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]

[1]KAIST [2]University of Washington [3]AITRICS [4]IBM T.J. Watson Research Center

arcprime@kaist.ac.kr

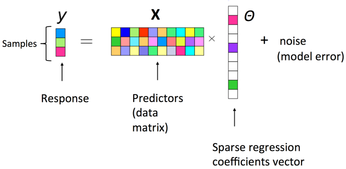
International Conference on Machine Learning
June 12, 2019

Table of Contents

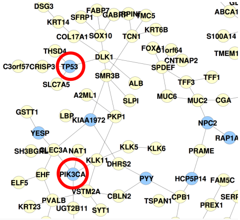
Table of Contents

ℓ_1 Regularization is Popular

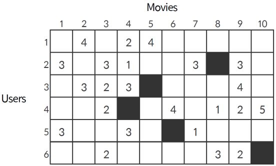
- High-dimensional data with ℓ_1 regularization ($n \ll p$)
  - Genomic Data, Matrix Completion, Deep Learning, etc.



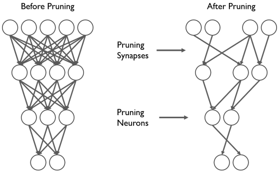
(a) Sparse linear models



(b) Sparse graphical models



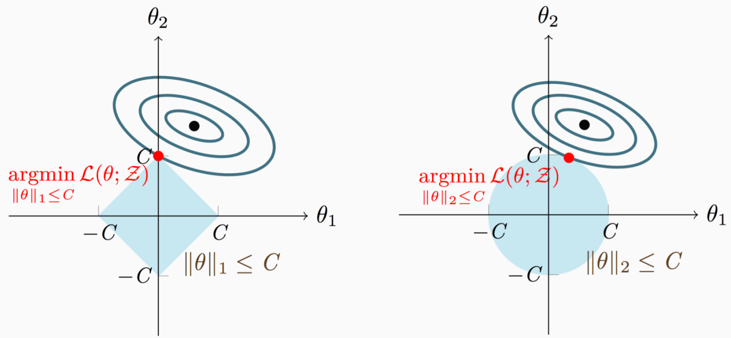
(c) Matrix Completion



(d) Sparse neural networks

Example 1: Lasso* (Sparse Linear Regression)

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\Omega}{\operatorname{argmin}} \; \frac{1}{2n}\|y - X\boldsymbol{\theta}\|_2^2 + \lambda_n \|\boldsymbol{\theta}\|_1$$



* R. Tibshirani. Regression shrinkage and selection via the lasso. JRSS, Series B,1996.

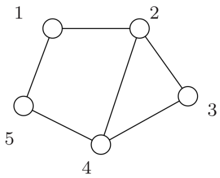
Concrete Example 2

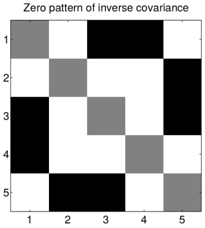
Graphical Lasso

Example 2: Graphical Lasso* (Sparse Concentration Matrix)

$$\widehat{\Theta} \in \underset{\mathbf{\Theta} \in \mathcal{S}^p_{++}}{\operatorname{argmin}} \; \color{red}{\operatorname{trace}(\widehat{\Sigma}\Theta) - \log\det(\Theta)} \color{black}{+ \lambda_n} \color{blue}{\|\Theta\|_{1,\mathrm{off}}}$$

where $\widehat{\Sigma}$ is a sample covariance matrix, $\mathcal{S}^p_{++}$ the symmetric and strictly positive definite matrices, and $\|\Theta\|_{1,\mathrm{off}}$ the ℓ_1-norm on the off-diagonal elements of Θ.





* P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing l1-penalized log-determinant divergence. EJS, 2011

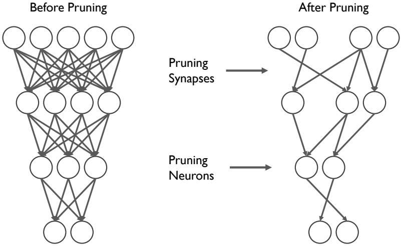
Concrete Example 3

Group ℓ_1 on Network Pruning Task

Example 3: Group ℓ_1* (Structured Sparsity of Weight Parameters)

$$\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \Omega}{\operatorname{argmin}} \; \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) + \lambda_n \|\boldsymbol{\theta}\|_{\mathcal{G}}$$

where $\widehat{\boldsymbol{\theta}}$ is a collection of weight parameters of neural networks, $\mathcal{L}$ the neural network loss (ex. softmax), and $\|\boldsymbol{\theta}\|_{\mathcal{G}}$ the group sparsity regularizer.



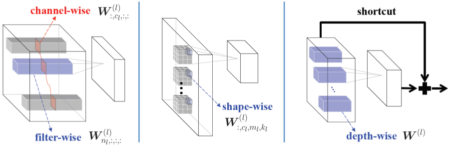


Figure: Encouraging **group sparsity**. For example, $\|\theta\|_{\mathcal{G}} = \sum_{g \in \mathcal{G}} \|\theta_g\|_2$ with each group g.

*W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning Structured Sparsity in Deep Neural Networks. NIPS, 2016

Shrinkage Bias of Standard ℓ_1 Penalty

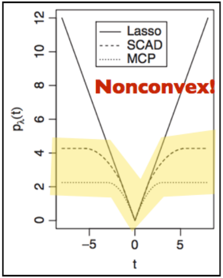
- As parameter size gets larger, the shrinkage bias effect also tends to be larger.
  - The ℓ_1 penalty is proportional to the size of parameters.

Despite the popularity of ℓ_1 penalty
(and also strong statistical guarantees),
Is it really good enough?

Non-convex Regularizers

Previous Work

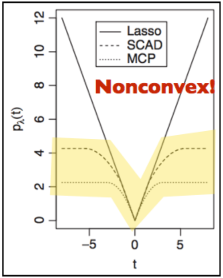
- For *amenable* non-convex regularizers (such as SCAD* and MCP**),
  - ▷ *Amenable* regularizer: Resembles ℓ_1 at the origin and has vanishing derivatives at the tail. $\rightarrow$ *coordinate-wise decomposable*.
  - ▷ (Loh & Wainwright)*** provide the statistical analysis on amenable regularizers.



Non-convex Regularizers

Previous Work

- For *amenable* non-convex regularizers (such as SCAD* and MCP**),
  - ▷ *Amenable* regularizer: Resembles ℓ_1 at the origin and has vanishing derivatives at the tail. $\rightarrow$ *coordinate-wise decomposable*.
  - ▷ (Loh & Wainwright)*** provide the statistical analysis on amenable regularizers.



What about more structurally complex regularizer?

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

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

*** P. Loh and M. J. Wainwright. Regularized M-estimators with non-convexity: statistical and algorithmic theory for local optima and algorithmic. JMLR, 2015.

*** P. Loh and M. J. Wainwright. Support recovery without incoherence: A case for nonconvex regularization. The Annals of Statistics, 2017.

Trimmed ℓ_1 Penalty

Definition

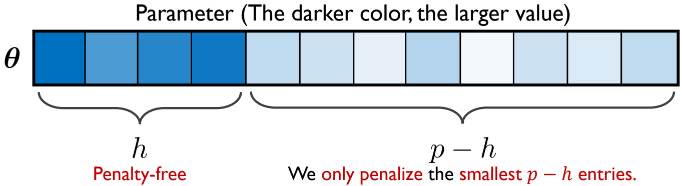
- In this paper, we study the Trimmed ℓ_1 penalty.
  - New class of regularizers.

- In this paper, we study the Trimmed ℓ_1 penalty.
  - New class of regularizers.

- **Definition:**
  For a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, we only ℓ_1-**penalize** each entry **except largest** h **entries** (We call h the trimming parameter).

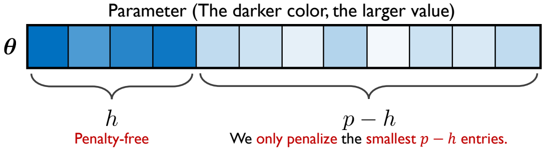
- In this paper, we study the Trimmed ℓ_1 penalty.
  - New class of regularizers.

- **Definition:**
  For a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^p$, we only ℓ_1-**penalize** each entry **except largest** h **entries** (We call h the trimming parameter).



Trimmed ℓ_1 Penalty

First Formulation



- We can *formalize* by defining the order statistics of the parameter vector $|\theta_{(1)}| > |\theta_{(2)}| > \cdots > |\theta_{(p)}|$, the M-estimation with the Trimmed ℓ_1 penalty is

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

  where the regularizer $\mathcal{R}(\boldsymbol{\theta};h) = \sum_{j=h+1}^{p} |\theta_{(j)}|$ (sum of smallest $p-h$ entries in absolute values).
  - Importantly, the Trimmed ℓ_1 is not amenable nor coordinate-wise separable.

- We can rewrite the M-estimation with the Trimmed ℓ_1 penalty by introducing additional variable $\boldsymbol{w}$:

$$\underset{\boldsymbol{\theta}\in\Omega,\boldsymbol{w}\in[0,1]^p}{\text{minimize}} \quad \mathcal{F}(\boldsymbol{\theta},\boldsymbol{w}) := \mathcal{L}(\boldsymbol{\theta};\mathcal{D}) + \lambda_n \sum_{j=1}^{p} w_j|\theta_j|$$

$$\text{such that } \mathbf{1}^T\boldsymbol{w} \geq p - h$$

- We can rewrite the M-estimation with the Trimmed ℓ_1 penalty by introducing additional variable $\boldsymbol{w}$:

$$\underset{\boldsymbol{\theta}\in\Omega,\boldsymbol{w}\in[0,1]^p}{\text{minimize}} \quad \mathcal{F}(\boldsymbol{\theta},\boldsymbol{w}) := \mathcal{L}(\boldsymbol{\theta};\mathcal{D}) + \lambda_n \sum_{j=1}^{p} w_j|\theta_j|$$

$$\text{such that } \mathbf{1}^T\boldsymbol{w} \geq p-h$$

- The variable $\boldsymbol{w}$ encodes the **sparsity pattern** and **order information** of $\boldsymbol{\theta}$. As an ideal case,
  - $w_j = 0$ for largest h entries
  - $w_j = 1$ for smallest $p-h$ entries

- We can rewrite the M-estimation with the Trimmed ℓ_1 penalty by introducing additional variable $\boldsymbol{w}$:

$$\underset{\boldsymbol{\theta}\in\Omega,\boldsymbol{w}\in[0,1]^p}{\text{minimize}} \quad \mathcal{F}(\boldsymbol{\theta},\boldsymbol{w}) := \mathcal{L}(\boldsymbol{\theta};\mathcal{D}) + \lambda_n \sum_{j=1}^{p} w_j|\theta_j|$$

such that $\mathbf{1}^T\boldsymbol{w} \geq p-h$

- The variable $\boldsymbol{w}$ encodes the **sparsity pattern** and **order information** of $\boldsymbol{\theta}$. As an ideal case,
  - $w_j = 0$ for largest h entries
  - $w_j = 1$ for smallest $p-h$ entries

- If we set the trimming parameter $h = 0$, it is just a standard ℓ_1.

$$\underset{\boldsymbol{\theta}\in\Omega, \boldsymbol{w}\in[0,1]^p}{\text{minimize}} \quad \mathcal{F}(\boldsymbol{\theta}, \boldsymbol{w}) := \mathcal{L}(\boldsymbol{\theta}; \mathcal{D}) + \lambda_n \sum_{j=1}^{p} w_j |\theta_j|$$

$$\text{such that } \mathbf{1}^T \boldsymbol{w} \geq p - h$$

- The objective function $\mathcal{F}$ is
  - Weighted ℓ_1-regularized if we fix $\boldsymbol{w}$.
  - Linear in $\boldsymbol{w}$ with fixing $\boldsymbol{\theta}$.
  - However, $\mathcal{F}$ is **non-convex** in jointly $(\boldsymbol{\theta}, \boldsymbol{w})$ because of coupling of $\boldsymbol{\theta}$ and $\boldsymbol{w}$.
- We use this second formulation for an **optimization.**
  - Since we don't need to sort the parameter.

Trimmed ℓ_1 Penalty

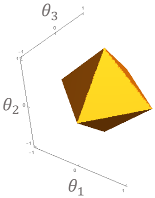
Unit Balls Visualization

- **Trimmed ℓ_1 Unit balls** of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ in the 3-dimensional space.

Trimmed ℓ_1 Penalty

Unit Balls Visualization

- **Trimmed ℓ_1 Unit balls** of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ in the 3-dimensional space.

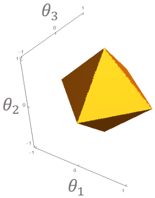


$h = 0$

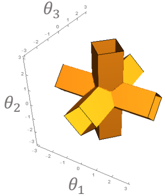
Trimmed ℓ_1 Penalty

Unit Balls Visualization

- **Trimmed** ℓ_1 **Unit balls** of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ in the 3-dimensional space.

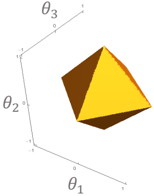


$h = 0$

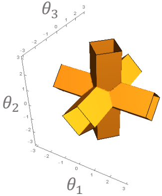


$h = 1$

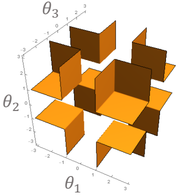
- **Trimmed ℓ_1 Unit balls** of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ in the 3-dimensional space.



$h = 0$

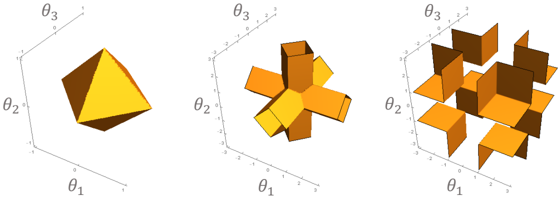


$h = 1$



$h = 2$

- **Trimmed ℓ_1 Unit balls** of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ in the 3-dimensional space.



$h = 0$ $\quad$ $h = 1$ $\quad$ $h = 2$

- For $h = 0$, the shape is the same as standard ℓ_1 unit ball.
- For $h > 0$, the penalty could be unbounded.
  - Since the largest h entries are not penalized, the unit ball could extend to infinity in these directions.
  - As h increases, the penalty would be more complicated.

Table of Contents

Statistical Analysis: Key Assumptions and Quantity

Assumptions:

(C1) The loss $\mathcal{L}$ is differentiable and convex.

(C2) **Restricted Strong Convexity on $\boldsymbol{\theta}$:** Let $\mathbb{D}$ be the set of all possible error vectors for $\boldsymbol{\theta}$. Then, for all $\boldsymbol{\theta} - \boldsymbol{\theta}^* \in \mathbb{D}$,

$$\langle \nabla \mathcal{L}(\boldsymbol{\theta}^*, \Delta) - \nabla \mathcal{L}(\boldsymbol{\theta}^*), \Delta \rangle \geq \kappa_l \|\Delta\|_2^2 - \tau_1 \frac{\log p}{n} \|\Delta\|_1^2,$$

where κ_l is a "curvature" parameter, and τ_1 a "tolerance".

- Allowing a small loss difference to be translated to a small error $\boldsymbol{\theta} - \boldsymbol{\theta}^*$.
- **RSC condition** is a **standard one** in this line of work.

Quantity:

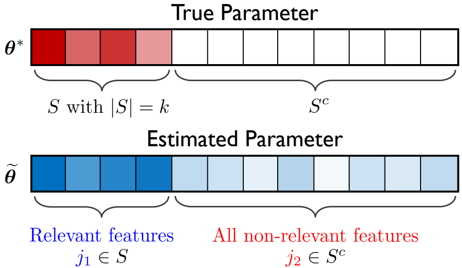
- Let $\widehat{Q} = \int_0^1 \nabla^2 \mathcal{L}(\boldsymbol{\theta}^* + t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*))dt$.

- Consider an M-estimation problem with the Trimmed ℓ_1 penalty.
- Under (C1)&(C2) and standard conditions, for *any* local minimum $\widetilde{\boldsymbol{\theta}}$, we have

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

- Consider an M-estimation problem with the Trimmed ℓ_1 penalty.
- Under (C1)&(C2) and standard conditions, for *any* local minimum $\widetilde{\boldsymbol{\theta}}$, we have

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



Statistical Analysis

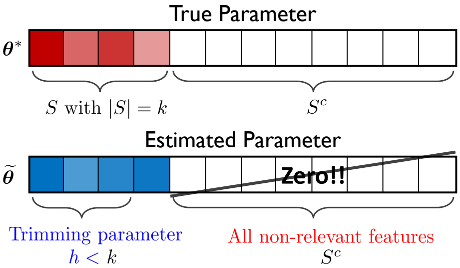
Theorem 1: General ℓ_∞-error Bound and Variable Selection

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

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \left\|(\widehat{Q}_{SS})^{-1} \nabla \mathcal{L}(\boldsymbol{\theta}^*)_S\right\|_\infty + \lambda_n \left\|\left|(\widehat{Q}_{SS})^{-1}\right|\right\|_\infty$$

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

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \left\|(\widehat{Q}_{SS})^{-1} \nabla \mathcal{L}(\boldsymbol{\theta}^*)_S\right\|_\infty + \lambda_n \left\|\left|(\widehat{Q}_{SS})^{-1}\right|\right\|_\infty$$



Statistical Analysis

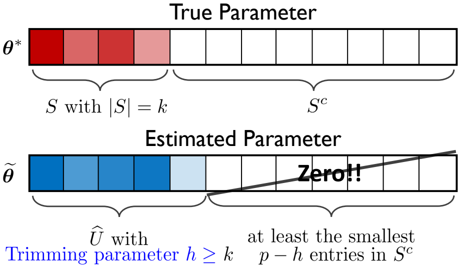
Theorem 1: General ℓ_∞-error Bound and Variable Selection

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
$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \left\|(\widehat{Q}_{SS})^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_S\right\|_\infty + \lambda_n \left|\left|\left|(\widehat{Q}_{SS})^{-1}\right|\right|\right|_\infty$$
3. If $h \geq k$, at least the smallest (in absolute) $p - h$ entries in S^c are exactly zero and $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_\infty \leq \|(\widehat{Q}_{\widehat{U}\widehat{U}})^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_{\widehat{U}}\|_\infty$ where $\widehat{U}$ is defined as the h largest absolute entries of $\widetilde{\boldsymbol{\theta}}$ including S.

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

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \left\|(\widehat{Q}_{SS})^{-1} \nabla \mathcal{L}(\boldsymbol{\theta}^*)_S\right\|_\infty + \lambda_n \left\|\left|(\widehat{Q}_{SS})^{-1}\right|\right\|_\infty$$

3. If $h \geq k$, at least the smallest (in absolute) $p - h$ entries in S^c are exactly zero and $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\|_\infty \leq \|(\widehat{Q}_{\widehat{U}\widehat{U}})^{-1} \nabla \mathcal{L}(\boldsymbol{\theta}^*)_{\widehat{U}}\|_\infty$ where $\widehat{U}$ is defined as the h largest absolute entries of $\widetilde{\boldsymbol{\theta}}$ including S.



Theorem 2

- Consider an M-estimation problem with Trimmed ℓ_1 regularization where all conditions in Theorem 1 hold.
- For any local minimum $\widetilde{\boldsymbol{\theta}}$, the parameter estimation error in terms of ℓ_2-norm is upper bounded as:

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

Theorem 2

- Consider an M-estimation problem with Trimmed ℓ_1 regularization where all conditions in Theorem 1 hold.
- For any local minimum $\widetilde{\boldsymbol{\theta}}$, the parameter estimation error in terms of ℓ_2-norm is upper bounded as:

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

- From our bound, $h = k$ is the best case!
  - We can choose $h \asymp k$ via cross-validation.

Table: ℓ_2-error bound for different h values.

	$h < k$	$h = k$	$h > k$
$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$	$C\lambda_n(\frac{\sqrt{k}}{2} + \sqrt{k-h})$	$C\lambda_n\frac{\sqrt{k}}{2}$	$C\lambda_n\frac{\sqrt{h}}{2}$

Statistical Analysis

Remarks: Other alternative penalties vs. Trimmed ℓ_1

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

- $\rho_\lambda(t)$: (μ, γ)-amenable
  - $\rho_\lambda(t) + \frac{1}{2}\mu t^2$ is convex.
  - $\rho'_\lambda(t) = 0$ for $|t| > \gamma$.

Statistical Analysis

Remarks: Other alternative penalties vs. Trimmed ℓ_1

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

- $\rho_\lambda(t)$: (μ, γ)-amenable
  - $\rho_\lambda(t) + \frac{1}{2}\mu t^2$ is convex.
  - $\rho'_\lambda(t) = 0$ for $|t| > \gamma$.

Table: ℓ_2-error bound comparison with universal constant c_0 in sub-Gaussian tail bounds.

	Standard ℓ_1 $(h = 0)$	(μ, γ)-amenable	Trimmed ℓ_1 $(h = k)$
$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$	$\frac{3c_0}{\kappa_l}\frac{\lambda_n\sqrt{k}}{2}$	$\frac{c_0}{\kappa_l - \frac{3}{2}\mu}\frac{\lambda_n\sqrt{k}}{2}$	$\frac{c_0}{\kappa_l}\frac{\lambda_n\sqrt{k}}{2}$

Statistical Analysis

Remarks: Other alternative penalties vs. Trimmed ℓ_1

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

- $\rho_\lambda(t)$: (μ, γ)-amenable
  - $\rho_\lambda(t) + \frac{1}{2}\mu t^2$ is convex.
  - $\rho'_\lambda(t) = 0$ for $|t| > \gamma$.

Table: ℓ_2-error bound comparison with universal constant c_0 in sub-Gaussian tail bounds.

	Standard ℓ_1 $(h = 0)$	(μ, γ)-amenable	Trimmed ℓ_1 $(h = k)$
$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$	$\boxed{\frac{3c_0}{\kappa_l}\frac{\lambda_n\sqrt{k}}{2}}$	$\frac{c_0}{\kappa_l - \frac{3}{2}\mu}\frac{\lambda_n\sqrt{k}}{2}$	$\boxed{\frac{c_0}{\kappa_l}\frac{\lambda_n\sqrt{k}}{2}}$

- Trimmed ℓ_1 can achieve **three times smaller bound** than standard one.

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

- $\rho_\lambda(t)$: (μ, γ)-amenable
  - $\rho_\lambda(t) + \frac{1}{2}\mu t^2$ is convex.
  - $\rho'_\lambda(t) = 0$ for $|t| > \gamma$.

Table: ℓ_2-error bound comparison with universal constant c_0 in sub-Gaussian tail bounds.

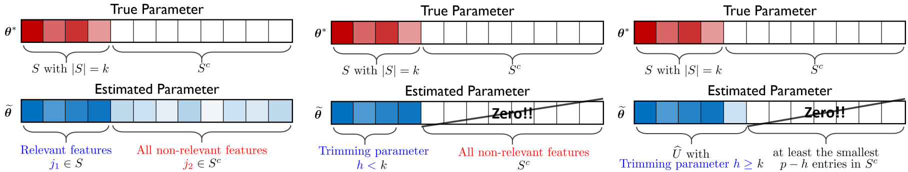
	Standard ℓ_1 $(h = 0)$	(μ, γ)-amenable	Trimmed ℓ_1 $(h = k)$
$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$	$\frac{3c_0}{\kappa_l}\frac{\lambda_n\sqrt{k}}{2}$	$\boxed{\frac{c_0}{\kappa_l - \frac{3}{2}\mu}\frac{\lambda_n\sqrt{k}}{2}}$	$\boxed{\frac{c_0}{\kappa_l}\frac{\lambda_n\sqrt{k}}{2}}$

- Also, **we have a smaller bound than non-convex regularizers** since (μ, γ)-amenable regularizers have (possibly large) μ in the denominator.

Statistical Analysis

Corollary 1: General ℓ_∞-error Bound for Linear Regression

- Consider a linear regression problem with sub-Gaussian error ϵ.
- Under standard conditions as in **Theorem 1** and incoherence condition on sample covariance, with high probability, *any* local minimum $\widetilde{\boldsymbol{\theta}}$ satisfies



1 The absolute value of relevant features is **always larger** than non-relevant features.

2 If we set the trimming parameter h smaller than true sparsity level k, all non-relevant parameters are estimated as zero.

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq c_1 \sqrt{\frac{\log p}{n}} + \lambda_n c_\infty$$

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq c_4 \sqrt{\frac{\log p}{n}} \left(\sqrt{k}/2 + \sqrt{k-h}\right)$$

3 If we set h larger than true sparsity level k, at least the smallest $p - h$ entries are estimated as zero.

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq c_1 \sqrt{\frac{\log p}{n}}$$

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{c_4}{2} \sqrt{\frac{h \log p}{n}}$$

$\boldsymbol{\theta}^*$: True parameter

$\widetilde{\boldsymbol{\theta}}$: Estimated parameter

Table of Contents

Optimization for Trimmed ℓ_1 Regularized Program

- For an optimization, we use our **second formulation** of trimmed regularization problem

$$\underset{\boldsymbol{\theta}\in\Omega,\boldsymbol{w}\in[0,1]^p}{\text{minimize}}\ \mathcal{F}(\boldsymbol{\theta},\boldsymbol{w}) := \mathcal{L}(\boldsymbol{\theta};\mathcal{D}) + \lambda\sum_{j=1}^{p} w_j|\theta_j| \quad \text{s.t.} \quad \mathbf{1}^T\boldsymbol{w} \geq p-h$$

Optimization for Trimmed ℓ_1 Regularized Program

- For an optimization, we use our **second formulation** of trimmed regularization problem

$$\underset{\boldsymbol{\theta}\in\Omega,\boldsymbol{w}\in[0,1]^p}{\text{minimize}}\ \mathcal{F}(\boldsymbol{\theta},\boldsymbol{w}) := \mathcal{L}(\boldsymbol{\theta};\mathcal{D}) + \lambda\sum_{j=1}^{p} w_j|\theta_j| \quad \text{s.t.} \quad \mathbf{1}^T\boldsymbol{w} \geq p-h$$

- We update $(\boldsymbol{\theta},\boldsymbol{w})$ in an alternating manner.

$$\boldsymbol{\theta}^{k+1} \leftarrow \mathrm{prox}_{\eta\lambda\mathcal{R}(\cdot,\boldsymbol{w}^k)}[\boldsymbol{\theta}^k - \eta\nabla_\theta\mathcal{L}(\boldsymbol{\theta}^k)]$$
$$\boldsymbol{w}^{k+1} \leftarrow \mathrm{proj}_{\mathcal{S}}[\boldsymbol{w}^k - \tau\boldsymbol{r}(\boldsymbol{\theta}^{k+1})]$$

  - Fixing $\boldsymbol{w}$, prox operator is weighted ℓ_1 norm.
  - By fixing $\boldsymbol{\theta}$, the objective function $\mathcal{F}$ is linear in $\boldsymbol{w}$.
  - $\mathrm{proj}_{\mathcal{S}}$ is a projection onto the constraint set $\mathcal{S} = \{\boldsymbol{w}\in[0,1]^p \mid \mathbf{1}^T\boldsymbol{w} = p-h\}$.

Optimization: Comparison with DC-based Approach

- Convergence history our algorithm vs. Algorithm 2 of (Khamaru & Wainwright, 2018)*.
  - Algorithm 2 of (Khamaru & Wainwright, 2018) is an optimization method for (non-convex and non-smooth) objective functions of the form **difference of convex functions** ($f := g + \phi - h$).
  - Trimmed regularized problem can be formulated as a DC.

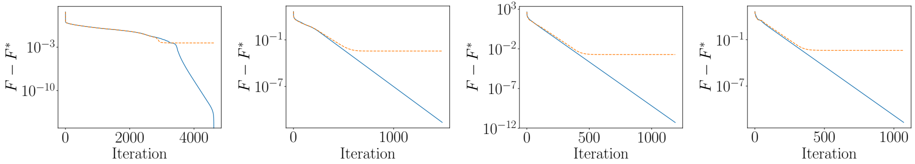
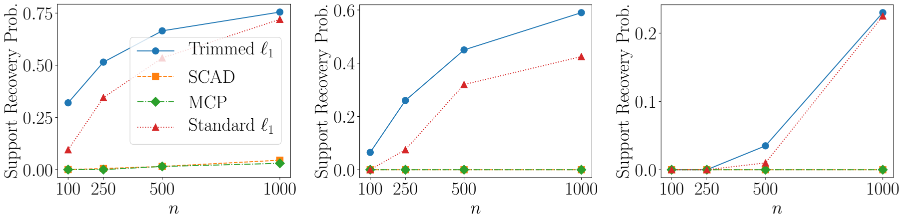


Figure: Algorithm comparison with $\lambda \in \{0.5, 5, 10, 20\}$.

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

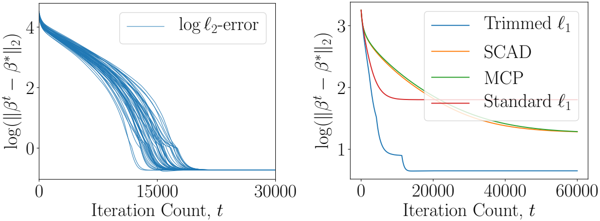
Table of Contents

- Scenario 1: Incoherence condition is satisfied



- Probability of successful support recovery for Trimmed Lasso, SCAD, MCP, and standard Lasso with $(p, k) = (128, 8), (256, 16), (512, 32)$.

- Scenario 1: Incoherence condition is satisfied

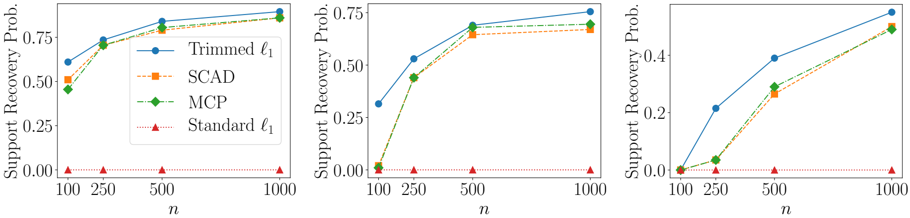


- **(Left)** 50 random initializations for a setting with $(n, p, k) = (160, 256, 16)$.
- **(Right)** $\log \ell_2$-error comparison.

Simulation Experiments

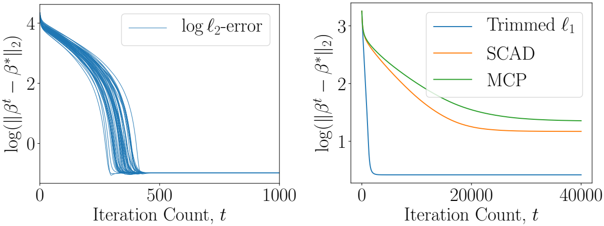
Nonincoherent Case: Support Recovery

- Scenario 2: Incoherence condition violated
  - Note that we need an incoherence condition in our Corollary 1.
  - Interestingly, the Trimmed Lasso outperforms all the other comparison regularizers even in this regime.



- Probability of successful support recovery for Trimmed Lasso, SCAD, MCP, and standard Lasso with $(p, k) = (128, 8), (256, 16), (512, 32)$.

- Scenario 2: Incoherence condition violated

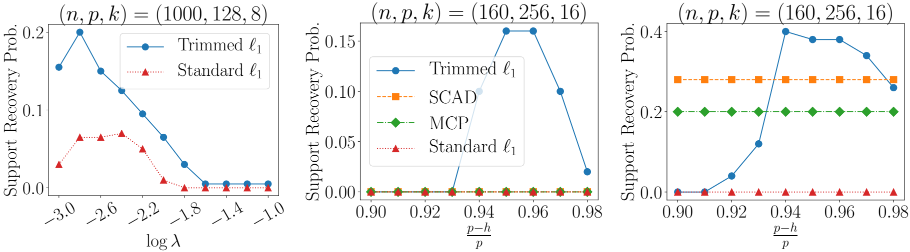


- **(Left)** 50 random initializations for a setting with $(n, p, k) = (160, 256, 16)$.
- **(Right)** $\log \ell_2$-error comparison.

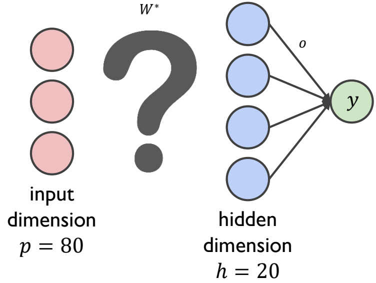
Simulation Experiments

Nonincoherent Case: Stationary

- Scenario 3
  - **(Left)** True signals and regularization parameter λ are both small (Small regime)
  - Investigating the choice of the trimming parameter h (**Middle**: Incoherent case, **Right**: Non-incoherent case).



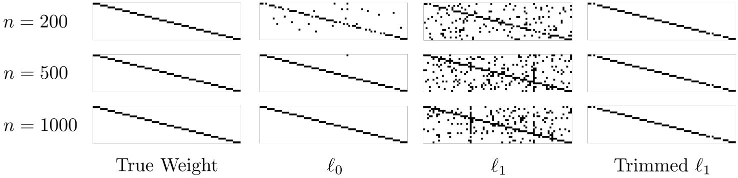
- We apply trimmed regularization to recover the weight structure of neural networks as parameter support recovery.
- Motivated by the recent work of Oymak (2018)*, we consider



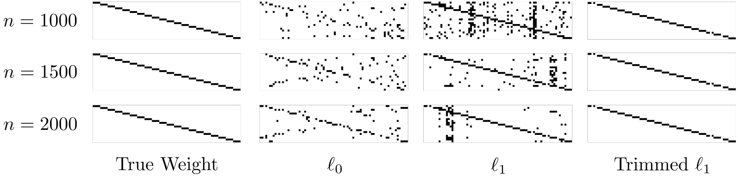
- The regression model, $y_i = \boldsymbol{o}^T \mathrm{ReLU}(\boldsymbol{W}^* \boldsymbol{x}_i)$ with $\boldsymbol{o} = \mathbf{1}$.
- Each hidden node is connected to only 4 input features.

* Samet Oymak. Learning Compact Neural Networks with Regularization. ICML, 2018.

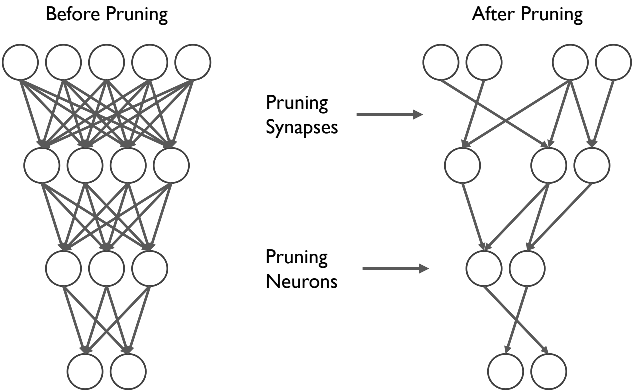
- With good initialization (**small perturbation** from true weight)



- With random initialization



Applications to Deep Learning 2: Pruning Deep Networks



- **Pruning neurons** is more computationally efficient than edge-wise pruning.

To encourage group sparsity on neural networks, we consider two cases:

- Neuron sparsity (for fully-connected layers)
  - Let $\boldsymbol{\theta}_l \in \mathbb{R}^{n_{\text{in}} \times n_{\text{out}}}$ be a weight parameter, then we can enforce group-wise sparsity via **Trimmed group ℓ_1 penalty** as

$$\mathcal{R}_l(\boldsymbol{\theta}_l, \boldsymbol{w}) = \lambda_l \sum_{j=1}^{n_{\text{in}}} w_j \sqrt{\theta_{j,1}^2 + \theta_{j,2}^2 + \cdots + \theta_{j,n_{\text{out}}}}$$

- Activation map sparsity (for convolutional layers)
  - Similarly, let $\boldsymbol{\theta}_l \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times H \times W}$ be a weight parameter, then

$$\mathcal{R}_l(\boldsymbol{\theta}, \boldsymbol{w}) = \lambda_l \sum_{j=1}^{C_{\text{out}}} w_j \sqrt{\sum_{m,n,k} \theta_{j,m,n,k}^2}$$

    for all possible indices (m, n, k).

with the constraint $\mathbf{1}^T \boldsymbol{w} = n_{\text{in}} - h_l$ or $C_{\text{out}} - h_l$ respectively.

- Comparison with vanilla group ℓ_1 penalty vs. Trimmed group ℓ_1 penalty on LeNet-300-100 structure

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**

- Most modern algorithms for network pruning are based on **Bayesian variational framework**. We propose a Bayesian neural network with Trimmed ℓ_1 regularization regarding only $\boldsymbol{\theta}$ as Bayesian.

- Most modern algorithms for network pruning are based on **Bayesian variational framework**. We propose a Bayesian neural network with Trimmed ℓ_1 regularization regarding only $\boldsymbol{\theta}$ as Bayesian.

- By relationship between **Bayesian neural networks** and **variational dropout**, we choose $q_{\boldsymbol{\theta},\boldsymbol{\alpha}}(\theta_{i,j}) = \mathcal{N}(\phi_{i,j}, \alpha_{i,j}\phi_{i,j}^2)$ as a variational distribution.

- Most modern algorithms for network pruning are based on **Bayesian variational framework**. We propose a Bayesian neural network with Trimmed ℓ_1 regularization regarding only $\boldsymbol{\theta}$ as Bayesian.
- By relationship between **Bayesian neural networks** and **variational dropout**, we choose $q_{\boldsymbol{\theta},\boldsymbol{\alpha}}(\theta_{i,j}) = \mathcal{N}(\phi_{i,j}, \alpha_{i,j}\phi_{i,j}^2)$ as a variational distribution.
- Combined with Trimmed ℓ_1 regularization, the objective is

$$\underbrace{\mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\alpha}}(\boldsymbol{\theta})}\Big[-\mathcal{L}(\mathcal{W};\mathcal{D})\Big] + \mathbb{KL}(q_{\boldsymbol{\phi},\boldsymbol{\alpha}}(\mathcal{W})\|p(\mathcal{W}))}_{\text{ELBO}} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi},\boldsymbol{\alpha}}(\boldsymbol{\theta})}\Big[\sum_{l=1}^{L+1}\lambda_l\mathcal{R}_l(\boldsymbol{\theta}_l,\boldsymbol{w}_l)\Big]}_{\text{Expected Trimmed group } \ell_1 \text{ penalty}}$$

- With Bayesian extensions on LeNet-300-100
  - We compare with a smoothed ℓ_0-norm under Bayesian variational framework proposed by Louizos et al. (2018)*

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

- With Bayesian extensions on LeNet-5-Caffe

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

* Louizos et al. Learning Sparse Neural Networks through ℓ_0 Regularization. ICLR, 2018

Concluding Remarks

- **High-dimensional M-estimators with Trimmed ℓ_1 penalty**: Alleviate the bias incurred by the vanilla ℓ_1 penalty by leaving the h largest parameter entries penalty-free.
- **Theoretical Results** on support recovery and ℓ_2-error hold for any local optima and are competitive with other non-convex regularizers.
- **Simulation experiments** demonstrated the value of approach compared to Lasso and non-convex penalties.
- Future work:
  - Trimming for other standard regularizers beyond sparsity
  - Bypassing incoherence condition in corollaries
  - More experiments and theories when RSC does not hold
  - Investigating the use of trimmed regularization in deep models.

THANK YOU!
Any Questions?

Poster Session at Pacific Ballroom #186
6:30pm – 9:00pm