPE (OWL) penalty. Trimmed regularization has an apparent resemblance to the SLOPE (or OWL) penalty (Bogdan et al., 2015; Figueiredo & Nowak, 2014), but the two are in fact distinct and pursue different goals. Indeed, the SLOPE penalty can be written as $\sum_{i=1}^p w_i |\beta_{(i)}|$ for a *fixed* set of weights $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$ and where $|\beta_{(1)}| > |\beta_{(2)}| > \cdots > |\beta_{(p)}|$ are the sorted entries of $\boldsymbol{\beta}$. SLOPE is convex and penalizes more those parameter entries with *largest amplitude*, while trimmed regularization is generally non-convex, and only penalizes entries with *smallest amplitude*; the weights are also optimization variables. While the goal of trimmed regularization is to alleviate bias, SLOPE is akin to a significance test where top ranked entries are subjected to a "tougher" threshold, and has been employed for clustering strongly correlated variables (Figueiredo & Nowak, 2014). Finally from a robust optimization standpoint, Trimmed regularization can be viewed as using an optimistic (min-min) model of uncertainty and SLOPE a pessimistic (min-max) counterpart. We refer the interested reader to Bertsimas et al. (2017) for an in-depth exploration of these connections.

Relationship with ℓ_0 regularization. The ℓ_0 norm can be written as $\|\boldsymbol{\theta}\|_0 = \sum_{j=1}^p z_j$ with reparameterization $\theta_j = z_j \tilde{\theta}_j$ such that $z_j \in \{0, 1\}$ and $\tilde{\theta}_j \neq 0$. Louizos et al. (2018) suggest a smoothed version via continuous relaxation on $\boldsymbol{z}$ in a variational inference framework. The variable $\boldsymbol{z}$ plays a similar role to $\boldsymbol{w}$ in our formulation in that they both learn sparsity patterns. In Section 5 we consider a Bayesian extension of the trimmed regularization problem where $\boldsymbol{\theta}$ only is be treated as Bayesian, since we can optimize $\boldsymbol{w}$ without any approximation, in contrast to previous work which needs to relax the discrete nature of $\boldsymbol{z}$.

3. Statistical Guarantees of M-Estimators with Trimmed Regularization

Our goal is to estimate the *true* k-sparse parameter vector (or matrix) $\boldsymbol{\theta}^*$ that is the minimizer of expected loss: $\boldsymbol{\theta}^* := \text{argmin}_{\boldsymbol{\theta} \in \Omega} \mathbb{E}[\mathcal{L}(\boldsymbol{\theta})]$. We use S to denote the support set of $\boldsymbol{\theta}^*$, namely the set of non-zero entries (i.e., $k = |S|$). In this section, we derive support recovery, ℓ_∞ and ℓ_2 guarantees under the following standard assumptions:

- **(C-1)** The loss function $\mathcal{L}$ is differentiable and convex.
- **(C-2)** **(Restricted strong convexity on $\boldsymbol{\theta}$)** Let $\mathbb{D}$ be the possible set of error vector on the parameter $\boldsymbol{\theta}$. Then, for all $\Delta := \boldsymbol{\theta} - \boldsymbol{\theta}^* \in \mathbb{D}$, $\Big\langle \nabla\mathcal{L}(\boldsymbol{\theta}^* + \Delta) - \nabla\mathcal{L}(\boldsymbol{\theta}^*),\, \Delta \Big\rangle \geq \kappa_l \|\Delta\|_2^2 - \tau_1 \frac{\log p}{n} \|\Delta\|_1^2$, where κ_l is a "curvature" parameter, and τ_1 is a "tolerance" constant.

In the high-dimensional setting ($p > n$), the loss function $\mathcal{L}$ cannot be strongly convex in general. (C-2) imposes strong curvature only in some limited directions where the ratio $\frac{\|\Delta\|_1}{\|\Delta\|_2}$ is small. This condition has been extensively studied and known to hold for several popular high dimensional problems (see Raskutti et al. (2010); Negahban et al. (2012); Loh & Wainwright (2015) for instance). The convexity condition of $\mathcal{L}$ in (C-1) can be relaxed as shown in (Loh & Wainwright, 2017). For clarity, however, we focus on convex loss functions.

We begin with ℓ_∞ guarantees. We use a primal-dual witness (PDW) proof technique, which we adapt to the trimmed regularizer $\mathcal{R}(\boldsymbol{\theta}; h)$. The PDW method has been used to analyze the support set recovery of ℓ_1 regularization (Wainwright, 2009c; Yang et al., 2015) as well as decomposable and amenable non-convex regularizers (Loh & Wainwright, 2017). However, the trimmed regularizer $\mathcal{R}(\boldsymbol{\theta}; h)$ is neither decomposable nor amenable, thus the results of Loh & Wainwright (2017) cannot be applied. The key step of PDW is to build a restricted program: Let T be an arbitrary subset of $\{1, \ldots, p\}$ of size h. Denoting $U := S \cup T$ and $V := S - T$, we consider the following restricted program: $\widehat{\boldsymbol{\theta}} \in \text{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^U:\, \boldsymbol{\theta} \in \Omega}\ \mathcal{L}(\boldsymbol{\theta}) + \lambda_n \mathcal{R}(\boldsymbol{\theta}; h)$ where we fix $\widehat{\boldsymbol{\theta}}_j = 0$ for all $j \in U^c$. We further construct the dual variable $\widehat{\boldsymbol{z}}$ to satisfy the zero sub-gradient condition

$$\nabla\mathcal{L}(\widehat{\boldsymbol{\theta}}) + \lambda_n \widehat{\boldsymbol{z}} = 0 \tag{5}$$

where $\widehat{\boldsymbol{z}} = (0, \widehat{\boldsymbol{z}}_V, \widehat{\boldsymbol{z}}_{U^c})$ for $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_T, \widehat{\boldsymbol{\theta}}_V, 0_{U^c})$ (after reordering indices properly) and $\widehat{\boldsymbol{z}}_V \in \partial\|\widehat{\boldsymbol{\theta}}_V\|_1$. We suppress the dependency on T in $\widehat{\boldsymbol{z}}$ and $\widehat{\boldsymbol{\theta}}$ for clarity. In order to derive the final statement, we will establish the strict dual feasibility of $\widehat{\boldsymbol{z}}_{U^c}$, i.e., $\|\widehat{\boldsymbol{z}}_{U^c}\|_\infty < 1$.

The following theorem describes our main theoretical result concerning *any* local optimum of the non-convex program (2). The theorem guarantees under strict dual feasibility that non-relevant parameters of local optimum have smaller absolute values than relevant parameters; hence relevant parameters are not penalized (as long as $h \geq k$).

Theorem 1. *Consider the problem with trimmed regularizer (2) that satisfies (C-1) and (C-2). Let $\widetilde{\boldsymbol{\theta}}$ be an any local minimum of (2) with a sample size $n \geq \frac{2\tau_1}{\kappa_l}(k+h)\log p$ and $\lambda_n \geq 2\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty$. Suppose that:*

(a) given **any** *selection of $T \subseteq \{1,\ldots,p\}$ s.t. $|T| = h$, the dual vector $\widehat{\boldsymbol{z}}$ from the PDW construction (5) satisfies the strict dual feasibility with some $\delta \in (0,1]$, $\|\widehat{\boldsymbol{z}}_{U^c}\|_\infty \leq 1-\delta$ where U is the union of true support S and T,*

(b) letting $\widehat{Q} := \int_0^1 \nabla^2\mathcal{L}\big(\boldsymbol{\theta}^* + t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\big)dt$, the minimum absolute value $\boldsymbol{\theta}^*_{\min} := \min_{j\in S}|\boldsymbol{\theta}^*_j|$ is lower bounded by $\frac{1}{2}\boldsymbol{\theta}^*_{\min} \geq \|(\widehat{Q}_{UU})^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\|_\infty + \lambda_n |||(\widehat{Q}_{UU})^{-1}|||_\infty$ where $|||\cdot|||_\infty$ denotes the maximum absolute row sum of the matrix.

Then, the following properties hold:

(1) For every pair $j_1 \in S, j_2 \in S^c$, we have $|\widetilde{\boldsymbol{\theta}}_{j_1}| > |\widetilde{\boldsymbol{\theta}}_{j_2}|$,

(2) If $h < k$, all $j \in S^c$ are successfully estimated as zero and $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty$ is upper bounded by

$$\big\|\big(\widehat{Q}_{SS}\big)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_S\big\|_\infty + \lambda_n \big|\big|\big|\big(\widehat{Q}_{SS}\big)^{-1}\big|\big|\big|_\infty, \tag{6}$$

(3) If $h \geq k$, at least the smallest (in absolute value) $p-h$ entries in S^c are estimated exactly as zero and we have a simpler (possibly tighter) bound:

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \big\|\big(\widehat{Q}_{\widehat{U}\widehat{U}}\big)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_{\widehat{U}}\big\|_\infty \tag{7}$$

where $\widehat{U}$ is defined as the h largest absolute entries of $\widetilde{\boldsymbol{\theta}}$ including S.

Remarks. The above theorem will be instantiated for the specific cases of sparse linear and sparse graphical models in subsequent corollaries (for which we will bound terms involving $\nabla\mathcal{L}(\boldsymbol{\theta}^*)$, $\widehat{\boldsymbol{z}}$ and $\widehat{Q}$). Though conditions (a) and (b) in Theorem 1 seem apparently more stringent than the case where $h = 0$ (vanilla Lasso), we will see in corollaries that they are uniformly upper bounded for all selections, under the asymptotically same probability as $h = 0$.

Note also that for $h = 0$, we recover the results for the vanilla ℓ_1 norm. Furthermore, by the statement (1) in the theorem, if $h < k$, $\widehat{U}$ only contains relevant feature indices and some relevant features are not penalized. If $h \geq k$, $\widehat{U}$ includes all relevant indices (and some non-relevant indices). In this case, the second term in (6) disappears, but the term $\big\|\big(\widehat{Q}_{\widehat{U}\widehat{U}}\big)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_{\widehat{U}}\big\|_\infty$ increases as $|\widehat{U}|$ gets larger. Moreover, the condition that $n \asymp (k+h)\log p$ will be violated as h approaches p. While we do not know the true sparsity k a priori in many problems, we implicitly assume that we can set $h \asymp k$ (i.e., by cross-validation).

Now we turn to ℓ_2 bound under the same conditions:

Theorem 2. *Consider the problem with trimmed regularizer (2) where all conditions in Theorem 1 hold. Then, for any local minimum of (2), the parameter estimation error in terms of ℓ_2 norm is upper bounded: for some constant C,*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \begin{cases} C\lambda_n\left(\sqrt{k}/2 + \sqrt{k-h}\right) & \text{if } h < k \\ C\lambda_n\sqrt{h}/2 & \text{otherwise} \end{cases}$$

Remarks. The benefit of using trimmed ℓ_1 over standard ℓ_1 can be clearly seen in Theorem 2. Even though both have the same asymptotic convergence rates (in fact, standard ℓ_1 is already information theoretically optimal in many cases such as high-dimensional least squares), trimmed ℓ_1 has a smaller constant: $\frac{3C\lambda_n\sqrt{k}}{2}$ for standard ℓ_1 ($h = 0$) vs. $\frac{C\lambda_n\sqrt{k}}{2}$ for trimmed ℓ_1 ($h = k$). Comparing with non-convex (μ,γ)-amenable regularizers SCAD or MCP, we can also observe that the estimation bounds are asymptotically the same: $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq c\|(\widehat{Q}_{SS})^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_S\|_\infty$ and $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq c\lambda_n\sqrt{k}$. However, the constant c here for those regularizers might be too large if μ is not small enough, since it involves $\frac{1}{\kappa_l - \mu}$ term (vs. $\frac{1}{\kappa_l}$ for the trimmed ℓ_1.) Moreover amenable non-convex regularizers require the additional constraint $\|\boldsymbol{\theta}\|_1 \leq R$ in their optimization problems for theoretical guarantees, along with further assumptions on $\boldsymbol{\theta}^*$ and tuning parameter R, and the true parameter must be feasible for their modified program (see Loh & Wainwright (2017)). The condition $\|\boldsymbol{\theta}^*\|_1 \leq R$ is stringent with respect to the analysis: as p and k increase, in order for R to remain constant, $\|\boldsymbol{\theta}^*\|_\infty$ must shrink to get satisfactory theoretical bounds. In contrast, while choosing the trimming parameter h requires cross-validation, it is possible to set h on a similar order as k.

We are now ready to apply our main theorem to the popular high-dimensional problems introduced in Section 2: sparse linear regression and sparse graphical model estimation. Due to space constraint, the results for sparse graphical models are provided in the supplementary materials.

3.1. Sparse Linear Regression

Motivated by the information theoretic bound for arbitrary methods, all previous analyses of sparse linear regression assume $n \geq c_0 k\log p$ for sufficiently large constant c_0. We also assume $n \geq c_0\max\{k,h\}\log p$, provided $h \asymp k$.

Corollary 1. *Consider the model (3) where $\boldsymbol{\epsilon}$ is sub-Gaussian. Suppose we solve (4) with the selection of:*

(a) $\lambda_n \geq c_\ell\sqrt{\frac{\log p}{n}}$ for some constant c_ℓ depending only on the sub-Gaussian parameters of X and $\boldsymbol{\epsilon}$

(b) h *satisfying: for any selection of* $T \subseteq [p]$ *s.t.* $|T| = h$,

$$\left\|\left|(\widehat{\Gamma}^{-1})_{UU}\right\|\right|_\infty \le c_\infty, \qquad \left\|\left|\widehat{\Gamma}_{U^cU}\left(\widehat{\Gamma}_{UU}\right)^{-1}\right\|\right|_\infty \le \eta,$$
$$\max\left\{\lambda_{\max}(\widehat{\Gamma}_{U^cU^c}), \lambda_{\max}\big((\widehat{\Gamma}_{UU})^{-1}\big)\right\} \le c_u \tag{8}$$

where $\widehat{\Gamma} = \frac{X^\top X}{n}$ *is the sample covariance matrix and* $\lambda_{\max}$ *is the maximum singular value of a matrix.*

Further suppose $\frac{1}{2}\boldsymbol{\theta}^*_{\min} \ge c_1\sqrt{\frac{\log p}{n}} + \lambda_n c_\infty$ *for some constant* c_1. *Then with high probability at least* $1 - c_2\exp(-c_3\log p)$*, any local minimum* $\widetilde{\boldsymbol{\theta}}$ *of* (4) *satisfies*

(a) for every pair $j_1 \in S, j_2 \in S^c$*, we have* $|\widetilde{\boldsymbol{\theta}}_{j_1}| > |\widetilde{\boldsymbol{\theta}}_{j_2}|$,

(b) if $h < k$*, all* $j \in S^c$ *are successfully estimated as zero and we have*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \le c_1\sqrt{\frac{\log p}{n}} + \lambda_n c_\infty,$$
$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \le c_4\sqrt{\frac{\log p}{n}}\left(\sqrt{k}/2 + \sqrt{k-h}\right).$$

(c) if $h \ge k$*, at least the smallest* $p - h$ *entries in* S^c *have exactly zero and we have*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \le c_1\sqrt{\frac{\log p}{n}}, \quad \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \le \frac{c_4}{2}\sqrt{\frac{h\log p}{n}}.$$

Remarks. The conditions in Corollary 1 are also used in previous work and may be shown to hold with high probability via standard concentration bounds for sub-Gaussian matrices. In particular (8) is known as an incoherence condition for sparse least square estimators (Wainwright, 2009b). In the case of vanilla Lasso, estimation will fail if the incoherence condition is violated (Wainwright, 2009b). In contrast, we confirm by simulations in Section 5 that the trimmed ℓ_1 problem (4) can succeed even when this condition is not met. Therefore we conjecture that the incoherence condition could be relaxed in our case, similarly to the case of non-convex μ-amenable regularizers such as SCAD or MCP (Loh & Wainwright, 2017). Proving this conjecture is highly non-trivial, since our penalty is based on a sum of absolute values, which is not μ-amenable; we leave the proof for future work.

4. Optimization

We develop and analyze a block coordinate descent algorithm for solving objective (1), which is highly nonconvex problem because of the coupling of w and θ in the regularizer. The block-coordinate descent algorithm uses simple nonlinear operators: $\text{proj}_{\mathcal{S}}(\boldsymbol{z}) := \arg\min_{\boldsymbol{w}\in\mathcal{S}} \frac{1}{2}\|\boldsymbol{z} - \boldsymbol{w}\|^2$ and $\text{prox}_{\eta\lambda\mathcal{R}(\cdot,\boldsymbol{w}^{k+1})}(\boldsymbol{z}) := \arg\min_{\boldsymbol{\theta}} \frac{1}{2\eta\lambda}\|\boldsymbol{\theta} - \boldsymbol{z}\|^2 + \sum_{j=1}^p w_j^{k+1}|\theta_j|$. Adding a block of weights $\boldsymbol{w}$ decouples the problem into simply computable pieces. Projection onto a polyhedral set is straightforward, while the prox operator is a weighted soft thresholding step.

Algorithm 1 Block Coordinate Descent for (1)

Input: λ, η, and τ.
Initialize: $\boldsymbol{\theta}^0$, $\boldsymbol{w}^0$, and $k = 0$.
while not converged **do**
  $\boldsymbol{w}^{k+1} \leftarrow \text{proj}_{\mathcal{S}}[\boldsymbol{w}^k - \tau\boldsymbol{r}(\boldsymbol{\theta}^k)]$
  $\boldsymbol{\theta}^{k+1} \leftarrow \text{prox}_{\eta\lambda\mathcal{R}(\cdot,\boldsymbol{w}^{k+1})}[\boldsymbol{\theta}^k - \eta\nabla\mathcal{L}(\boldsymbol{\theta}^k)]$
  $k \leftarrow k + 1$
end while
Output: $\boldsymbol{\theta}^k$, $\boldsymbol{w}^k$.

We analyze Algorithm 1 using the structure of (1) instead of relying on the DC formulation for (2). The convergence analysis is summarized in Theorem 3 below. The analysis centers on the general objective function

$$\min_{\boldsymbol{\theta},\boldsymbol{w}} F(\boldsymbol{\theta}, \boldsymbol{w}) := \mathcal{L}(\boldsymbol{\theta}) + \lambda\sum_{i=1}^p w_i r_i(\boldsymbol{\theta}) + \delta(\boldsymbol{w}|\mathcal{S}), \tag{9}$$

where $\delta(\boldsymbol{w}|\mathcal{S})$ enforces $w \in \mathcal{S}$. We let

$$\boldsymbol{r}(\boldsymbol{\theta}) = \begin{bmatrix} r_1(\boldsymbol{x}) & \ldots & r_p(\boldsymbol{x}) \end{bmatrix}^T, \mathcal{R}(\boldsymbol{\theta}, \boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{r}(\boldsymbol{\theta})\rangle.$$

In the case of trimmed ℓ_1, r is the ℓ_1 norm, $r_i(x) = |x_i|$ and $\mathcal{S}$ encodes the constraints $0 \le w_i \le 1$, $\mathbf{1}^T\boldsymbol{w} = p - h$.

We make the following assumptions.

Assumption 1. *(a)* $\mathcal{L}$ *is a smooth closed convex function with an* L_f*-Lipchitz continuous gradient; (b)* r_i *are convex, and* L_r*-Lipchitz continuous and (c)* $\mathcal{S}$ *is a closed convex set and* F *is bounded below.*

In the non-convex setting, we do not have access to distances to optimal iterates or best function values, as we do for strongly convex and convex problems. Instead, we use distance to stationarity to analyze the algorithm. Objective (9) is highly non-convex, so we design a stationarity criterion, which goes to 0 as we approach stationary points. The analysis then shows Algorithm 1 drives this measure to 0, i.e. converges to stationarity. In our setting, every stationary point of (1) corresponds to a local optimum in $\boldsymbol{w}$ with $\boldsymbol{\theta}$ fixed, and a local optimum in $\boldsymbol{\theta}$ with $\boldsymbol{w}$ fixed.

Definition 1 (Stationarity)**.** *Define the stationarity condition* $T(\boldsymbol{\theta}, \boldsymbol{w})$ *by*

$$T(\boldsymbol{\theta}, \boldsymbol{w}) = \min\{\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 : \boldsymbol{u} \in \partial_\theta F(\boldsymbol{\theta}, \boldsymbol{w}), \; \boldsymbol{v} \in \partial_w F(\boldsymbol{\theta}, \boldsymbol{w})\}. \tag{10}$$

The pair $(\boldsymbol{\theta}, \boldsymbol{w})$ *is a stationary point when* $T(\boldsymbol{\theta}, \boldsymbol{w}) = 0$.

Theorem 3. *Suppose Assumptions 1 (a-c) hold, and define the quantity* $\mathcal{G}$ *as follows:*

$$\mathcal{G}_k := \frac{L_f}{2}\|\boldsymbol{\theta}^{k+1} - \boldsymbol{\theta}^k\|^2 + \frac{\lambda}{\tau}\|\boldsymbol{w}^{k+1} - \boldsymbol{w}^k\|^2.$$

With step size $\eta = 1/L_f$, we have,

$$\min_k \mathcal{G}_k \leq \frac{1}{K}\sum_{k=1}^{K}\mathcal{G}_k \leq \frac{1}{K}(F(\boldsymbol{\theta}^1) - F^*)$$

$$T(\boldsymbol{\theta}^{k+1}, \boldsymbol{w}^{k+1}) \leq (4 + 2\lambda L_r/L_f)\mathcal{G}_k,$$

and therefore

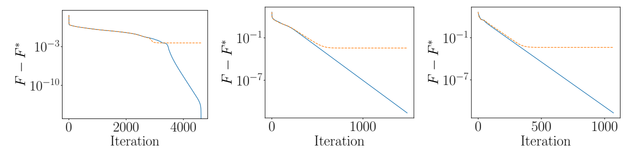
$$\min_{k=1:K}\{T(\boldsymbol{\theta}^k, \boldsymbol{w}^k)\} \leq \frac{4 + 2\lambda L_r/L_f}{K}(F(\boldsymbol{\theta}^1) - F^*).$$

The trimmed ℓ_1 problem satisfies Assumption 1 and hence Theorem 3 holds. Algorithm 1 for (1) converges at a sublinear rate measured using the distance to stationarity T (10), see Theorem 3. In the simulation experiments of Section 5, we will observe that the iterates converge to very close points regardless of initializations. Khamaru & Wainwright (2018) use similar concepts to analyze their DC-based algorithm, since it is also developed for a nonconvex model.

We include a small numerical experiment, comparing Algorithm 1 with Algorithm 2 of (Khamaru & Wainwright, 2018). The authors proposed multiple approaches for DC programs; the prox-type algorithm (Algorithm 2) did particularly well for subset selection, see Figure 2 of (Khamaru & Wainwright, 2018). We generate Lasso simulation data with variables of dimension 500, and 100 samples. The number of nonzero elements in the true generating variable is 10. We take $h = 25$, and apply both Algorithm 1 and Algorithm 2 of (Khamaru & Wainwright, 2018). Initial progress of the methods is comparable, but Algorithm 1 continues at a linear rate to a lower value of the objective, while Algorithm 2 of (Khamaru & Wainwright, 2018) tapers off at a higher objective value. We consistently observe this phenomenon for a broad range of settings, regardless of hyperparameters; see convergence comparisons in Figure 1 for $\lambda \in \{0.5, 5, 20\}$. This comparison is very brief; we leave a detailed study comparing Algorithm 1 with DC-based algorithms to future algorithmic work, along with further analysis of Algorithm 1 and its variants under the Kurdyka-Lojasiewicz assumption (Attouch et al., 2013).

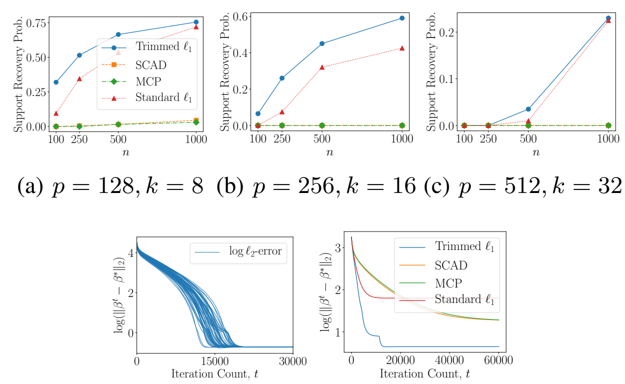
5. Experimental Results

Simulations for sparse linear regression. We design four experiments. For all experiments except the third one where we investigate the effect of small regularization parameters, we choose the regularization parameters via cross-validation from the set: $\log_{10}\lambda \in \{-3.0, -2.8, \ldots, 1.0\}$. For non-convex penalties requiring additional parameter, we just fix their values (2.5 for MCP and 3.0 for SCAD respectively) since they are not sensitive to results. When we generate feature vectors, we consider two different covariance matrices of normal distribution as introduced in Loh & Wainwright (2017) to see how regularizers are affected by the incoherence condition.



(a) $\lambda = 0.5$ (b) $\lambda = 5$ (c) $\lambda = 20$

Figure 1. Convergence of Algorithm 1 (blue solid) vs. Algorithm 2 of (Khamaru & Wainwright, 2018) (orange dot). We see consistent results across parameter settings.



(a) $p = 128, k = 8$ (b) $p = 256, k = 16$ (c) $p = 512, k = 32$

(d) Stationarity (e) $\log \ell_2$-errors

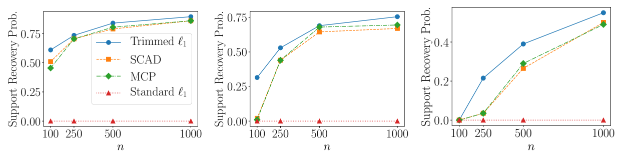
Figure 2. Results for the incoherent case of the first experiments. **(a)**$\sim$**(c)**: Probability of sucessful support recovery for Trimmed ℓ_1, SCAD, MCP, and standard ℓ_1 as sample size n increases. For **(d)**, **(e)**, we adopt the high-dimensional setting with $(n, p, k) = (160, 256, 16)$, and use 50 random initializations.

In our first experiment, we generate i.i.d. observations from $x_i \sim N(0, M_2(\theta))$ where $M_2(\theta) = \theta\mathbf{1}\mathbf{1}^T + (1-\theta)I_p$ with θ = 0.7.[1] This choice of $M_2(\theta)$ satisfies the incoherence condition Loh & Wainwright (2017). We give non-zero values β^* with the magnitude sampled from $N(0, 5^2)$, at k random positions, and the response variables are generated by $y_i = x_i^T\beta^* + \epsilon_i$, where $\epsilon_i \sim N(0, 1^2)$. In Figure 2 (a) $\sim$ (c), we set $(p, k) = (128, 8), (256, 16), (512, 32)$ and increase the sample size n. The probability of correct support recovery for trimmed Lasso is higher than baselines for all samples in all cases. Figure 2(d) corroborates Corollary 1: any local optimum with trimmed ℓ_1 is close to points with correct support regardless of initialization; see comparisons against baselines with same setting in Figure 2(e).

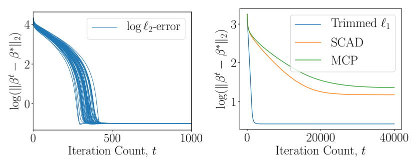
In the second experiment, we replace $M_2(\theta)$ with $M_1(\theta)$, which does not satisfy the incoherence condition.[2] Trimmed still outperforms comparison approaches (Figure 3). Lasso

[1] M_1 and M_2 as defined in Loh & Wainwright (2017).

[2] $M_1(\theta)$ is a matrix with 1's on the diagonal, θ's in the first k positions of the $(k+1)^{\text{st}}$ row and column, and 0's elsewhere.

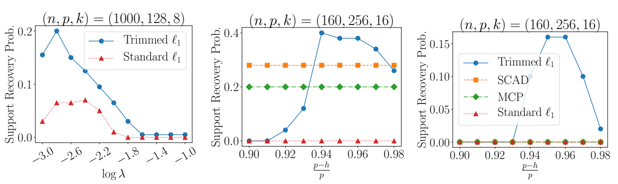


(a) $p = 128, k = 8$ (b) $p = 256, k = 16$ (c) $p = 512, k = 32$



(d) Stationarity (e) log ℓ_2-errors

Figure 3. Results for the non-incoherent case. **(a)**∼**(e)**: same as Figure 2.



(a) Small Regime (b) Non-incoherent (c) Incoherent

Figure 4. Plots for third and last experiments. **(a)**: Trimmed Lasso versus standard one in a small regime. We set $h = \lceil 0.05p \rceil$. **(b)**, **(c)**: Performance of the trimmed Lasso as the value of h varies.

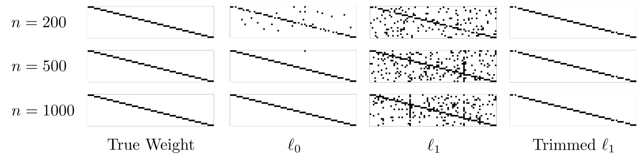
is omitted from Figure 3(e) as it always fails in this setting.

Our next experiment compares Trimmed Lasso against vanilla Lasso where both λ and true non-zeros are small: $\log \lambda \in \{-3.0, -2.8, \ldots, -1.0\}$ and $\beta^* \sim N(0, 0.8^2)$. When the magnitude of $\boldsymbol{\theta}^*$ is large, standard Lasso tends to choose a small value of λ to reduce the bias of the estimate while Trimmed Lasso gives good performance even for large values of λ as long as h is chosen suitably. Figure 4(a) also confirms the superiority of Trimmed Lasso in a small regime of λ with a proper choice of h.

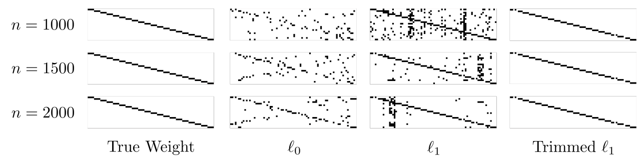
In the last experiment, we investigate the effect of choosing the trimming parameter h. Figure 4(b) and (c) show that Trimmed ℓ_1 outperforms if we set $h = k$ (note $(p-h)/p \approx 0.94$). As $h \downarrow 0$ (when $(p-h)/p = 1$), the performance approaches that of Lasso, as we can see in Corollary 1.

Additional experiments on sparse Gaussian Graphical Models are provided as supplementary materials.

Input Structure Recovery of Compact Neural Networks. We apply the Trimmed ℓ_1 regularizer to recover input structures of deep models. We follow Oymak (2018) and consider the regression model $y_i = \mathbf{1}^T \sigma(\boldsymbol{W}^* \boldsymbol{x}_i)$ with input dimension $p = 80$, hidden dimension $z = 20$, and ReLU



(a) with good initialization (small perturbation from true signal)



(b) with random initialization

Figure 5. Results for sparsity pattern recovery of deep models.

activation $\sigma(\cdot)$. We generate i.i.d. data $\boldsymbol{x}_i \sim N(0, I_p)$ and $\boldsymbol{W}^* \in \mathbb{R}^{z \times p}$ such that ith row has exactly 4 non-zero entries from $N(0, \frac{p}{4z})$ to ensure that $\mathbb{E}[\|\boldsymbol{W}^* \boldsymbol{x}\|_{\ell_2}^2] = \|\boldsymbol{x}\|_{\ell_2}^2$ at only $4(i-1)+1 \sim 4i$ positions. For ℓ_0 and ℓ_1 regularizations, we optimize $\boldsymbol{W}$ using a projected gradient descent with prior knowledge of $\|\boldsymbol{W}^*\|_0$ and $\|\boldsymbol{W}^*\|_1$, and we use Algorithm 1 for trimmed ℓ_1 regularization with $h = 4z$ and $(\lambda, \tau) = (0.01, 0.1)$ obtained by cross-validation. We set the step size $\eta = 0.1$ for all approaches. We consider two sets of simulations with varying sample size n where the initial $\boldsymbol{W}_0$ is selected as (a) a small perturbation of $\boldsymbol{W}^*$ and (b) at random, as in Oymak (2018). Figure 5 shows the results where black dots indicate nonzero values in the weight matrix, and we can confirm that Trimmed ℓ_1 outperforms alternatives in terms of support recovery for both cases.

Pruning Deep Neural Networks. Several recent studies have shown that neural networks are highly overparameterized, and we can prune the weight parameters/neurons with marginal effect on performance. Toward this, we consider trimmed regularization based network pruning. Suppose we have deep neural networks with L hidden layers. Let n_i be the number of neurons in the layer $\boldsymbol{h}_i$. The parameters we are interested in are $\mathcal{W} := \{\boldsymbol{\theta}_l, \boldsymbol{b}_l\}_{l=1}^{L+1}$ for $\boldsymbol{\theta}_l \in \mathbb{R}^{n_{l-1} \times n_l}$ and $\boldsymbol{b}_l \in \mathbb{R}^{n_l}$ where $\boldsymbol{h}_0$ is the input feature $\boldsymbol{x}$ and $\boldsymbol{h}_{L+1}$ is the output $\boldsymbol{y}$. Then, for $l = 1, \ldots, L$, $\boldsymbol{h}_l = \mathrm{ReLU}(\boldsymbol{h}_{l-1}\boldsymbol{\theta}_l + \boldsymbol{b}_l)$. Since the edge-wise pruning will not give actual benefit in terms of computation, we prune unnecessary *neurons* through group-sparse encouraging regularizers. Specifically, given the weight parameter $\boldsymbol{\theta} := \boldsymbol{\theta}_l$ between $\boldsymbol{h}_{l-1}$ and $\boldsymbol{h}_l$, we consider the group norm extension of trimmed ℓ_1: $\mathcal{R}_l(\boldsymbol{\theta}, \boldsymbol{w}) := \lambda \sum_{j=1}^{n_{l-1}} w_j \sqrt{\theta_{j,1}^2 + \cdots + \theta_{j,n_l}^2}$ with the constraint of $\mathbf{1}^T \boldsymbol{w} = n_{l-1} - h$. Moreover, we can naturally make an extension to a convolutional layer with encouraging *activation map sparsity* as follows. If $\boldsymbol{\theta}$ is a weight parameter for 2-dimensional convolutional layer (most generally

Table 1. Results on MNIST using LeNet-300-100.

Method	Pruned Model	Error (%)
No Regularization	784-300-100	**1.6**
grp ℓ_1	784-241-67	1.7
grp $\ell_{1_{\text{trim}}}$, $h =$ half of original	**392-150-50**	**1.6**

Table 2. Results on MNIST classification for LeNet 300-100 with Bayesian approaches. $h = \circ$ means that the trimming parameter h is set to the same sparsity level of $\circ$, and λ sep. indicates that different λ values are employed on each layer.

Method	Pruned Model	Error (%)
ℓ_0 (Louizos et al., 2018)	219-214-100	**1.4**
ℓ_0, λ sep. (Louizos et al., 2018)	266-88-33	1.8
Bayes grp $\ell_{1_{\text{trim}}}$, $h = \ell_0$	219-214-100	**1.4**
Bayes grp $\ell_{1_{\text{trim}}}$, $h = \ell_0$, λ sep.	266-88-33	1.6
Bayes grp $\ell_{1_{\text{trim}}}$, $h < \ell_0$, λ sep.	**245-75-25**	1.7

Table 3. Results on MNIST classification for LeNet-5-Caffe with Bayesian approaches.

Method	Pruned Model	Error (%)
ℓ_0 (Louizos et al., 2018)	20-25-45-462	**0.9**
ℓ_0, λ sep. (Louizos et al., 2018)	9-18-65-25	1.0
Bayes grp $\ell_{1_{\text{trim}}}$, $h < \ell_0$	20-25-45-150	**0.9**
Bayes grp $\ell_{1_{\text{trim}}}$, $h = \ell_0$, λ sep.	9-18-65-25	1.0
Bayes grp $\ell_{1_{\text{trim}}}$, $h < \ell_0$, λ sep.	**8-17-53-19**	1.0

used) with $\boldsymbol{\theta} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times H \times W}$, the trimmed regularization term that induces activation map-wise sparsity is given by $\mathcal{R}_l(\boldsymbol{\theta}, \boldsymbol{w}) := \lambda \sum_{j=1}^{C_{\text{out}}} w_j \sqrt{\sum_{m,n,k} \theta_{j,m,n,k}^2}$ for all possible indices (m, n, k). Finally, we add all penalizing terms to a loss function to have $\mathcal{L}(\mathcal{W}; \mathcal{D}) + \sum_{l=1}^{L+1} \lambda_l \mathcal{R}_l(\boldsymbol{\theta}_l, \boldsymbol{w}_l)$ where we allow different hyperparameters λ_l and h_l for each layer.

In Table 1, we compare trimmed group ℓ_1 regularization against vanilla group ℓ_1 on MNIST dataset using LeNet-300-100 architecture (Lecun et al., 1998). Here, we set the trimming parameter h to half sparsity level of the original model. For the vanilla group ℓ_1, we need larger λ values to obtain sparser models, for which we pay a significant loss of accuracy. In contrast, we can control the sparsity level using trimming parameters h with little or no drop of accuracy.

Most algorithms for network pruning recently proposed are based on a variational Bayesian approach (Dai et al., 2018; Louizos et al., 2018). Motivated by learning sparse structures via smoothed version of ℓ_0 norm (Louizos et al., 2018), we propose a Bayesian neural network with trimmed regularization where we regard only $\boldsymbol{\theta}$ as Bayesian. Inspired by a relation between variational dropout and Bayesian neural networks (Kingma et al., 2015), we specifically choose a fully factorized Gaussian as a variational distribution, $q_{\boldsymbol{\phi},\boldsymbol{\alpha}}(\theta_{i,j}) = \mathcal{N}(\phi_{i,j}, \alpha_{i,j}\phi_{i,j}^2)$, to approximate the true posterior and leave $\boldsymbol{w}$ to directly learn sparsity patterns. Then the problem is cast to maximizing corresponding evidence lower bound (ELBO), $\mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\alpha}}}[\mathcal{L}(\mathcal{W}; \mathcal{D})] - \mathbb{KL}\big(q_{\boldsymbol{\phi},\boldsymbol{\alpha}}(\mathcal{W}) \| p(\mathcal{W})\big)$. Combined with trimmed ℓ_1 regularization, the objective is

$$\mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\alpha}}(\boldsymbol{\theta})}\Big[-\mathcal{L}(\mathcal{W}; \mathcal{D}) + \sum_{l=1}^{L+1} \lambda_l \mathcal{R}_l(\boldsymbol{\theta}_l, \boldsymbol{w}_l)\Big] + \mathbb{KL}(q_{\boldsymbol{\phi},\boldsymbol{\alpha}}(\mathcal{W}) \| p(\mathcal{W})) \quad (11)$$

which can be interpreted as a sum of expected loss and expected trimmed group ℓ_1 penalizing term. Kingma & Welling (2014) provide the efficient unbiased estimator of stochastic gradients for training $(\boldsymbol{\phi}, \boldsymbol{\alpha})$, via the reparameterization trick to avoid computing gradient of sampling process. In order to speed up our method, we approximate expected loss term in (11) using a local reparameterization trick (Kingma et al., 2015) while the standard reparameterization trick is used for the penalty term.

Trimmed group ℓ_1 regularized Bayesian neural networks have smaller capacity with less error than other baselines (Table 2). Our model has lower error rate and better sparsity even for convolutional network, LeNet-5-Caffe[3] (Table 3).[4]

The code is available at `https://github.com/abcdxyzpqrst/Trimmed_Penalty`.

6. Concluding Remarks

In this work we studied statistical properties of high-dimensional M-estimators with the trimmed ℓ_1 penalty, and demonstrated the value of trimmed regularization compared to convex and non-convex alternatives. We developed a provably convergent algorithm for the trimmed problem, based on specific problem structure rather than generic DC structure, with promising numerical results. A detailed comparison to DC based approaches is left to future work. Going beyond M-estimation, we showed that trimmed regularization can be beneficial for two deep learning tasks: input structure recovery and network pruning. As future work we plan to study trimming of general decomposable regularizers, including ℓ_1/ℓ_q norms, and further investigate the use of trimmed regularization in deep models.

Acknowledgement. This work was supported by the National Research Foundation of Korea (NRF) grants (No.2018R1A5A1059921, No.2019R1C1C1009192), Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No.2016-0-00563, No.2017-0-01779,

[3] `https://github.com/BVLC/caffe/tree/master/examples/mnist`

[4] We only consider methods based on sparsity encouraging regularizers. State-of-the-art VIBNet (Dai et al., 2018) exploits the mutual information between each layer.

No.2019-0-01371) funded by the Korea government (MSIT) and Samsung Research Funding & Incubation Center via SRFC-IT1702-15.

References

Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

Bannerjee, O., , Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Jour. Mach. Lear. Res.*, 9:485–516, March 2008.

Bertsimas, D., Copenhaver, M. S., and Mazumder, R. The trimmed lasso: Sparsity and robustness. *arXiv preprint arXiv:1708.04527*, 2017.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. Slope?adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103, 2015.

Breheny, P. and Huang, J. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232, 2011.

Candes, E. J. and Tao, T. The dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

Dai, B., Zhu, C., Guo, B., and Wipf, D. Compressing neural networks using the variational information bottleneck. In *International Conference on Machine learning (ICML)*, 2018.

Fan, J. and Li, R. Variable selection via non-concave penalized likelihood and its oracle properties. *Jour. Amer. Stat. Ass.*, 96(456):1348–1360, December 2001.

Figueiredo, M. A. and Nowak, R. D. Sparse estimation with strongly correlated variables using ordered weighted l1 regularization. *arXiv preprint arXiv:1409.4005*, 2014.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 2007.

Gotoh, J.-y., Takeda, A., and Tono, K. Dc formulations and algorithms for sparse optimization problems. *Mathematical Programming*, pp. 1–36, 2017.

Khamaru, K. and Wainwright, M. J. Convergence guarantees for a class of non-convex and non-smooth optimization problems. *arXiv preprint arXiv:1804.09629*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *International Conference on Learning Representation (ICLR)*, 2014.

Kingma, D. P., Salimans, T., and Welling, M. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Lauritzen, S. *Graphical models*. Oxford University Press, USA, 1996.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pp. 2278–2324, 1998.

Loh, P. and Wainwright, M. J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Neur. Info. Proc. Sys. (NIPS)*, 26, 2013.

Loh, P. and Wainwright, M. J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research (JMLR)*, 16:559–616, 2015.

Loh, P. and Wainwright, M. J. Support recovery without incoherence: A case for nonconvex regularization. *Annals of Statistics*, 45(6):2455–2482, 2017.

Louizos, C., Welling, M., and Kingma, D. P. Learning sparse neural networks through ℓ_0 regularization. In *International Conference on Learning Representation (ICLR)*, 2018.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

Meinshausen, N. and Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2009.

Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Oymak, S. Learning compact neural networks with regularization. In *International Conference on Machine learning (ICML)*, 2018.

Raskutti, G., Wainwright, M. J., and Yu, B. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research (JMLR)*, 99:2241–2259, 2010.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

van de Geer, S. and Buhlmann, P. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009a.

Wainwright, M. J. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Info. Theory*, 55:2183–2202, 2009b.

Wainwright, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55:2183–2202, May 2009c.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research (JMLR)*, 16:3813–3847, 2015.

Yang, E., Lozano, A., and Aravkin, A. General family of trimmed estimators for robust high-dimensional data analysis. *Electronic Journal of Statistics*, 12:3519–3553, 2018.

Yuan, M. and Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Zhang, C.-H. and Zhang, T. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, pp. 576–593, 2012.

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

Zhao, P. and Yu, B. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.