Sparse + Group-Sparse Dirty Models: Statistical Guarantees without Unreasonable Conditions and a Case for Non-Convexity

Eunho Yang [1 2] **Aurélie C. Lozano** [3]

Abstract

Imposing sparse + group-sparse superposition structures in high-dimensional parameter estimation is known to provide flexible regularization that is more realistic for many real-world problems. For example, such a superposition enables partially-shared support sets in multi-task learning, thereby striking the right balance between parameter overlap across tasks and task specificity. Existing theoretical results on estimation consistency, however, are problematic as they require too stringent an assumption: the incoherence between sparse and group-sparse superposed components. In this paper, we fill the gap between the practical success and suboptimal analysis of sparse + group-sparse models, by providing the first consistency results that do not require unrealistic assumptions. We also study non-convex counterparts of sparse + group-sparse models. Interestingly, we show that these are guaranteed to recover the true support set under much milder conditions and with smaller sample size than convex models, which might be critical in practical applications as illustrated by our experiments.

1. Introduction

We consider high-dimensional statistical models where the ambient dimension p is much larger than the number of observations n. Under such high-dimensional scaling, it is still possible to obtain consistent estimators by imposing low-dimensional structural constraints upon the statistical models, such as sparsity (e.g. in compressed sensing (Baraniuk, 2007) and Lasso (Tibshirani, 1996)), low-rank structure (Recht et al., 2007; Negahban & Wainwright, 2010), sparse graphical model structure (Friedman et al., 2007; Ravikumar et al., 2008), and sparse additive structure for non-parametric models (Ravikumar et al., 2009). A widely used approach to structured learning is via specific regularization functions. For instance, ℓ_1-regularization is employed for sparse models (Tibshirani, 1996), ℓ_1/ℓ_q norms for group sparsity (Yuan & Lin, 2006), and nuclear norm for low-rank matrix-structured models (Candès & Tao, 2010). Much attention has been devoted to the study of these structured norms and their theoretical properties.

Such a "clean" regularization approach, however, might be too stringent in practice. For instance in linear regression, a blend of element-wise sparsity and group-sparsity might be more appropriate than a purely sparse or purely group-sparse solution. In multitask learning, while some parameters might be shared across tasks, others might only be relevant to a subset of tasks or a single task. To overcome this limitation, a line of work on so-called *dirty models* has emerged, which addresses this caveat by "mixing and matching" different structures. One basic approach consists in decomposing the model parameters as a sum of two components, each penalized separately: one component captures the common structure across tasks and the other task-specific characteristics (Jalali et al., 2010; Gong et al., 2012). For instance the dirty model in Jalali et al. (2010) employs $\ell_{1,1}$ and $\ell_{1,\infty}$ regularizers to the two components. Chandrasekaran et al. (2011) consider the problem of recovering unknown low-rank and sparse matrices, given the sum of their sum, with application such as optical imaging systems. Robust principal component analysis and related extensions (Candès et al., 2011; Agarwal et al., 2012; Hsu et al., 2011) estimate a covariance matrix that is the sum of a low-rank matrix and a structured (e.g. sparse, column sparse) matrix.

A general framework for studying dirty models was recently proposed in Yang & Ravikumar (2013), which bridges and extends several analyses for specific pairs of superposition structures and specific statistical models (e.g., Jalali et al. (2010); Chandrasekaran et al. (2011); Candès et al. (2011); Agarwal et al. (2012); Hsu et al.

[1]School of Computing, KAIST, Daejeon, South Korea [2]AItrics, Seoul, South Korea [3]IBM T.J. Watson Research Center, Yorktown Heights, NY, USA. Correspondence to: Eunho Yang <eunhoy@kaist.ac.kr>, Aurelie C. Lozano <aclozano@us.ibm.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

(2011)) Specifically, this framework applies to a general class of M-estimators employing a so-called *hybrid* regularization function, which is the infimal convolution of weighted regularization functions, one for each structural component. This formulation is equivalent to an M-estimator that combines a loss function applied to the sum of multiple parameter vectors (one per structural component) and a weighted sum of regularization functions (one per parameter vector).

For the *sparse + group sparse* decomposition, however existing analyses are highly problematic. The key weakness is that they require some form of *structural incoherence condition* which captures the interaction between the different structured components. While such a structural incoherence is a reasonable assumption for e.g. sparse + low rank superposition, it is what too stringent for the sparse+group sparse case because the two structures are completely coherent for this case! This yields a key motivating question for this paper: *Under the sparse + group sparse setting, can we bypass structural incoherence conditions and yet obtain tight error bounds?*

In this paper we provide a positive answer by developing a novel proof technique. Prior analyses require 'local' restricted strong convexity conditions (RSC): one condition for the sparse component and one for the group sparse component. The use of structural incoherence between sparse and group sparse components in then needed to show 'global' RSC for the vector concatenating sparse and group sparse components. To avoid the need for structural incoherence, we use RSC in the *summed space* directly (namely for the summed sparse + group-sparse structure). However, this brings in a new issue: in this case, the dirty regularizer for the parameter vector is *not* decomposable. To circumvent this issue, our key ingredient is to introduce "surrogate" sparse and group sparse components depending on our estimators such that i) their sum equals the sum of the true parameter components and ii) corresponding error vectors are decomposable even though the regularizer itself is not decomposable. Using the decomposability of error vectors, we are then able to show ℓ_2 consistency for general loss functions.

As an additional key contribution of this paper, we consider the extension of sparse+group sparse dirty models to non-convex regularizers, and show their ℓ_∞ consistency. Interestingly, these models are guaranteed to recover the true support set under much milder conditions and with smaller sample size than convex models. In particular, our ℓ_∞ consistency results require *neither* incoherence in the loss function *nor* structural incoherence between sparse and group sparse parameters. We illustrate the practical impact of this superior theoretical results with simulation experiments.

The remainder of this paper is organized as follows. In Section 2 we review sparse+group-sparse dirty models with convex penalties and introduce their non-convex counterparts. In Section 3 we discuss the incoherence assumption required by prior analyses and explain why such an assumption is unreasonable. Section 4 introduces the key ingredient of our novel proof technique. Section 5 presents the convergence bounds for models with convex penalties. Those for non-convex penalties are stated in Section 6. Finally, simulation experiments are provided in Section 7 to illustrate the remarkable practical advantage of non-convex penalties, agreeing with their superior convergence rates.

2. Sparse + Group-Sparse Dirty Models: Setup and Formulations

Consider a data collection $Z = \{Z_1, \ldots, Z_n\}$, where each element is drawn independently from distribution $\mathbb{P}$, and a loss function $\mathcal{L}(\cdot\,; Z) : \Omega \to \mathbb{R}$ where $\mathcal{L}(\boldsymbol{\theta}\,; Z)$ measures the goodness of fit of parameter $\boldsymbol{\theta} \in \Omega$ to the given data collection Z. Typically $\Omega = \mathbb{R}^p$ (parameters are vectors) or $\mathbb{R}^{p\times r}$ (parameters are matrices). Assume there are some *known* groups $\mathcal{G} = \{G_1, \ldots, G_q\}$ that partition the parameter index set: $G_i \cap G_j = \phi$ and $\cup_{g=1}^{q} G_g = \{1, \ldots, p\}$.

We aim at recovering parameter $\boldsymbol{\theta}^*$ which is the unique minimizer of the population risk: $\boldsymbol{\theta}^* := \operatorname{argmin}_{\boldsymbol{\theta}\in\Omega} \mathbb{E}_Z[\mathcal{L}(\boldsymbol{\theta}; Z)]$ in cases where

$$\boldsymbol{\theta}^* = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*, \tag{1}$$

where $\boldsymbol{\alpha}^*$ is a *sparse* component and $\boldsymbol{\beta}^*$ is a *group-sparse* component obeying the group structure $\mathcal{G}$. For that purpose, we focus on regularized M-estimators under a *dirty* learning setting that combines sparsity and group-sparsity. We consider both convex and non-convex regularizers as follows.

2.1. Dirty models with convex regularizers

We focus on regularized M-estimators of the form

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\text{minimize}}\ \mathcal{L}(\boldsymbol{\alpha}+\boldsymbol{\beta}; Z) + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_{1,a}, \tag{2}$$

where the loss function $\mathcal{L}(\cdot\,; Z)$ is possibly non-convex. Here, given known parameter groups $\mathcal{G} = \{G_1, G_2, \ldots, G_q\}$, the group regularizer is defined as $\|\boldsymbol{\beta}\|_{1,a} := \sum_{t=1}^{q} \|\boldsymbol{\beta}_{G_t}\|_a$ for $a \geq 2$, where $\boldsymbol{\beta}_{G_t}$ denotes the parameter subset in group G_t. The constant a determines how the elements within each group are combined.

We provide examples for the popular settings of linear regression and inverse covariance estimation.

Linear regression. Consider the standard linear model $\boldsymbol{y} = X\boldsymbol{\theta}^* + \boldsymbol{w}$ where $\boldsymbol{y} \in \mathbb{R}^n$ is the observation vector, $\boldsymbol{\theta}^*$

is the true regression parameter which is the sum of sparse $\boldsymbol{\alpha}^*$ and group sparse $\boldsymbol{\beta}^*$, $X \in \mathbb{R}^{n\times p}$ is the design matrix, and $\boldsymbol{w} \in \mathbb{R}^n$ is the observation noise. The "dirty" regularized least squares solves

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \ \frac{1}{2n}\|\boldsymbol{y} - X\big(\boldsymbol{\alpha}+\boldsymbol{\beta}\big)\|_2^2 + \lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_{1,a} \tag{3}$$

where groups are defined within a (single) parameter vector space via $\boldsymbol{\beta}$. The formulation can be seamlessly extended to cover the dirty multitask learning setting of Jalali et al. (2010):

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^{p\times m}}{\text{minimize}} \ \sum_{k=1}^{m}\frac{1}{2n}\big\|\boldsymbol{y}^{(k)} - X^{(k)}\big([\boldsymbol{\alpha}+\boldsymbol{\beta}]_{(\bullet,k)}\big)\big\|_2^2 \tag{4}$$
$$+\lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_{1,\infty}$$

where we have m related tasks in columns: $\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathbb{R}^{p\times m}$, and the groups can be defined across tasks in rows. E.g. for predictor j, $\boldsymbol{\beta}_{(j,1)},\ldots,\boldsymbol{\beta}_{(j,m)}$ belong to the same group. Here, $[\boldsymbol{\alpha}+\boldsymbol{\beta}]_{(\bullet,k)}$ indicates k-th column of matrix input $\boldsymbol{\alpha}+\boldsymbol{\beta}$.

Graphical Model Estimation. Another key example is a modified graphical Lasso where the goal is to estimate the structure of the underlying graphs representing conditional independences across variables. Assume that there are some known set of edge groups and that the true parameter Θ^* has only a small number of active edge groups plus some individual edges. To recover Θ^* we solve

$$\underset{S+B\succ 0}{\text{minimize}} \ \text{trace}\big((S+B)\,\widehat{\Sigma}\big) - \log\det(S+B) \tag{5}$$
$$+\lambda_1\|S\|_1 + \lambda_2\|B\|_{1,a}$$

where $\widehat{\Sigma}$ is the sample covariance matrix and regularizers are applied to off-diagonal entries of S and B. As done in (4) for the linear model, the formulation (5) can be seamlessly extended to the multitask setting where we wish to estimate multiple precision matrices jointly, encouraging similar structure while allowing for some discrepancy across them. This estimator is discussed in Hara & Washio (2013).

Equivalent Program. As shown in Yang & Ravikumar (2013), the formulation (2) can be rewritten as:

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \ \mathcal{L}(\boldsymbol{\theta};Z) + \|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}} \tag{6}$$

where $\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}$ is the infimal convolution of two regularizers

$$\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}} := \inf_{\boldsymbol{\alpha},\boldsymbol{\beta}}\Big\{\lambda_1\|\boldsymbol{\alpha}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_{1,a} \ : \ \boldsymbol{\alpha}+\boldsymbol{\beta}=\boldsymbol{\theta}\Big\}. \tag{7}$$

It is known that $\|\cdot\|_{\boldsymbol{\lambda}}$ is a norm and its dual is defined as $\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}^* := \max\{\|\boldsymbol{\theta}\|_\infty/\lambda_1, \|\boldsymbol{\theta}\|_{\infty,a^*}/\lambda_2\}$ where $1/a + 1/a^* = 1$ so that $\|\cdot\|_{\infty,a^*}$ is the dual norm of $\|\cdot\|_{1,a}$ (see Yang & Ravikumar (2013) for details).

2.2. Dirty models with non-convex regularizers

In this paper, we introduce and study estimators of the form

$$\underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\text{minimize}} \ \mathcal{L}(\boldsymbol{\alpha}+\boldsymbol{\beta}) + \rho_{\lambda_1}\big(\boldsymbol{\alpha}\big) + \phi_{\lambda_2,a}\big(\boldsymbol{\beta}\big). \tag{8}$$

Here $\rho_{\lambda_1}(\cdot)$ is any regularizer inducing sparsity beyond the ℓ_1-norm (note that the notation encapsulates the regularization parameter λ_1 itself within the regularizer) satisfying the following conditions (Loh & Wainwright, 2014):

(C1) $\rho_{\lambda_1}(0) = 0$ and is symmetric. For $t > 0$, $\rho_{\lambda_1}(t)$ is non-decreasing but $\rho_{\lambda_1}(t)/t$ is non-increasing in t. Besides, $\rho_{\lambda_1}(t)$ is differentiable for $t \neq 0$ with $\lim_{t\to 0^+}\rho'_{\lambda_1}(t) = \lambda_1$, and is $\rho_{\lambda_1}(t) + \frac{\mu}{2}t^2$ is convex for some $\mu > 0$.

(C2) There exists some scalar $\gamma \in (0,\infty)$ such that $\rho'_{\lambda_1}(t) = 0$ when $t \geq \gamma\lambda_1$.

Following the notation of Loh & Wainwright (2014), we call $\rho_{\lambda_1}(\cdot)$ μ-amenable if it satisfies (C1) and (μ,γ)-amenable if it additionally satisfies (C2). The popular non-convex regularizers SCAD (Fan & Li, 2001) and MCP (Zhang, 2010) are both (μ,γ)-amenable (Loh & Wainwright, 2014).

The regularizer $\phi_{\lambda_2,a}(\cdot)$ a non-convex counterpart of the group regularizer $\lambda_2\|\cdot\|_{1,a}$ employed in (2) where we use $\rho_{\lambda_2}(\cdot)$ instead of $\lambda_2\|\cdot\|_1$, over groups:

$$\phi_{\lambda_2,a}(\boldsymbol{\beta}) := \rho_{\lambda_2}(\mathcal{G}(\boldsymbol{\beta}))$$

where $\mathcal{G}(\boldsymbol{\beta}) := (\|\boldsymbol{\beta}_{G_1}\|_a,\ldots,\|\boldsymbol{\beta}_{G_q}\|_a)^\top$. Example of non-convex regularizers include the Group-SCAD and Group-MCP penalties where SCAD and MCP penalties are respectively used on the norm of each group.

Remarkably, the proof techniques developed in this paper make it possible to provide not only ℓ_2-error bounds under milder conditions than prior work on convex problem (2), but also support set recovery guarantees for non-convex one (8). In fact we shall see that dirty models with non-convex regularizers (8) enjoy strictly better statistical guarantees than their convex counterpart (2), with practical consequences.

3. Structural Incoherence: essential in prior work, yet an unreasonable assumption

As our starting point, we focus on the case of convex dirty models in (2) or equivalently in (6). A key ingredient for showing statistical guarantees of regularized M-estimators is the *decomposability* of regularizer (Negahban et al., 2012). However, considering the form of regularizer in (6), it is not obvious to find the model space and its orthogonal complement with which we could directly derive

error bounds with optimal rates. To circumvent this problem, Yang & Ravikumar (2013) utilize the decomposability of each component separately, but this requires restricted strong convexity (RSC) to hold *jointly* for all component parameters. In order to have the "joint" RSC property from "local" RSC with respect to each individual component, Yang & Ravikumar (2013) assume a *structural incoherence* condition. Even if the loss function is strongly convex with respect to each component, such incoherence across components is essential for the joint RSC due to the linearity across components. To see this more clearly, suppose we have the function z^2 for $z \in \mathbb{R}$, which is obviously strongly convex. If we assume, however, that z is the sum of two components $x, y \in \mathbb{R}$, then one can immediately see that $(x+y)^2$ is not strongly convex jointly in x and y because x and y are completely coherent in this one dimensional example.

The problem is that the structural incoherence condition for the $\ell_1+\ell_{1,a}$ setting is way too restrictive because the sparse and the group-sparse structures essentially share the same model and its orthogonal spaces[1]. In order to see this, we consider the popular linear model setting (3) for example. Let $\boldsymbol{s}^*$ (and $\boldsymbol{b}^*$) be the support set of true sparse (group-sparse) component and $\boldsymbol{s}^c$ be its complement. Furthermore, $[\frac{1}{n}X^\top X]_{(\boldsymbol{s}^c \cap \boldsymbol{b}^*)}$ denotes the projection of the sample covariance onto $\boldsymbol{s}^c \cap \boldsymbol{b}^*$-coordinate space ($j$-th coordinate becomes zero if $j \notin \boldsymbol{s}^c \cap \boldsymbol{b}^*$). Projections on other spaces are defined similarly. Then, the structural incoherence condition for joint RSC can be reduced as: for all $(\boldsymbol{s}, \boldsymbol{b}) \in \{(\boldsymbol{s}^c, \boldsymbol{b}^*), (\boldsymbol{s}^*, \boldsymbol{b}^c), (\boldsymbol{s}^c, \boldsymbol{b}^c)\}$,

$$\sigma_{\max}\big([\tfrac{1}{n}X^\top X]_{(\boldsymbol{s}\cap\boldsymbol{b})}\big) \leq C\kappa_1 \tag{9}$$

where $\sigma_{\max}(\cdot)$ is the maximum singular value of a matrix, κ_1 is the curvature of (restricted) eigenvalue condition, and C is some fixed constant. Informally, this condition requires the maximum singular value of sample covariance (modulo the projection onto the true model and its orthogonal space) to be smaller than its minimum singular value (Note that for linear models, the curvature parameter of the eigenvalue condition is related to the minimum singular value of the sample covariance). This condition can be easily shown to fail in many cases. For instance consider the popular setting where the design matrix X is a set of samples from Gaussian ensemble with covariance Σ, and the true parameter is the sum of group sparse + a *single* nonzero component as depicted in Figure 1. Then, the incoherence condition in (9) implies $\max_{i,j} |[\frac{1}{n}X^\top X]_{ij}| \leq 1/128\sigma_{\min}(\Sigma)$, which can be easily violated in many natural setting of Σ because the minimum eigenvalue of Σ is smaller than the maximum element of Σ by the Rayleigh quotient.

[1]Note that the sparse + group sparse setting is outstanding. The structural incohence assumption makes sense in other dirty models settings, e.g. sparse + low rank dirty models.

3	1	1
1	2	1
0	0	0
0	1	0
0	0	0

Figure 1. Example illustrating why the incoherence condition required by previous work fails to hold.

This naturally leads to the following question:

Can we provide tight error bounds for the problem (2) not requiring the joint RSC across individual structures and hence bypassing the incoherence condition?

4. Our key strategy: Constructing surrogate components that are always decomposable.

In order to address the above question, our key proof technique is to establish the decomposability between two components of error vectors, by making the target components dependent of our estimation. Consider arbitrary target parameter $\boldsymbol{\theta}^*$ such that $\boldsymbol{\theta}^* = \boldsymbol{\alpha}^* + \boldsymbol{\beta}^*$. Note that we do not impose additional constraints on defining the sparse component $\boldsymbol{\alpha}^*$ and the group sparse component $\boldsymbol{\beta}^*$, hence the possible combination of $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ is not unique. As we will see later, we provide estimation error bounds that depend on the selection of $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$–more precisely on the sparsity level of $\boldsymbol{\alpha}^*$ and the group sparsity level of $\boldsymbol{\beta}^*$. In that sense our theorems provide sets of estimation bounds. However, it is important to note that we still do not need to worry about the identifiability between structures, because we only care about the ℓ_2 and ℓ_∞ error rates of the final or "summed" estimator (we do not recover (nor care about) the individual components).

Suppose we compute $\widetilde{\boldsymbol{\theta}}$ from the program (6) where $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\beta}}$ are minimizing its dirty regularizer (7). Then, rather than directly deriving error bounds of $\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ from $\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*$ and $\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$, which are *not* decomposable, we introduce an additional set of vectors, $\bar{\boldsymbol{\alpha}}$, $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\theta}}$ from the following rules:

1. If $\boldsymbol{\alpha}^*_j = \boldsymbol{\beta}^*_j = 0$, then $\bar{\boldsymbol{\alpha}}_j = \bar{\boldsymbol{\beta}}_j := 0$.
2. If $\boldsymbol{\alpha}^*_j \neq 0$ and $\boldsymbol{\beta}^*_j = 0$, then $\bar{\boldsymbol{\beta}}_j := \widetilde{\boldsymbol{\beta}}_j$, and $\bar{\boldsymbol{\alpha}}_j := \boldsymbol{\theta}^*_j - \widetilde{\boldsymbol{\beta}}_j$.
3. If $\boldsymbol{\beta}^*_j \neq 0$, then $\bar{\boldsymbol{\alpha}}_j := \widetilde{\boldsymbol{\alpha}}_j$ and $\bar{\boldsymbol{\beta}}_j := \boldsymbol{\theta}^*_j - \widetilde{\boldsymbol{\alpha}}_j$.
4. $\bar{\boldsymbol{\theta}}$ is defined as the sum of $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\beta}}$.

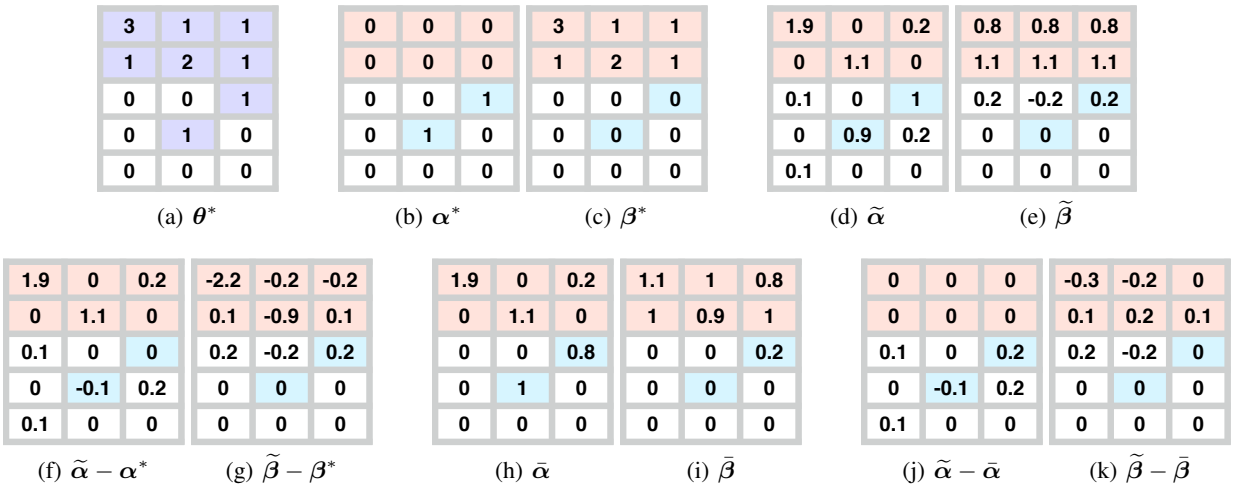

Figure 2. Example of constructing surrogate target parameters given (b) $\boldsymbol{\alpha}^*$, (c) $\boldsymbol{\beta}^*$, (d) $\widetilde{\boldsymbol{\alpha}}$ and (e) $\widetilde{\boldsymbol{\beta}}$ via transformation $\mathcal{T}(\cdot)$. Error vectors based on surrogates are decomposable (see text for details).

By construction, $\bar{\boldsymbol{\theta}}$ is same as $\boldsymbol{\theta}^*$, but $\bar{\boldsymbol{\alpha}}$ has different sparsity pattern and values from $\boldsymbol{\alpha}^*$, depending on $\widetilde{\boldsymbol{\alpha}}$. The same holds for $\bar{\boldsymbol{\beta}}$ as well. We denote the above transformation as $(\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{\beta}}) := \mathcal{T}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*; \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}})$.

It turns out that the error vectors computed based on the surrogate $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\beta}}$ are always decomposable as described in the following proposition, and the consequence of this decomposability plays a key role for showing i) ℓ_2-error bounds without incoherence condition and ii) support set recovery guarantee for non-convex $\ell_1 + \ell_{1,a}$ dirty regularizers (with faster estimation rates than for convex dirty regularizers).

Proposition 1. *Consider any local optimum $\widetilde{\boldsymbol{\theta}}$ of convex or non-convex dirty models, and corresponding $\bar{\boldsymbol{\theta}} := \mathcal{T}(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*; \widetilde{\boldsymbol{\alpha}}, \widetilde{\boldsymbol{\beta}})$. Then, the error vectors for individual components, $\boldsymbol{\Delta} := \widetilde{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}$ and $\boldsymbol{\Gamma} := \widetilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}$, are decomposable in the sense that $\big|[\boldsymbol{\Delta} + \boldsymbol{\Gamma}]_j\big| = |\boldsymbol{\Delta}_j| + |\boldsymbol{\Gamma}_j|$ for all j, and the overall ℓ_2 error $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2$ is lower bounded as follows:*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2 \geq \|\boldsymbol{\Delta}\|_2^2 + \|\boldsymbol{\Gamma}\|_2^2 \,. \tag{10}$$

Moreover, let $\boldsymbol{s}^* := supp(\boldsymbol{\alpha}^*)$ (the support set of $\boldsymbol{\alpha}^*$), $\bar{\boldsymbol{s}} := supp(\boldsymbol{\alpha}^*) \cup supp(\bar{\boldsymbol{\alpha}})$ and $U := supp(\boldsymbol{\alpha}^*) \cup supp(\boldsymbol{\beta}^*)$. Similarly, we also define $\boldsymbol{b}^* := supp(\boldsymbol{\beta}^*)$ and $\bar{\boldsymbol{b}} := supp(\bar{\boldsymbol{\beta}})$. Then, by construction of $\mathcal{T}$, $\boldsymbol{s}^* \subseteq \bar{\boldsymbol{s}} \subseteq U$ and $\boldsymbol{b}^* \subseteq \bar{\boldsymbol{b}} \subseteq U$. However, it is always guaranteed that

$$\boldsymbol{\Delta}_{\boldsymbol{s}^*} = \boldsymbol{\Delta}_{\bar{\boldsymbol{s}}} = \boldsymbol{\Delta}_U \quad \textit{and} \quad \boldsymbol{\Gamma}_{\boldsymbol{b}^*} = \boldsymbol{\Gamma}_{\bar{\boldsymbol{b}}} = \boldsymbol{\Gamma}_U \tag{11}$$

where $\boldsymbol{\Delta}_{\boldsymbol{s}^*}$ represents the projection of $\boldsymbol{\Delta}$ onto the $\boldsymbol{s}^*$-coordinate space; that is, $[\boldsymbol{\Delta}_{\boldsymbol{s}^*}]_j$ is $\boldsymbol{\Delta}_j$ if $j \in \boldsymbol{s}^*$, and 0 otherwise.

Illustrative example. Figure 2 describes an example: consider a 5×3 matrix parameter with 5 known groups in rows. Suppose (i) the target parameter is given by (a), (ii) we define (b) and (c) as the sparse and group sparse components of $\boldsymbol{\theta}^*$, and (iii) the minimizer of program (6) are computed as in (d) and (e), respectively for $\widetilde{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\beta}}$. Then, (f) and (g) show the error vectors for sparse and group-sparse components, which are not decomposable ((10) does not hold for (f) and (g)). On the other hand, for $\bar{\boldsymbol{\alpha}}$ in (h) and $\bar{\boldsymbol{\beta}}$ in (i) derived from $\mathcal{T}(\cdot)$, we can verify that surrogate error vectors (shown in (j) and (k)) are decomposable; at every position, $\widetilde{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}$ and $\widetilde{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}$ are sign consistent (or at least one of them is zero).

5. Statistical Guarantees of Models with Convex Regularizers

Throughout our analysis, we assume that the loss function $\mathcal{L}(\cdot)$ is twice differentiable and and satisfies the restricted strong convexity condition

(RSC) For any vector $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$, the loss function $\mathcal{L}(\cdot)$ satisfies

$$\big\langle \nabla\mathcal{L}(\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2) - \nabla\mathcal{L}(\boldsymbol{\theta}_1), \boldsymbol{\theta}_2 \big\rangle$$

$$\geq \begin{cases} \kappa_1 \|\boldsymbol{\theta}_2\|_2^2 - \tau_1 \|\boldsymbol{\theta}_2\|_{\boldsymbol{\eta}}^2 \,, & \text{for all } \|\boldsymbol{\theta}_2\|_2 \leq 1 \,, \quad (12) \\ \kappa_2 \|\boldsymbol{\theta}_2\|_2 - \tau_2 \|\boldsymbol{\theta}_2\|_{\boldsymbol{\eta}} \,, & \text{for all } \|\boldsymbol{\theta}_2\|_2 \geq 1 \,. \quad (13) \end{cases}$$

RSC of the loss is also used to guarantee ℓ_2-consistency (Negahban et al., 2012; Loh & Wainwright, 2015) or ℓ_∞-consistency (Loh & Wainwright, 2014) of "clean" structurally constrained problems (i.e. problems with a single

structure). Note that there are slight variations in the definition of RSC conditions in the literature. Here we adopt the form with tolerance terms in Loh & Wainwright (2014; 2015), to allow for a wide class of non quadratic and/or non-convex loss functions. We will show that RSC with tolerance in dirty norm holds with high probability under the popular setting of Gaussian ensembles, as an example.

For the analysis, we consider a slight modification of the program (6):

$$\underset{\|\boldsymbol{\theta}\|_{\boldsymbol{\eta}} \leq r}{\text{minimize}} \; \mathcal{L}(\boldsymbol{\theta}) + \|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}} \tag{14}$$

where $\mathcal{L}$ is possibly non-convex, but satisfies (RSC). The additional constraint $\|\boldsymbol{\theta}\|_{\boldsymbol{\eta}} \leq r$ also involves the dirty norm (7) but with a different parameter vector $\boldsymbol{\eta}$. This constraint is a safety radius commonly used for analyzing non-convex problems to ensure that the global minimum exists (see e.g. Loh & Wainwright (2014; 2015)). In practice, we can disregard this additional constraint.

Theorem 1. *Consider the dirty model for problem (14) where $\mathcal{L}(\cdot)$ is possibly non-convex but satisfies the restricted strong convexity (RSC). Suppose that $\boldsymbol{\theta}^*$ is feasible and the regularization parameters are set so that $\lambda_1 \geq 4\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_\infty$ and $\lambda_2 \geq 4\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|_{\infty,a^*}$. Suppose furthermore that $r \leq \min\left\{\frac{\kappa_2}{4\tau_2}, \frac{\kappa_2}{5\max\{\lambda_1/\eta_1, \lambda_2/\eta_2\}}, \frac{\lambda_1}{8\tau_1\eta_1}, \frac{\lambda_2}{8\tau_1\eta_2}\right\}$. Then, any local optimum $\widetilde{\boldsymbol{\theta}}$ of (14) is guaranteed to be ℓ_2 consistent:*

$$\left\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_2 \leq \frac{3}{\kappa_1}\max\left\{\lambda_1\sqrt{s}\,,\, \lambda_2\sqrt{s_{\mathcal{G}}}\right\} \tag{15}$$

where s is the number of nonzero elements in $\boldsymbol{\alpha}^*$ and $s_{\mathcal{G}}$ is the number of nonzero groups in $\boldsymbol{\beta}^*$.

Remarks. The error bound in (15) scales with $(n, p, s, s_{\mathcal{G}})$ at the same rate as previous analysis (Yang & Ravikumar, 2013) for the sparse plus group sparse setting, which required a much stringent incoherence condition, as we already discussed. It is also instructive to note that Theorem 1 holds for any combination of $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ such that $\boldsymbol{\alpha}^* + \boldsymbol{\beta}^* = \boldsymbol{\theta}^*$, but different views of $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ constructing $\boldsymbol{\theta}^*$ give different bounds depending on sparsity/group sparsity levels of $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ (i.e. s and $s_{\mathcal{G}}$). In this sense, Theorem 1 provides a set of ℓ_2 estimation upper bounds.

Linear model and modified graphical Lasso. In the following corollaries, we apply Theorem 1 to the linear model (3) and the modified graphical Lasso problem (5) and derive their corresponding ℓ_2 estimation bounds.

Corollary 1. *Consider the linear model (3). Assume that (i) each row X_i of the observation matrix X is independently sampled from $N(0, \Sigma)$, (ii) X is (group) column normalized by scaling as in (Negahban et al., 2012), and (iii) $\boldsymbol{w}$ is independent sub-Gaussian with parameter σ. Now suppose that in (14) we set $a := 2$ (where a is the parameter for the group norm both for $\|\boldsymbol{\theta}\|_{\boldsymbol{\eta}}$ and $\|\boldsymbol{\theta}\|_{\boldsymbol{\lambda}}$), r constant (r only depends on Σ and σ), $\lambda_1 = \eta_1 := 8\sigma\sqrt{\log p/n}$ and $\lambda_2 = \eta_2 := 8\sigma(\sqrt{m/n} + \sqrt{\log q/n})$ for q groups and maximum group size m ($\max_{g=1,\ldots,q}|G_g|$). Suppose that $\boldsymbol{\theta}^*$ is feasible to program (14) with these settings. Then with probability at least $1 - c_1\exp(-c_2 n\lambda_s^2) - c_3/q^2$, any local optimum $\widetilde{\boldsymbol{\theta}}$ satisfies*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2 \leq \frac{24\sigma}{\kappa_1}\max\left\{\sqrt{\frac{s\log p}{n}}, \sqrt{\frac{s_{\mathcal{G}} m}{n}} + \sqrt{\frac{s_{\mathcal{G}}\log q}{n}}\right\}.$$

where κ_1 is some constant depending on Σ.

Corollary 2. *Consider the modified graphical Lasso (5) to estimate inverse covariance Θ^*. Suppose we set the parameters of (14) as $\lambda_1 = \eta_1 := 4\max_{i\neq j}\left|\widehat{\Sigma}_{ij} - (\Theta^*)^{-1}_{ij}\right|$, $\lambda_2 = \eta_2 := 4\max_{g\in\mathcal{G}}\left\|\widehat{\Sigma}_g - (\Theta^*)^{-1}_g\right\|_{a^*}$ and $r \leq \frac{1}{5(|||\Theta^*|||_2+1)^2}$ where $|||\cdot|||_2$ is the spectral norm of the matrix and $a \geq 2$. In addition, assume that Θ^* is feasible for this problem. Then, any local optimum $\widetilde{\Theta}$ satisfies*

$$\|\widetilde{\Theta} - \Theta^*\|_{\mathrm{F}} \leq 3(|||\Theta^*|||_2 + 1)^2\max\left\{\lambda_1\sqrt{s}\,,\, \lambda_2\sqrt{s_{\mathcal{G}}}\right\}. \tag{16}$$

Remark. Since $\|\boldsymbol{\theta}\|_\eta$ scales with $\frac{1}{\sqrt{n}}$ for the specified values of $\boldsymbol{\eta}$, the constraint $\|\boldsymbol{\theta}\|_\eta \leq r$ gets milder as n increases. It is also important to note that this constraint is no more stringent than those of non-convex analyses with a single regularizer (Loh & Wainwright, 2015; 2014): their constraints can be written as $\eta_1\|\boldsymbol{\theta}\|_1 \leq r$ (since $\eta_1 \asymp \sqrt{\log p/n}$ for linear models for example.) in our notation, which directly implies $\|\boldsymbol{\theta}\|_\eta \leq r$ since $\|\boldsymbol{\theta}\|_{\boldsymbol{\eta}} \leq \eta_1\|\boldsymbol{\theta}\|_1$ by the definition of $\|\cdot\|_{\boldsymbol{\eta}}$.

6. Statistical Guarantees of Models with Non-convex Penalties

A natural extension of (14) is to incorporate non-convex regularizers that have some advantages such as unbiasedness. For that purpose, in this section we consider the following formulation

$$\underset{\|\boldsymbol{\theta}\|_{\boldsymbol{\eta}} \leq r}{\text{minimize}} \; \mathcal{L}(\boldsymbol{\theta}) + \mathcal{R}(\boldsymbol{\theta}; \boldsymbol{\lambda}) \tag{17}$$

where $\mathcal{R}(\boldsymbol{\theta}; \boldsymbol{\lambda}) = \inf_{\boldsymbol{\alpha},\boldsymbol{\beta}}\{\rho_{\lambda_1}(\boldsymbol{\alpha}) + \phi_{\lambda_2,a}(\boldsymbol{\beta}) \,:\, \boldsymbol{\alpha} + \boldsymbol{\beta} = \boldsymbol{\theta}\}$. While the ℓ_2 analysis in Theorem 1 can be extended to non-convex regularizers following proof techniques recently developed in Loh & Wainwright (2015), using non-convex unbiased regularizers has no benefit in terms of asymptotic convergence rates of ℓ_2 estimation errors. Instead, we here investigate the ℓ_∞-norm bound and related support set recovery guarantees where non-convex unbiased regularizers help. To derive ℓ_∞ bounds, we use the

primal-dual witness method described in the supplementary materials.

Theorem 2. *Consider the dirty program with non-convex penalties in (17), under (RSC). Suppose* $2r(\tau_2 + 2\max\{\frac{\lambda_1}{\eta_1}, \frac{\lambda_2}{\eta_2}\}) \leq 1$. *Also suppose that for some* $\delta \in \big[\max\{\frac{4r\tau_1\eta_1}{\lambda_1}, \frac{4r\tau_1\eta_2}{\lambda_2}\}, 1\big]$, *the strict dual feasibility of primal-dual witness holds. Then, any stationary point* $\widehat{\boldsymbol{\theta}}$ *of (17) is supported by* U *(recall* $U := supp(\boldsymbol{\alpha}^*) \cup supp(\boldsymbol{\beta}^*)$*) if the number of samples is large enough to satisfy* $\max\{\lambda_1\sqrt{s}, \lambda_2\sqrt{s_{\mathcal{G}}}\}^2 < c$ *for some constant* c *depending only on* κ_1, τ_1 *and* δ.

As in Theorem 1, the decomposability in (10) and (11) with respect to the surrogates $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\beta}}$, plays a crucial role in establishing the support set recovery guarantee of any local optimum in Theorem 2.

Based on Theorem 2, we can derive the ℓ_∞ bounds following the standard steps in (Loh & Wainwright, 2014):

Corollary 3. *Suppose the assumptions in Theorem 2 hold. Then,*

1. *If* $\frac{\kappa_1 - \mu}{2} \geq \tau_1\big(\max\{\eta_1\sqrt{s}, \eta_2\sqrt{s_{\mathcal{G}}}\}\big)^2$ *holds for large enough sample size* n, *the program (17) has a unique stationary point* $\widehat{\boldsymbol{\theta}}$, *specified by the construction of the primal dual witness.*
2. *Letting* $\widehat{Q} := \int_0^1 \nabla^2\mathcal{L}\big(\boldsymbol{\theta}^* + t(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\big)dt$, *it holds that* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \big\|\big(\widehat{Q}_{UU}\big)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\big\|_\infty + \min\{\lambda_1, \lambda_2\}\big\|\big(\widehat{Q}_{UU}\big)^{-1}\big\|_\infty$ *where* $\|\cdot\|_\infty$ *denotes a matrix induced norm (maximum absolute row sum).*
3. *Moreover, if* ρ *is* (μ, γ)*-amenable, and the minimum absolute value* $\boldsymbol{\theta}^*_{\min} := \min_j |\boldsymbol{\theta}^*_j|$ *is lower-bounded by* $\boldsymbol{\theta}^*_{\min} \geq \big\|\big(\widehat{Q}_{UU}\big)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\big\|_\infty + \min\{\lambda_1, \lambda_2\}\big\|\big(\widehat{Q}_{UU}\big)^{-1}\big\|_\infty + 2\max\{\lambda_1, \lambda_2\}\gamma$. *Then, the error bound in the statement 2 is reduced to tighter bound as* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_\infty \leq \big\|\big(\widehat{Q}_{UU}\big)^{-1}\nabla\mathcal{L}(\boldsymbol{\theta}^*)_U\big\|_\infty$.

Multi-task high-dimensional linear regression. We consider the multi-task high-dimensional linear regression, as a concrete example of using non-convex dirty regularizers. This is the counterpart of model (4) which uses convex dirty regularizer. In the following corollary, we analyze the sparsistency of dirty multi-task linear regression with non-convex regularizers:

$$\underset{\Theta\in\mathbb{R}^{p\times m}\text{s.t.}\|\Theta\|_{\boldsymbol{\eta}}\leq r}{\text{minimize}} \sum_{k=1}^{m}\frac{1}{2n}\big\|\boldsymbol{y}^{(k)} - X^{(k)}\Theta_{(\bullet,k)}\big\|_2^2 + \mathcal{R}(\Theta;\boldsymbol{\lambda}) \tag{18}$$

where $\mathcal{R}(\Theta;\boldsymbol{\lambda}) = \inf_{\boldsymbol{\alpha},\boldsymbol{\beta}}\{\rho_{\lambda_1}(\boldsymbol{\alpha}) + \phi_{\lambda_2,a}(\boldsymbol{\beta}) : \boldsymbol{\alpha} + \boldsymbol{\beta} = \Theta\}$. Now, we derive a corollary for this particular non-convex dirty model.

Corollary 4. *Consider the multitask regression model. Suppose that for each task, design matrix* $X^{(k)}$ *is a zero-mean Gaussian ensemble and is column normalized,* $\boldsymbol{w}^{(k)}$ *is independent sub-Gaussian with parameter* σ. *Now suppose we set parameters of (18) as* $a := \infty$, r *constant (only depends on* Σ *and* σ.*),* $\lambda_1 := c_1\sigma\sqrt{\log(pm)/n}$, $\lambda_2 := c_2\sigma\sqrt{(\log p + m\log 2)/n}$, *and* Θ^* *is feasible to program (14) with these settings. Then, for any local optimum* $\widetilde{\Theta}$, *with probability at least* $1 - c_1\exp(-c_2\log(pm)) - c_3\exp\big(-c_4(\log p + m\log 2)\big)$ *(which is approaching to 1) for some positive constants* $c_1 - c_4$,

1. $supp(\widetilde{\Theta}) \subseteq supp(\Theta^*)$,
2. *if additionally the regularizer* ρ_λ *is* (μ, γ)*-amenable with* $\mu < \lambda_{\min}(\Sigma)$ *(the minimum eigenvalue of* Σ*) and* $C_{\min} := \min_{k=1,\ldots,m}\lambda_{\min}\big(\Sigma^{(k)}_{U_kU_k}\big) > 0$, *then* $supp(\widetilde{\Theta}) = supp(\Theta^*)$ *and the element-wise difference is bounded as follows:*

$$\big\|\widetilde{\Theta} - \Theta^*\big\|_{\max} := \max_{i,j}|[\widetilde{\Theta} - \Theta^*]_{i,j}| \leq \sigma\sqrt{\frac{100\log(pm)}{nC_{\min}}}$$

provided that $\boldsymbol{\theta}^*_{\min} \geq \sigma\sqrt{\frac{100\log(pm)}{nC_{\min}}} + \min\{\lambda_1, \lambda_2\}\max_{k=1,\ldots,m}\big\|(\Sigma^{(k)}_{U_kU_k})^{-1}\big\|_\infty + 2\max\{\lambda_1, \lambda_2\}\gamma$.

Remark. In order to highlight the benefit of using (μ, γ)-amenable regularizers, we briefly compare the result of Corollary 4 with that of $\ell_1 + \ell_{1,a}$ case in (Jalali et al., 2010). Not only the result in (Jalali et al., 2010) requires the incoherence on X (specifically, $\max_{j\in U^c}\sum_{k=1}^m\big\|\Sigma^{(k)}_{j,U_k}\big(\Sigma^{(k)}_{j,U_k}\big)^{-1}\big\|_1 < 1$), but it also has an additional $\frac{s\lambda_1}{C_{\min}\sqrt{n}}$ term in $\big\|\widetilde{\Theta} - \Theta^*\big\|_{\max}$ bound. Moreover, the required λ_1 and λ_2 there can converge to zero more slowly: $\lambda_1 \asymp \frac{\sqrt{\log(pm)}}{\sqrt{n} - \sqrt{s\log(pm)}}$ and $\lambda_2 \asymp \frac{\sqrt{m(m+\log p)}}{\sqrt{n} - \sqrt{sm(m+\log p)}}$.

7. Experiments

To illustrate the practical consequences of the superior statistical guarantees of models with non-convex penalties, we perform experiments on both simulated and real-world data and compare convex and non-convex dirty models for sparse + group-sparse structures.

Simulated data. We consider multitask regression problems with $m = 10$ tasks and $p = 260$ variables for settings of parameters $(s, s_{\mathcal{G}}) \in \{(2p/10, p/20), (p/10, p/10)\}$ with respectively less / more support overlap across tasks (recall s and $s_{\mathcal{G}}$ are the number of nonzero elements in $\boldsymbol{\alpha}^*$ and the number of nonzero groups in $\boldsymbol{\beta}^*$, respectively). The

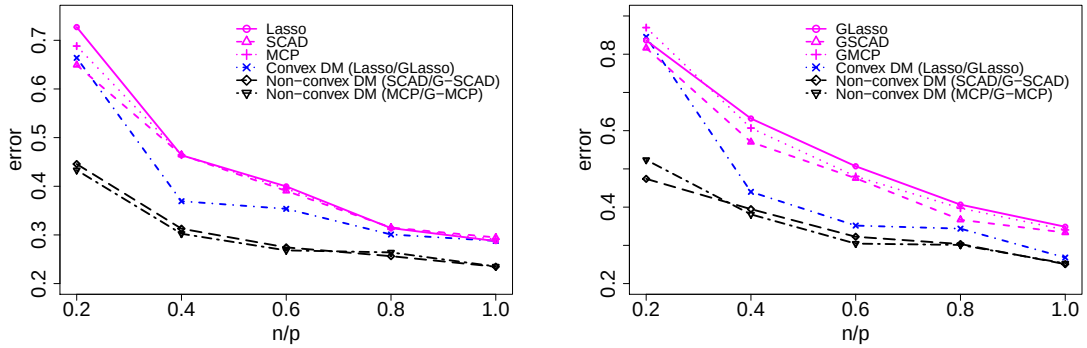

Figure 3. ℓ_∞-error for comparison methods for varying sample size n. Left: less sharing across tasks. Right: more sharing across tasks

rows of the design matrices X are sampled i.i.d. from a zero-mean Gaussian distribution with correlation of 0.2 between feature pairs. For each set of parameters $(s, s_\mathcal{G})$, we generate 100 instances of the problem where for each instance the non-zero entries of the true model parameter matrix are i.i.d. zero-mean Gaussian to agree with s and $s_\mathcal{G}$. Gaussian error with standard deviation of 4 is added to each observation. For varying sample size n we measure the ℓ_∞ error of parameters estimated by (i) convex dirty model (Jalali et al., 2010), (ii) non-convex dirty model with SCAD + Group-SCAD penalty, and (iii) non-convex dirty model with MCP + Group-MCP penalty. We also evaluate the following baselines: Lasso, MCP, SCAD, Group-Lasso, Group-MCP and Group-SCAD [2]. The regularization parameters of each method are tuned via 5-folds cross-validation. The results are presented in Figure 3 (To avoid cluttering the graphs, we do not display standard errors as these are much lower than the gaps between the pertinent groups of methods, and we only display the best group of baselines for each setting). As can be seen from the figure, dirty models with non-convex penalties enjoy superior performance over their counterparts with convex penalties as a function of the sample size. In terms of computational cost, (Group) coordinate descent steps for (group) lasso, (group) MCP and (group) SCAD all have simple closed-form expressions (Huang et al., 2012), similarly for proximal-based approaches. We noticed that for a wide range of (λ_1, λ_2), non-convex procedures took less time and converged faster (See supplements). As future work it would be interesting to study their theoretical numerical convergence rates.

Real data analysis. We consider the problem of predicting biological activities of molecules given features extracted from their chemical structures. We analyze three biological activity datasets from the "molecular activity challenge" (`http://www.kaggle.com/c/MerckActivity`). Specifically we consider multitask regression with three tasks corresponding to predicting the raw value $(-\log(IC50))$ of three different types of biological activities : 'binding to cannabinoid receptor 1', 'inhibition of dipeptidyl peptidase 4' and 'time dependent 3A4 inhibitions'. For each task we used $n = 200$ observations with $p = 3000$ molecular features. We consider 20 random data splits into training and validation sets, using 2/3 of the data for tranining and 1/3 for validation, and report the average R^2 over these random splits. As shown in table1, dirty models outperformed "clean" models suggesting the importance to strike a balanc e between task specificity and sharing for this data. Non-convex dirty models achieved the best R^2, which illustrate their capability as a valuable tool for high-dimensional data analysis.

Table 1. Average R^2 for comparison methods on molecular activity data

Method	R^2
Lasso	0.36 ± 0.03
SCAD	0.37 ± 0.03
MCP	0.36 ± 0.04
GLasso	0.35 ± 0.03
GSCAD	0.37 ± 0.03
GMCP	0.38 ± 0.03
Convex DM (Lasso/GLasso)	0.43 ± 0.04
Nonconvex DM (SCAD/GSCAD)	$\mathbf{0.49 \pm 0.04}$
Nonconvex DM (MCP/GMCP)	$\mathbf{0.49 \pm 0.03}$

8. Concluding Remarks

This paper finally resolved the outstanding case of sparse + group-sparse dirty models with convex penalties: we provided the first satisfactory consistency results that do not require implausible assumptions, thereby fully justifying their practical success. In addition we proposed and studied dirty models with non-convex penalties and showed that they enjoy superior theoretical guarantees that translate into significant practical impact. An interesting direction for future work is to investigate whether our proof technique might be applicable to other dirty models and beyond.

[2] Our theorem on ℓ_∞ consistency requires the sample size to be larger than the maximum of *two* terms, which precludes from presenting graphs with curve alignment across p (by rescaling the x-axis with a control parameter as in Jalali et al. (2010)).

Acknowledgments

E.Y. acknowledges the support of MSIP/NRF (National Research Foundation of Korea) via NRF-2016R1A5A1012966 and MSIP/IITP (Institute for Information & Communications Technology Promotion of Korea) via ICT R&D program 2016-0-00563, 2017-0-00537.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40 (2):1171–1197, 2012.

Baraniuk, R. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

Candès, E. J. and Tao, T. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM*, 58(3), May 2011.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

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.

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

Gong, P., Ye, J., and Zhang, C. Multi-stage multi-task feature learning. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 25*, pp. 1988–1996. 2012.

Hara, S. and Washio, T. Learning a common substructure of multiple graphical gaussian models. *Neural Networks*, 38:23–38, 2013.

Hsu, D., Kakade, S. M., and Zhang, T. Robust matrix decomposition with sparse corruptions. *Information Theory, IEEE Transactions on*, 57(11):7221–7234, 2011.

Huang, J., Breheny, P., and Ma, S. A selective review of group selection in high-dimensional models. *Statist. Sci.*, 27(4):481–499, 11 2012. doi: 10.1214/12-STS392. URL http://dx.doi.org/10.1214/12-STS392.

Jalali, A., Ravikumar, P., Sanghavi, S., and Ruan, C. A dirty model for multi-task learning. In *Neur. Info. Proc. Sys. (NIPS)*, 23, 2010.

Loh, P. and Wainwright, M. J. Support recovery without incoherence: A case for nonconvex regularization. *Arxiv preprint arXiv:1412.5632*, 2014.

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.

Negahban, S. and Wainwright, M. J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. In *Inter. Conf. on Machine learning (ICML)*, 2010.

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.

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. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle. In *Neur. Info. Proc. Sys. (NIPS)*, 21, 2008.

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) (JRSSB)*, (5):1009–1030, 2009.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *Allerton Conference 07, Allerton House, Illinois*, 2007.

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

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 2009.

Yang, E. and Ravikumar, P. Dirty statistical models. In *Neur. Info. Proc. Sys. (NIPS)*, 26, 2013.

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.

Yuan, M. and Lin, Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):49, 2006.

Zhang, C. H. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(8):894–942, 2010.