Review:
A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers

Jiaxin Hu

June 9, 2021

1 Abstract

This paper organizes the convergence results for the penalized M-estimators of the high-dimensional models, which satisfies

$$\hat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \mathbb{R}^p} \left\{ \mathcal{L}(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \right\}, \tag{1}$$

where Z_1^n is the n i.i.d. observations, θ is the p-dimensional parameter of interests, $\mathcal{L}$ is the loss function, $\mathcal{R}$ is the regularizer, and λ_n is the tuning parameter. Let θ^* denote the true parameter. The unified conclusion requires only two key properties of the objective functions:

1. **Decomposable Regularizer** respect to the model subspace $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$, i.e.,

   $$\mathcal{R}(\theta + \gamma) = \mathcal{R}(\theta) + \mathcal{R}(\gamma),$$

   for all $\theta \in \mathcal{M}$ and $\gamma \in \bar{\mathcal{M}}^{\perp}$. A "small" model subspace $\mathcal{M}$ would bring a better convergence rate. This assumption leads to a key consequence of the space of the estimator (1).

   Lemma 1. *Suppose $\mathcal{L}$ is convex and differentiable. The optimizer (1) with strictly positive regularization parameter satisfying*

   $$\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*; Z_1^n)). \tag{2}$$

   Then, for any pair $(\mathcal{M}, \bar{\mathcal{M}}^{\perp})$ over which $\mathcal{R}$ is decomposable, the error $\hat{\Delta} = \hat{\theta}_{\lambda_n} - \theta^*$ belongs to the set

   $$\mathbb{C}(\mathcal{M}, \bar{\mathcal{M}}^{\perp}; \theta^*) := \left\{ \Delta \in \mathbb{R}^p | \mathcal{R}(\Delta_{\bar{\mathcal{M}}^{\perp}}) \leq 3\mathcal{R}(\Delta_{\bar{\mathcal{M}}}) + 4\mathcal{R}(\theta^*_{\mathcal{M}^{\perp}}) \right\}.$$

   Particularly, if $\theta^* \in \mathcal{M}$, then the last term $\mathcal{R}(\theta^*_{\mathcal{M}^{\perp}}) = 0$.

2. **Restricted strong convexity (RSC) of the loss $\mathcal{L}$**, i.e.,

   $$\delta\mathcal{L}(\Delta, \theta^*) := \mathcal{L}(\theta^* + \Delta) - \mathcal{L}(\theta^*) - \langle \nabla \mathcal{L}(\theta^*), \Delta \rangle \geq \kappa_{\mathcal{L}} \|\Delta\|^2 + \tau^2_{\mathcal{L}}(\theta^*), \tag{3}$$

   for all $\Delta \in \mathbb{C}(\mathcal{M}, \bar{\mathcal{M}}^{\perp}; \theta^*)$, and $\kappa_{\mathcal{L}} > 0$. Particularly, if $\theta^* \in \mathcal{M}$, the last term is usually $\tau^2_{\mathcal{L}}(\theta^*) = 0$. In many loss functions, we usually consider the error bound

   $$\delta\mathcal{L}(\Delta, \theta^*) \geq \kappa_1 \|\Delta\|^2 - \kappa_2 g(n,p) \mathcal{R}^2(\Delta),$$

for all $\|\Delta\| \leq 1$, and $g(n,p)$ is a function usually increasing in p and decreasing in n. To obtain the RSC in form (3), we define the subspace compatibility constant, denoted $\Psi(\mathcal{M})$, to bridge the regularizer and the error norm, where

$$\Psi(\mathcal{M}) := \sup_{u \in \mathcal{M}/\{0\}} \frac{\mathcal{R}(u)}{\|u\|}.$$

With the definition of $\mathbb{C}$, we have $\mathcal{R}(\Delta) \leq 4\Psi(\bar{\mathcal{M}}) \|\Delta\|$ and

$$\delta\mathcal{L}(\Delta, \theta^*) \geq \left\{\kappa_1 - 16\kappa_2 \Psi^2(\bar{\mathcal{M}}) g(n,p)\right\} \|\Delta\|^2.$$

Note that the loss function with RSC property may not be strongly convex on all the directions.

With the mentioned assumptions, we obtain the unified convergence conclusion.

Theorem 1.1. *Under the conditions of **Decomposability** and **RSC**, consider the problem (1) with $\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}(\theta^*; Z_1^n))$. Then, any optimal solution $\hat{\theta}_{\lambda_n}$ satisfies the bound*

$$\left\|\hat{\theta}_{\lambda_n} - \theta^*\right\|^2 \leq 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}^2}\Psi^2(\bar{\mathcal{M}}) + \frac{\lambda_n}{\kappa_{\mathcal{L}}}\left\{2\tau_{\mathcal{L}}^2(\theta^*) + 4\mathcal{R}(\theta^*_{\mathcal{M}^\perp})\right\}.$$

Corollary 1. Suppose that, in addition to the conditions for Theorem 1.1, the unknown θ^* belongs to the subspace $\mathcal{M}$ and RSC holds with $\tau_{\mathcal{L}}^2(\theta^*) = 0$. Then, any optimal solution $\hat{\theta}_{\lambda_n}$ satisfies the bounds

$$\left\|\hat{\theta}_{\lambda_n} - \theta^*\right\| \leq 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}}\Psi^2(\bar{\mathcal{M}}),$$

and

$$\mathcal{R}(\hat{\theta}_{\lambda_n} - \theta^*) \leq 12\frac{\lambda_n}{\kappa_{\mathcal{L}}}\Psi^2(\bar{\mathcal{M}}).$$

2 Application to precision matrix estimation

Here we consider the simple penalized precision matrix estimation problem. Let $S_n \in \mathbb{R}^{p\times p}$ denote the sample covariance matrix with n observations, Σ denote the true covariance matrix, and $\Theta^* = \Sigma^{-1}$ denote the true precision matrix. Consider the penalized optimization problem

$$\hat{\Theta}_{\lambda_n} = \underset{\Theta\in\mathbb{R}^{p\times p}}{\arg\min}\left\{\langle\Theta, S_n\rangle - \log\det(\Theta) + \lambda_n \|\Theta\|_1\right\}, \tag{4}$$

which is in the class of the estimation problem (1) with $\mathcal{L}(\Theta, S_n) = \langle\Theta, S_n\rangle - \log\det(\Theta)$ and $\mathcal{R}(\Theta) = \|\Theta\|_1$. By (Negahban et al., 2012), we obtain the convergence rate for the precision estimation.

Our definition of s equals (s + p) in Guo et al

Corollary 2. Suppose the true precision matrix has s sparsity, i.e., $\|\Theta^*\|_0 = s$. With $\lambda_n \geq C\sqrt{\log p/n}$, with high probability tends to 1, the optimal solution to the optimization (4) satisfies the bounds

$$\left\|\hat{\Theta}_{\lambda_n} - \Theta\right\|_F^2 \leq \frac{16C_1^2\tau^4 s\log p}{n},$$

and

$$\left\|\hat{\Theta}_{\lambda_n} - \Theta^*\right\|_1 \leq C_2\sqrt{\frac{\log p}{n}}\tau^2 s,$$

for some constant C_1, C_2, with high probability.

Proof. To obtain the convergence rate, we need to verify the decomposability of the ℓ_1 norm with a particular model subspace, the RSC parameters for $\mathcal{L}(\Theta, S_n)$, and the valid region for λ_n.

1. **Decomposability.** Note that the matrix ℓ_1 norm is equal to the vector ℓ_1 norm for the vectorized matrix. By Example 1 in (Negahban et al., 2012), the vector ℓ_1 norm is decomposable with model subspace $\mathcal{M}(T)$, where

$$\mathcal{M} = \left\{\Theta \in \mathbb{R}^{p\times p} | \Theta_{ij} = 0, (i,j) \notin T\right\}, \quad T = \left\{(i,j) | \Theta^*_{ij} \neq 0\right\}, \quad |T| = s.$$

Thus, the term $\mathcal{R}(\Theta^*_{\mathcal{M}^\perp}) = 0$. Besides, with subspace $\mathcal{M}$, the subspace compatibility constant with respect to the pair $(\|\cdot\|_1, \|\cdot\|_F)$ is

$$\Psi(\mathcal{M}) = \sup_{A \in \mathcal{M}/\{0\}} \frac{\|A\|_1}{\|A\|_F} = \sqrt{s}.$$

2. **RSC.** Let $\Delta = \hat{\Theta}_{\lambda_n} - \Theta^*$. By the definition of Taylor series error in (3) , we have

$$\begin{aligned}
\delta\mathcal{L}(\Delta, \Theta^*) &= \langle \Delta, S_n \rangle - \log\det(\Theta^* + \Delta) + \log\det(\Theta^*) - \langle \Delta, S_n - \Sigma \rangle \\
&= \mathrm{vec}(\Delta)^T \left\{ \int_0^1 (1-v)(\Theta^* + v\Delta)^{-1} \otimes (\Theta^* + v\Delta)^{-1} dv \right\} \mathrm{vec}(\Delta) \\
&\geq \frac{1}{4\tau^2} \|\Delta\|_F^2 ,
\end{aligned}$$

where τ is the largest singular value of Θ^*, and the last equation and inequality follows by the Lemma A1 in (Guo et al., 2011).Thus, the loss function $\mathcal{L}$ satisfies the RSC with $\kappa_{\mathcal{L}} = \frac{1}{4\tau^2}$.

not the true reason.

3. **Valid range for λ_n.** The final step is to verify the valid the range for λ by (2). Note that, we do not know the true parameter Θ^*, and thereof the convergence rate with reasonable choice of λ is valid with high probability whereas the main theorem 1.1 is deterministic.

the randomness comes from ``green''

The dual norm is $\mathcal{R}^*(\Theta) = \|\Theta\|_{\max}$, which is identical to the maximal absolute value in Θ. This follows by the fact that vector max norm $\|\cdot\|_\infty$ is the dual norm of the vector ℓ_1 norm. Therefore, we have

$$\lambda \geq 2 \|(S_n - \Sigma)\|_{\max} ,$$

where

$$\|(S_n - \Sigma)\|_{\max} \leq C\sqrt{\frac{\log p}{n}},$$

This bound holds only in ``high probability''
this is the reason of randomness.

with high probability by the Lemma 1 of (Rothman et al., 2009).

In fact, in the main paper, Corollary 2
is also a high-probability result. Same reason
as here.

Therefore, we obtain the convergence rate

$$\left\|\hat{\Theta}_{\lambda_n} - \Theta\right\|_F^2 \leq \frac{16C^2\tau^2 s \log p}{n},$$

and

$$\left\|\hat{\Theta}_{\lambda_n} - \Theta^*\right\|_1 \leq C_2\sqrt{\frac{\log p}{n}}\tau^2 s,$$

for some constant C_1, C_2, with high probability.

□

Remark 1. The conclusion of in Corollary 2 is identical to the previous results in (Guo et al., 2011), except the upper bound requirement for the λ_n.

References

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

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

Rothman, A. J., Levina, E., and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186.