A Unified Dynamic Approach to Sparse Model Selection

Chendi Huang
School of Mathematical Sciences,
Peking University,
Beijing, China
cdhuang@pku.edu.cn

Yuan Yao
Department of Mathematics,
Hong Kong University of Science and Technology,
HKSAR, China
yuany@ust.hk

Abstract

Sparse model selection is ubiquitous from linear regression to graphical models where regularization paths, as a family of estimators upon the regularization parameter varying, are computed when the regularization parameter is unknown or decided data-adaptively. Traditional computational methods rely on solving a set of optimization problems where the regularization parameters are fixed on a grid that might be inefficient. In this paper, we introduce a simple iterative regularization path, which follows the dynamics of a sparse Mirror Descent algorithm or a generalization of Linearized Bregman Iterations with nonlinear loss. Its performance is competitive to `glmnet` with a further bias reduction. A path consistency theory is presented that under the Restricted Strong Convexity (RSC) and the Irrepresentable Condition (IRR), the path will first evolve in a subspace with no false positives and reach an estimator that is sign-consistent or of minimax optimal ℓ_2 error rate. Early stopping regularization is required to prevent overfitting. Application examples are given in sparse logistic regression and Ising models for NIPS coauthorship.

1 Introduction

In high dimensional statistics and machine learning, the data $\mathbf{z}$ is often assumed to be generated from a statistical model $\mathcal{P}(\alpha^\star, \beta^\star)$ with a sparse parameter $\beta^\star$, and the purpose is to estimate $\beta^\star$ typically via the following optimization approach,

$$\min_{\alpha,\beta} \left(\ell\left(\alpha,\beta;\mathbf{z}\right) + \lambda P\left(\beta\right)\right), \qquad (1.1)$$

where $\ell(\alpha,\beta;\mathbf{z})$ is a loss function depending on data $\mathbf{z}$ and parameter (α,β), usually based on likelihood, and $P(\beta)$ is a penalty function. For simplicity, we shall omit the dependence on $\mathbf{z}$ for the loss when it is clear from the context.

Example 1 (Sparse linear regression model)**.** Let $X = (x^{(1)},\ldots,x^{(n)})^T \in \mathbb{R}^{n\times p}$ be a fixed design matrix, and $y = (y^{(1)},\ldots,y^{(n)})^T \in \mathbb{R}^n$,

$$y^{(i)} = \alpha^\star + \beta^{\star T} x^{(i)} + \epsilon^{(i)} \ (1 \le i \le n),$$

with $\epsilon^{(i)}$'s i.i.d. drawn from $N(0,\sigma^2)$, and $\beta^\star$ sparse. Let

$$\ell(\alpha,\beta;\mathbf{z}) = \|y - \alpha - X\beta\|_2^2/(2n)$$

be the loss function for data $\mathbf{z} = (x^{(i)}, y^{(i)})_{i=1}^n$ and parameter (α,β) (intercept α and linear parameter β), as well as the Lasso penalty $P(\beta) = \|\beta\|_1$. For model selection consistency, Zhao and Yu (2006) and Wainwright (2009) showed it under Restricted Strong Convexity (RSC) and Irrepresentable Condition (IRR); under a weaker restricted eigenvalue condition, Bickel et al. (2009) established the ℓ_2-error at minimax optimal rates.

Example 2 (Sparse logistic regression model)**.** Let $x^{(i)} \in \mathbb{R}^p$ $(1 \le i \le n)$, and $y^{(i)} \in \{1,-1\}$,

$$\mathbb{P}\left(y^{(i)} = 1|x^{(i)}\right) = 1/\left(1 + \exp\left(-\left(\alpha^\star + \beta^{\star T} x^{(i)}\right)\right)\right),$$

with $\beta^\star$ sparse. Ravikumar et al. (2010) considered (1.1) with the loss function for data $\mathbf{z} = (y^{(i)}, x^{(i)})_{i=1}^n$ and parameter (α,β)

$$\begin{aligned}\ell\left(\alpha,\beta;\mathbf{z}\right) &= -\frac{1}{n}\sum_{i=1}^n \log \mathbb{P}_{\alpha,\beta}\left(y = y^{(i)}|x^{(i)}\right)\\ &= \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp\left(-\left(\alpha + \beta^T x^{(i)}\right) y^{(i)}\right)\right),\end{aligned} \qquad (1.2)$$

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. Copyright 2018 by the author(s).

as well as $P(\beta) = \|\beta\|_1$. They also showed its selection/estimation consistency.

Example 3 (Sparse Ising model)**.** $x^{(i)}$ $(1 \leq i \leq n)$ are drawn from $x \in \{1, -1\}^p$ whose population satisfies

$$\begin{aligned}\mathbb{P}&\left(x = (x_1, \ldots, x_p)^T\right)\\ &\propto \exp\left(\frac{1}{2}\sum_{j=1}^{p}\alpha_j^\star x_j + \frac{1}{2}\sum_{j<j'}\beta_{j,j'}^\star x_j x_{j'}\right), \quad (1.3)\end{aligned}$$

where $\alpha^\star \in \mathbb{R}^p$, $\beta^\star \in \mathbb{R}^{p\times p}$,[1] and $\beta^\star$ is sparse. Ravikumar et al. (2010) studied sparse Ising model (1.3) by the so-called *neighborhood-based logistic regression*, based on the discussion on sparse logistic models in their paper. Specifically, despite the difficulty to deal with the whole $(\alpha^\star, \beta^\star)$ by using likelihood-based loss functions of Ising model, they noticed that

$$\mathbb{P}\left(x_j | x_{-j}\right) = 1/\left(1 + \exp\left(-\left(\alpha_j^\star + \beta_{-j,j}^{\star T} x_{-j}\right) x_j\right)\right).$$

Thus each j corresponds to a sparse logistic regression problem, i.e. Example 2, with y, x, α, β replaced by $x_j, x_{-j}, \alpha_j, \beta_{-j,j}$. Thus they learned $(\alpha_j^\star, \beta_{-j,j}^\star)$ (by ℓ_1 regularized logistic regression) for each j, instead of dealing with $(\alpha^\star, \beta^\star)$ directly. Xue et al. (2012) considered (1.1) with the loss $\ell(\alpha, \beta)$ being the *negative composite conditional log-likelihood*

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\log\left(1 + \exp\left(-\left(\alpha_j + \beta_{-j,j}^T x_{-j}^{(i)}\right) x_j^{(i)}\right)\right). \quad (1.4)$$

$P(\cdot)$ can be ℓ_1 penalty, SCAD penalty or other positive penalty function defined on $[0, +\infty)$. Alternatively Sohl-Dickstein et al. (2011) proposed an approach of Minimum Probability Flow (MPF) which in the case of Ising model uses the following loss

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\exp\left(-\frac{1}{2}\left(\alpha_j + \beta_{-j,j}^T x_{-j}^{(i)}\right) x_j^{(i)}\right). \quad (1.5)$$

The minimizer of this function is a reasonable estimator of $(\alpha^\star, \beta^\star)$. However their work did not treat sparse models in high-dimensional setting. When facing sparse Ising model, one may consider (1.1) with the loss $\ell(\alpha, \beta)$ being the expression in (1.5) and $P(\beta) = \|\beta\|_1$, which is not seen in literature to the best of our knowledge.

Example 4 (Sparse Gaussian graphical model)**.** $x^{(i)}$ $(1 \leq i \leq n)$ are drawn from a multivariate Gaussian distribution with covariance $\Sigma^\star \in \mathbb{R}^{p\times p}$, and the *precision matrix* $\Omega^\star = \Sigma^{\star -1}$ is assumed to be sparse. Yuan and Lin (2007) and Ravikumar et al. (2008) studied (1.1), with the loss function being the *negative scaled log-likelihood*, and the penalty being the sum of the absolute values of the off-diagonal entries of the precision matrix.

In general, Negahban et al. (2009) provided a unified framework for analyzing the statistical consistency of the estimators derived by solving (1.1) with a proper choice of λ. However in practice, since λ is unknown, one typically needs to compute the regularization path β_λ as regularization parameter λ varies on a grid, e.g. the `lars` (Efron et al., 2004) or the coordinate descent in `glmnet`. Such regularization path algorithms can be inefficient in solving many optimization problems.

In this paper, we look at the following three-line iterative algorithm which, despite its simplicity, leads to a novel unified scheme of regularization paths for all cases above,

$$\alpha_{k+1} = \alpha_k - \kappa\delta_k \nabla_\alpha \ell\left(\alpha_k, \beta_k\right), \quad (1.6a)$$
$$z_{k+1} = z_k - \delta_k \nabla_\beta \ell\left(\alpha_k, \beta_k\right), \quad (1.6b)$$
$$\beta_{k+1} = \kappa \mathcal{S}\left(z_{k+1}, 1\right), \quad (1.6c)$$

where $z_0 = \beta_0 = 0$, α_0 can be arbitrary and is naturally set $\arg\min_\alpha \ell\left(\alpha, \beta_0\right)$, step size $\delta_k = \delta$ and κ are parameters whose selection to be discussed later, and the shrinkage operator $\mathcal{S}(\cdot, 1)$ is defined element-wise as $\mathcal{S}\left(z, 1\right) = \mathrm{sign}(z) \cdot \max\left(|z| - 1, 0\right)$. Such an algorithm is easy for parallel implementation, with linear speed-ups demonstrated in experiment Section 3 below.

To see the regularization paths returned by the iteration, Figure 1 compared it against the `glmnet`. Such simple iterative regularization paths exhibit competitive or even better performance than the Lasso regularization paths by `glmnet` in reducing the bias and improving the accuracy (Section 3.2 for more details).

How does this simple iteration algorithm work?

There are two equivalent views on algorithm (1.6). First of all, it can be regarded as a mirror descent algorithm (MDA) (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003; Nemirovski, 2012)

$$\begin{aligned}&(\alpha_{k+1}, \beta_{k+1})\\ &= \arg\min_z \left\{\langle z, \delta\nabla_{\alpha,\beta}\ell(\alpha_k, \beta_k)\rangle + B_\Phi(z, (\alpha_k, \beta_k))\right\}\\ &:= \mathrm{prox}_\Phi(\delta\nabla_{\alpha,\beta}\ell(\alpha_k, \beta_k))\end{aligned}$$

where B_Φ is the bregman divergence associated with Φ, i.e. defined by

$$B_\Phi(u, v) = \Phi(u) - \Phi(v) - \langle\partial\Phi(v), u - v\rangle. \quad (1.7)$$

Now set $\Phi(\alpha, \beta) = \|\alpha\|_2^2/(2\kappa) + \|\beta\|_1 + \|\beta\|_2^2/(2\kappa)$ involving a Ridge (ℓ_2) penalty on α and an elastic net type (ℓ_1 and ℓ_2) penalty on β. Hence $\partial_\alpha\Phi(\alpha, \beta) = \alpha/\kappa$

[1] We assume $\mathrm{diag}(\beta^\star) = 0$ and $\beta^\star$ is symmetric.

and $\partial_\beta \Phi(\alpha, \beta) = \rho + \beta/\kappa$ where $\rho \in \partial\|\beta\|_1$. With this, the optimization in MDA leads to (1.6a) and

$$\rho_{k+1} + \frac{1}{\kappa}\beta_{k+1} = \rho_k + \frac{1}{\kappa}\beta_k - \delta \nabla_\beta \ell(\alpha_k, \beta_k), \ \rho_k \in \partial\|\beta_k\|_1, \tag{1.8}$$

which is equivalent to (1.6b). There has been extensive studies on the convergence $\ell(\alpha_k, \beta_k) - \min_{\alpha,\beta} \ell(\alpha, \beta) \leq O(k^{-r})$ $(r > 0)$, which are however not suitable for statistical estimate above as such convergent solutions lead to overfitting estimators.

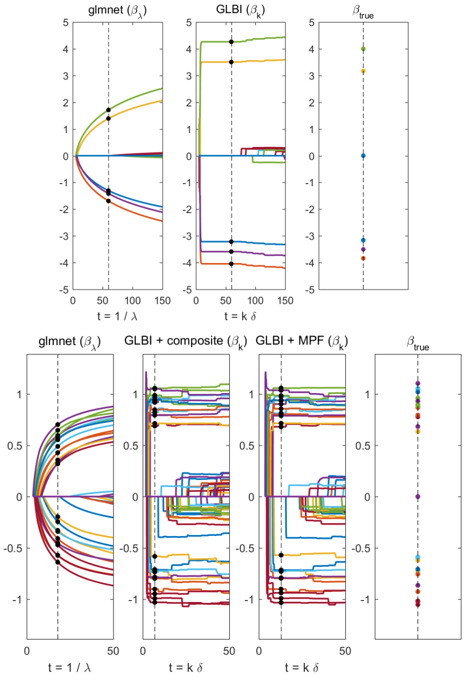

Figure 1: Top: path comparison between $\{\beta_\lambda\}$ $(t = 1/\lambda)$ by `glmnet` (left) and $\{\beta_k\}$ $(t = k\delta)$ by GLBI (middle), for logistic models with true parameters (right). Bottom: path comparison of $\{\beta_\lambda\}$ $(t = 1/\lambda)$ by `glmnet` (left), $\{\beta_k\}$ $(t = k\delta)$ by GLBI + composite loss (middle left), and $\{\beta_k\}$ $(t = k\delta)$ by GLBI + MPF loss (middle right), for Ising models with true parameters (right). In both cases, `glmnet` selects biased estimates while GLBI finds more accurate ones.

An alternative dynamic view may lead to a deeper understanding of the regularization path. In fact for Example 1, (1.6) reduces to the *Linearized Bregman Iteration (LBI)* proposed by Yin et al. (2008) and analyzed by Osher et al. (2016) via its limit differential inclusions. It shows that equipped with the standard conditions as Lasso, an early stopping rule can find a point on the regularization path of (1.6) with the same sign pattern as true parameter (sign-consistency) and gives the unbiased oracle estimate, hence better than Lasso or any convex regularized estimates which are always biased. This can be generalized to our setting where (1.8) is a discretization of the following dynamics

$$\dot{\alpha}(t)/\kappa = -\nabla_\alpha \ell\left(\alpha(t), \beta(t)\right), \tag{1.9a}$$

$$\dot{\rho}(t) + \dot{\beta}(t)/\kappa = -\nabla_\beta \ell\left(\alpha(t), \beta(t)\right), \tag{1.9b}$$

$$\rho(t) \in \partial\|\beta(t)\|_1. \tag{1.9c}$$

It is a restricted gradient flow (differential inclusion) where $\beta(t)$ has its sparse support controlled by $\rho(t)$. As $\kappa \to \infty$, it gives a sequence of estimates by minimizing ℓ with the sign pattern of $\beta(t)$ restricted on $\rho(t)$. Thus if an estimator $\beta(t)$ has the same sign pattern as $\beta^\star$, it must returns the *unbiased oracle estimator* which is optimal. So it is natural to ask if there is a point on the path $\beta(t)$ (or β_k) which meets the sparsity pattern of true parameter $\beta^\star$. This is the **path consistency** problem to be addressed in this paper. In Section 2, we shall present a theoretical framework as an answer, and Section 3 gives more applications, including Ising model learning for NIPS coauthorship.

Note that for Example 2, (1.6) reduces to the linearized Bregman iterations for logistic regression proposed by Shi et al. (2013) without a study of statistical consistency. A variable splitting scheme in comparison to generalized Lasso is studied in Huang et al. (2016) which shows improved model selection consistency in some scenarios. Hence in this paper, we shall call the general form (1.6) as *Generalized Linear Bregman Iterations (GLBI)*, in addition to (sparse) Mirror Descent flows.

2 Path Consistency of GLBI

Let $\theta^\star = (\alpha^\star, \beta^{\star T})^T$ denotes the true parameter, with sparse $\beta^\star$. Define $S := \mathrm{supp}(\beta^\star)$ $(s := |S| \ll p)$ as the index set corresponding to nonzero entries of β, and S^c be its complement. Let $S_\alpha = (\alpha, S)$, and $S_\alpha = S$ when α drops. Let the *oracle estimator* be

$$\theta^o = (\alpha^o, \beta^{oT})^T \in \arg \min_{\substack{\alpha,\beta \\ \beta_{S^c}=0}} \ell\left(\alpha, \beta\right), \tag{2.1}$$

which is an optimal estimate of $\theta^\star$. GLBI starts within the oracle subspace $(\{\theta = (\alpha, \beta^T)^T : \ \beta_{S^c} = 0\})$, and we are going to prove that under an Irrepresentable Condition (IRR) the dynamics will evolve in the oracle subspace with high probability before the stopping time $\bar{k}$, approaching the oracle estimator exponentially fast due to the Restricted Strong Convexity (RSC). Thus if all the true parameters are large enough, then we can identify their sign pattern correctly; otherwise, such a stopping time still finds an estimator (possibly with false positives) at minimax optimal ℓ_2 error rate. Furthermore, if the algorithm continues beyond the stopping time, it might escape the oracle subspace and eventually reach overfitted estimates. Such a picture is illustrated in Figure 2.

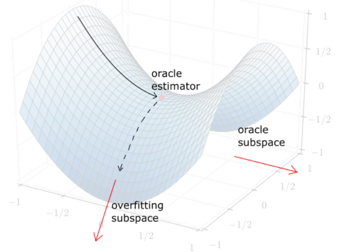

Figure 2: An illustration of global dynamics of the algorithm in this paper.

Hence, it is helpful to define the following ***oracle dynamics***:

$$\alpha'_{k+1} = \alpha'_k - \kappa\delta\nabla_\alpha \ell\left(\alpha'_k, \beta'_k\right), \tag{2.2a}$$

$$z'_{k+1,S} = z'_{k,S} - \delta\nabla_S \ell\left(\alpha'_k, \beta'_k\right), \tag{2.2b}$$

$$\beta'_{k+1,S} = \kappa\mathcal{S}\left(z'_{k+1,S}, 1\right), \tag{2.2c}$$

with $z'_{k,S^c} = \beta'_{k,S^c} \equiv 0_{p-s}$. Let $\theta'_k := (\alpha'_k, \beta'^T_k)^T$.

2.1 Basic Assumptions

Now we are ready to state the general assumptions that can be reduced to existing ones. We write

$$\ell(\theta) := \ell\left(\alpha, \beta\right),$$

$$\bar{H}(\theta) := \bar{H}(\alpha, \beta) := \int_0^1 \nabla^2 \ell\left(\theta^\star + \mu\left(\theta - \theta^\star\right)\right) \mathrm{d}\mu,$$

$$\bar{H}^o(\theta) := \bar{H}^o(\alpha, \beta) := \int_0^1 \nabla^2 \ell\left(\theta^o + \mu\left(\theta - \theta^o\right)\right) \mathrm{d}\mu.$$

Assumption 1 (Restricted Strong Convexity (RSC))**.** There exist $\lambda, \Lambda > 0$, such that for any $k \geq 0$, and for any θ on the line segment between θ'_k and θ^o, or on the line segment between $\theta^\star$ and θ^o,

$$\lambda I \preceq \nabla^2_{S_\alpha, S_\alpha} \ell(\theta) \preceq \Lambda I,$$

Assumption 2 (Irrepresentable Condition (IRR))**.** There exist $\eta \in (0, 1]$ and $C > 0$ such that

$$\sup_{K\geq 1} \left\| \sum_{k=0}^{K-1} \overline{\mathrm{irr}}_k \left(\begin{pmatrix} \alpha'_{k+1}/\kappa \\ z'_{k+1,S} \end{pmatrix} - \begin{pmatrix} \alpha'_k/\kappa \\ z'_{k,S} \end{pmatrix} \right) \right\|_\infty < 1 - \frac{\eta}{2},$$

$$\sup_{k\geq 0} \left\| \overline{\mathrm{irr}}_k \right\|_\infty \leq C,$$

where

$$\overline{\mathrm{irr}}_k := \bar{H}_{S^c, S_\alpha}\left(\theta'_k\right) \cdot \bar{H}_{S_\alpha, S_\alpha}\left(\theta'_k\right)^{-1}.$$

Remark 1. For sparse linear regression problem (Example 1) with no intercept (α drops), Assumption 1 reduces to $\lambda I \preceq X^*_S X_S \preceq \Lambda I$. The lower bound is exactly the RSC proposed in linear problems. Although the upper bound is not needed in linear problems, it arises in the analysis for logistic problem by Ravikumar et al. (2010) (see (A1) in Section 3.1 in their paper). Besides, $\overline{\mathrm{irr}}_k$ is constant and Assumption 2 reduces to

$$\sup_{K\geq 1} \left\| X^*_{S^c} X_S \left(X^*_S X_S\right)^{-1} z'_{K,S} \right\|_\infty < 1 - \frac{\eta}{2},$$

$$\left\| X^*_{S^c} X_S \left(X^*_S X_S\right)^{-1} \right\|_\infty \leq C,$$

which is true with high probability, as long as the classical Irrepresentable Condition (Zhao and Yu, 2006) $\|X^*_{S^c} X_S (X^*_S X_S)^{-1}\|_\infty \leq 1 - \eta$ holds along with $C \geq 1$ and κ is large, since by (C.6),

$$\begin{aligned}
\left\|z'_{K,S}\right\|_\infty &\leq \left\|z'_{K,S} - \mathcal{S}\left(z'_{K,S}, 1\right)\right\|_\infty + \left\|\mathcal{S}\left(z'_{K,S}, 1\right)\right\|_\infty \\
&\leq 1 + \left\|\beta'_{K,S}\right\|_\infty / \kappa \\
&\leq 1 + \left(\left\|\beta'_{K,S} - \beta^o_S\right\|_2 + \|\beta^o_S\|_2\right) / \kappa \\
&\leq 1 + \left(\sqrt{\Lambda/\lambda} + 1\right) \|\beta^o_S\|_2 / \kappa \\
&< (1 - \eta/2)/(1 - \eta).
\end{aligned}$$

Remark 2. For sparse logistic regression problem (Example 2), we have the following proposition stating that Assumption 1 and 2 hold with high probability under some natural setting, along with condition (2.3). See its proof in Appendix D. A slightly weaker condition compared to (2.3b), and a same version of (2.3c), can be found in Ravikumar et al. (2010), where $x^{(i)}$ are discrete.

Proposition 1. *In Example 2, we suppose $x^{(i)}$'s are i.i.d. drawn from some $X \sim N(0, \Sigma)$, where $\Sigma_{j,j} \leq 1$ $(1 \leq j \leq p)$. Then there exist constants $C_0, C_1, C_2 > 0$, such that Assumption 1 and 2 hold with probability not less than $1 - C_0/p$, as long as κ is sufficiently large and*

$$\left\|\Sigma_{S^c,S}\Sigma^{-1}_{S,S}\right\|_\infty \leq 1 - \eta, \tag{2.3a}$$

$$n/(\log n)^2 \geq C_1 s^4 \log p, \tag{2.3b}$$

$$\beta^\star_{\min} := \min_{j\in S} \left|\beta^\star_j\right| \geq C_2\sqrt{(s \log p)/n}. \tag{2.3c}$$

2.2 Path Consistency Theorem

Theorem 1 (Consistency of GLBI)**.** *Under Assumption 1 and 2, suppose $\kappa \geq 2\|\theta^o_{S_\alpha}\|_2$, and $\bar{k} \in \mathbb{N}$ such that*

$$\left(\bar{k} - 1\right)\delta < \frac{\eta}{2(C+1)} \cdot \frac{1}{\|\nabla\ell\left(\alpha^\star, \beta^\star\right)\|_\infty} \leq \bar{k}\delta. \tag{2.4}$$

Define λ' as in (C.3). We have the following properties.

No-false-positive: For all $0 \leq k \leq \bar{k}$, the solution path of GLBI has no false-positive, i.e. $\beta_{k,S^c} = 0$.

Sign consistency: If $\bar{k} \geq 5$ and

$$\beta^\star_{\min} := \min_{j \in S} \left|\beta^\star_j\right| \geq \max\left(2\left\|\beta^o_S - \beta^\star_S\right\|_\infty, \right.$$
$$\left. \frac{(8\log s + 18)(C+1)}{\lambda'\eta\left(1 - 4/\bar{k}\right)}\left\|\nabla\ell\left(\alpha^\star, \beta^\star\right)\right\|_\infty\right), \quad (2.5)$$

then $\operatorname{sign}(\beta_{\bar{k}}) = \operatorname{sign}(\beta^\star)$.

ℓ_2 consistency: The ℓ_2 error

$$\left\|\begin{pmatrix}\alpha_{\bar{k}} - \alpha^\star \\ \beta_{\bar{k}} - \beta^\star\end{pmatrix}\right\|_2 \leq \frac{20(C+1)\sqrt{s}}{\lambda'\eta}\left\|\nabla\ell\left(\alpha^\star, \beta^\star\right)\right\|_\infty.$$

The proof of Theorem 1 is collected in Appendix C, which largely follows the analysis of differential inclusion (1.6), given in Appendix B, as its discretization.

Remark 3. For sparse linear regression problem (Example 1), with high probability we have

$$\left\|\nabla\ell\left(\alpha^\star, \beta^\star\right)\right\|_\infty \lesssim \sqrt{\frac{\log p}{n}}, \; \left\|\beta^o_S - \beta^\star_S\right\|_\infty \lesssim \sqrt{\frac{\log s}{n}}.$$

Hence pick $\bar{k}\delta \sim \sqrt{n/\log p}$ satisfying (2.4), by Theorem 1 the sign consistency is guaranteed at $\bar{k}$ if

$$\beta^\star_{\min} \gtrsim (\log s)\sqrt{(\log p)/n}.$$

The ℓ_2 error bound reaches the minimax optimal rate:

$$\left\|\begin{pmatrix}\alpha_{\bar{k}} - \alpha^o \\ \beta_{\bar{k}} - \beta^o\end{pmatrix}\right\|_2 \lesssim \sqrt{\frac{s\log p}{n}}.$$

Remark 4. For sparse logistic regression problem (Example 2), with high probability we have

$$\begin{aligned}&\left\|\nabla\ell\left(\alpha^\star, \beta^\star\right)\right\|_\infty \lesssim \sqrt{(\log p)/n}, \\ &\quad \left\|\beta^o_S - \beta^\star_S\right\|_\infty \lesssim \left\|\beta^o_S - \beta^\star_S\right\|_2 \lesssim \sqrt{s}\left\|\nabla\ell\left(\alpha^\star, \beta^\star\right)\right\|_\infty \\ &\quad\quad \lesssim \sqrt{(s\log p)/n}.\end{aligned}$$

Hence the sign consistency is guaranteed at some $\bar{k} \sim \sqrt{n/\log p}$ if

$$\beta^\star_{\min} \gtrsim \sqrt{(s\log p)/n}$$

(meeting Condition (19) in Ravikumar et al. (2010)). The ℓ_2 error rate $\lesssim \sqrt{(s\log p)/n}$ is minimax optimal.

3 Experiments

As for the setting of algorithm parameters: κ should be large, and then $\delta \sim 1/(\kappa\Lambda)$ is automatically calculated based on κ (as long as $\kappa\delta\Lambda < 2$, such that λ' is positive in (C.3)). In practice, a small δ can prevent the iterations from oscillations.

3.1 Efficiency of Parallel Computing

Osher et al. (2016) has elaborated that LBI can easily be implemented in parallel and distributed manners, and applied on very large-scale datasets. Likewise, GLBI can be parallelized in many usual applications. We now take the logistic model Example 2 as an example to explain the details. The iteration (1.6) (generally taking $\delta_k = \delta$) can be written as

$$\begin{aligned}\alpha_{k+1} &= \alpha_k - \kappa\delta f\left(\alpha_k, X\beta_k\right), && (3.1a)\\ z_{k+1} &= z_k - \delta X^T g\left(\alpha_k, X\beta_k\right), && (3.1b)\\ \beta_{k+1} &= \kappa\mathcal{S}\left(z_{k+1}, 1\right), && (3.1c)\end{aligned}$$

where $f: \mathbb{R}^{n+1} \to \mathbb{R}, \; g: \mathbb{R}^{n+1} \to \mathbb{R}^n$ such that

$$g\left(\alpha, w\right)_i := -\frac{1}{n} \cdot \frac{1}{1 + \exp\left(-\left(\alpha + w_i\right)y^{(i)}\right)} y^{(i)} \in \mathbb{R}$$
$$(1 \leq i \leq n; \; w \in \mathbb{R}^n)$$
$$\begin{aligned}f\left(\alpha, w\right) &:= 1_n^T \cdot g(\alpha, w) \\ &= -\frac{1}{n}\sum_{i=1}^n \frac{1}{1 + \exp\left(-\left(\alpha + w_i\right)y^{(i)}\right)} y^{(i)}.\end{aligned}$$

Suppose

$$X = \left[X_1, X_2, \ldots, X_L\right] \in \mathbb{R}^{n \times p},$$

where X_l's are submatrices stored in a distributed manner on a set of networked workstations. The sizes of X_l's are flexible and can be chosen for good load balancing. Let each workstation l hold data y and X_l, and variables $z_{k,l}$ and $X_l\beta_{k,l}$ which are parts of z_k and summands of $w_k := X\beta_k$, respectively. The iteration (3.1) is carried out as

$$\begin{cases}\alpha_{k+1} = \alpha_k - \kappa\delta f(\alpha_k, w_k), \\ z_{k+1,l} = z_{k,l} - \delta X_l^T g(\alpha_k, w_k), \quad \text{in parallel for } l \\ w_{k+1,l} = \kappa X_l \mathcal{S}(z_{k+1,l}, 1)\end{cases}$$
$$w_{k+1} = \sum_{l=1}^L w_{k+1,l} \quad \text{(all-reduce summation)},$$

where the all-reduce summation step collects inputs from and then returns the sum to all the L workstations. It is the sum of L n-dimensional vectors. Therefore, the communication cost is independent of p no matter how the all-reduce step is implemented. It is important to note that the algorithm is not changed at all. particularly, increasing L, does not increase the number of iterations. So the parallel implementation is truly scalable.

If t_L denotes the time cost of a single GLBI run with L workstations under the same dataset and the same algorithmic settings, it is expected that $t_L \sim 1/L$.

Here we show this by an example. Construct a logistic model in Section 3.2 with $M = 1$, $r = 0.25$, and three settings for (p, s, n): (I) (2000, 200, 6000), (II) (5000, 500, 15000), (III) (10000, 1000, 30000). For each setting, we run our parallelized version of GLBI algorithm written in C++, with $\kappa = 10$, $\delta = 0.1$, $k_{\max} = 1000k_0$, where k_0 is the maximal k such that $\beta_1 = \cdots = \beta_k = 0$, and the path is early stopped at the $k_{\max}$-th iteration. The recorded t_L's are shown in Figure 3. The left panel shows t_L (in seconds) while the right panel shows t_1/t_L, for $L = 1, \ldots, 8$. We see truly $t_L \sim 1/L$, which is expected in our parallel and distributed treatment. When L is large, our package can deal with very large scale problems.

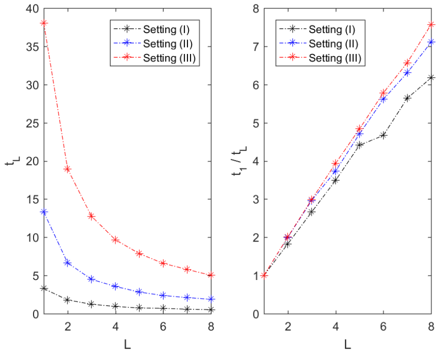

Figure 3: Time cost illustration for logistic model with three settings (black for Setting (I), blue for Setting (II) and red for Setting (III)). In each setting, the left panel shows t_L while the right panel shows t_1/t_L, for $L = 1, \ldots, 8$.

3.2 Application: Logistic Model

We do rep = 20 independent experiments, in each of which we construct a logistic model (Example 2), and then compare GLBI with other methods. Specifically, suppose that $\beta^\star$ has a support set $S = \{1, \ldots, s\}$ without loss of generality. $\alpha^\star, \beta_j^\star$ $(j \in S)$ are independent, each has a uniform distribution on $[-2M, -M] \cup [M, 2M]$. Each row of $X \in \mathbb{R}^{n \times p}$ is i.i.d. sampled from $N(0, \Sigma)$, where Σ is a Toeplitz matrix satisfying $\Sigma_{j,k} = r^{|j-k|}$. When X and $(\alpha^\star, \beta^\star)$ are determined, we generate $y \in \mathbb{R}^n$ as in Example 2.

After getting the sample (X, y), consider GLBI (1.6) and ℓ_1 optimization (1.1), both with logistic loss (1.2). For GLBI, set $\kappa = 10$. For (1.1), apply a grid search for differently penalized problems, for which we use `glmnet` – a popular package available in Matlab/R that can be applied on ℓ_1 regularization for sparse logistic regression models.

For each algorithm, we use K-fold ($K = 5$) cross validation (CV) to pick up an estimator from the path calculate based on the smallest CV estimate of *prediction error*. Specifically, we split the data into K roughly equal-sized parts. For a certain position on paths (t for GLBI, or λ for `glmnet`) and $k \in \{1, \ldots, K\}$, we obtain a corresponding estimator based on the data with the k-th part removed, use the estimator to build a classifier, and get the mis-classification error on the k-th part. Averaging the value for $k \in \{1, \ldots, K\}$, we obtain the CV estimate of prediction error, for the obtained estimator corresponding to a certain position on paths. Among all positions, we pick up the estimator producing the smallest CV estimate of prediction error. Besides, we calculate *AUC (Area Under Curve)*, for evaluating the path performance of learning sparsity patterns without choosing a best estimator.

Results for $p = 80$, $s = 20$, $M = 1$ are summarized in Table 1. We see that in terms of CV estimate of prediction error, GLBI is generally better than `glmnet`. Besides, GLBI is competitive with ℓ_1 regularization method in variable selection, in terms of AUC. Similar observations for more settings are listed in Table 4, 5 and 6 in Appendix G. Apart from these tables, we can also see the outperformance of CV estimate of prediction error Figure 5 in Appendix G, while in that figure we can see that GLBI further reduces bias, as well as provides us a relatively good estimator with small prediction error if a proper early stopping is equipped.

Table 1: Comparisons between GLBI and `glmnet`, for logistic models with $p = 80$, $s = 20$, $M = 1$. For each algorithm, we run rep = 20 independent experiments.

		AUC		prediction error	
r	n	GLBI	`glmnet`	GLBI	`glmnet`
0.25	400	.9902	**.9906**	**.1221**	.1355
		(.0065)	(.0062)	(.0218)	(.0223)
	800	**.9991**	.9990	**.1082**	.1132
		(.0020)	(.0022)	(.0125)	(.0104)
0.5	400	**.9690**	.9681	**.1321**	.1379
		(.0180)	(.0165)	(.0268)	(.0289)
	800	**.9925**	.9921	**.1139**	.1197
		(.0069)	(.0076)	(.0138)	(.0134)

3.3 Application: Ising Model with 4-Nearest-Neighbor Grid

We do rep = 20 independent experiments, in each of which we construct an ising model (Example 3), and then compare GLBI with other methods. Specifically, construct an $N \times N$ 4-nearest neighbor grid (with aperiodic boundary conditions) to be graph G, with node set V and edge set E. The distribution of a random vector x is given by (1.3) ($p = N^2$), where $\alpha_j^\star$'s and

Table 2: Comparisons of GLBI1 (GLBI + composite), GLBI2 (GLBI + MPF), and `glmnet`, for Ising models with $p = 36$. For each algorithm, we run rep = 20 independent experiments.

		AUC		
T	n	GLBI1	GLBI2	`glmnet`
1.25	500	.9754 (.0277)	**.9867** (.0128)	.9774 (.0265)
	750	.9868 (.0137)	**.9919** (.0082)	.9891 (.0134)
1.5	500	.9915 (.0110)	**.9963** (.0033)	.9929 (.0104)
	750	.9963 (.0041)	**.9980** (.0029)	.9975 (.0042)

		2nd order MDC		
T	n	GLBI1	GLBI2	`glmnet`
1.25	500	**.9762** (.0079)	.9758 (.0079)	.9744 (.0086)
	750	**.9840** (.0053)	.9830 (.0066)	.9827 (.0061)
1.5	500	**.9655** (.0087)	.9646 (.0099)	.9630 (.0094)
	750	**.9774** (.0060)	.9766 (.0066)	.9756 (.0070)

$\beta^\star_{j,j'}$'s ($(j,j') \in E$) are i.i.d. and each has a uniform distribution on $[-2/T, -1/T] \cup [1/T, 2/T]$. Let $X \in \mathbb{R}^{n\times p}$ represents n samples drawn from the distribution of x via Gibbs sampling.

After getting the sample X, consider GLBI1 (GLBI with composite loss (1.4)), GLBI2 (GLBI with MPF loss (1.5)) and ℓ_1 optimization (1.1) with logistic loss (see Example 3 for *neighborhood-based logistic regression* applied on Ising models, or see Ravikumar et al. (2010)). For GLBI1 and GLBI2, set $\kappa = 10$. For (1.1), apply a grid search for differently penalized problems; we still use `glmnet`.

For each algorithm, we calculate the *AUC (Area Under Curve)*, popular for evaluating the path performance of learning sparsity patterns. Besides, we apply K-fold ($K = 5$) cross validation (CV) to pick up an estimator from the path, with the *largest CV estimate of 2nd order marginal distribution correlation (2nd order MDC)* in the same way as the CV process done in Section 3.2, here the 2nd order MDC, defined in the next paragraph, is calculated based on two samples of the same size: the k-th part original data, and the newly sampled Ising model data (based on learned parameters) with the same size of the k-th part.

For any sample matrix $X' \in \{1,-1\}^{n'\times p}$, we construct $d_2(X')$, the 2nd marginal empirical distribution matrix of X', defined as follows. $d_2(X') = (d_2(X')_{[j_1,j_2]})_{p\times p} \in \mathbb{R}^{2p\times 2p}$, where

$$d_2\left(X'\right)_{[j_1,j_2]} = \frac{1}{n'}\sum_{i=1}^{n'}\begin{pmatrix} 1_{\left(x^{(i)}_{j_1},x^{(i)}_{j_2}\right)=(1,1)} & 1_{\left(x^{(i)}_{j_1},x^{(i)}_{j_2}\right)=(1,-1)} \\ 1_{\left(x^{(i)}_{j_1},x^{(i)}_{j_2}\right)=(-1,1)} & 1_{\left(x^{(i)}_{j_1},x^{(i)}_{j_2}\right)=(-1,-1)} \end{pmatrix}. \tag{3.2}$$

For any sample matrices X_1, X_2 with the same sample size, we call the correlation between $\mathrm{vec}(d_2(X_1))$ and $\mathrm{vec}(d_2(X_2))$ *the 2nd order marginal distribution correlation (2nd order MDC)*. This value is expected to be large as well as close to 1 if X_1, X_2 come from the same model.

Results for $p = N^2 = 36$ are summarized in Table 2. GLBI with composite/MPF loss are competitive with or better than `glmnet`. Similar observations are listed in Table 7 in Appendix G.

3.4 Application: Coauthorship Network in NIPS

Consider the information of papers and authors in *Advances in Neural Information Processing Systems* (NIPS) 1987–2016, collected from `https://www.kaggle.com/benhamner/nips-papers`. After preprocessing (e.g. author disambiguity), for simplicity, we restrict our analysis on the most productive $p = 30$ authors (Table 3) in the largest connected component of a coauthorship network that two authors are linked if they coauthored at least 2 papers (Coauthorship (2)). The first panel of Figure 4 shows this coauthorship network with edge width in proportion to the number of coauthored papers. There are $n = 1,028$ papers authored by at least one of these persons.

Let the j-th entry of $x^{(i)} \in \mathbb{R}^p$ be 1 if the j-th person is involved in the authors of the i-th paper, and -1 otherwise. Now we fit the data $x^{(i)}$ $(1 \le i \le n)$ by a sparse Ising model (1.3) with parameter $(\hat{\alpha}, \hat{\beta})$. Note that $\hat{\beta}_{j,j'} = 0$ indicates that j and j' are conditional independent on coauthorship, given all the other authors; $\hat{\beta}_{j,j'} > 0$ implies that j and j' coauthored more often than their averages, while $\hat{\beta}_{j,j'} < 0$ says the opposite.

The right three panels in Figure 4 compares some sparse Ising models chosen from three regularization paths at a similar sparsity level (the percentage of learned edges over the complete graph, here about $12\% \sim 14\%$): GLBI1 (GLBI with composite loss), GLBI2 (GLBI with MPF loss), and ℓ_1 regularization

Table 3: Most productive $p = 30$ authors in the largest connected component of Coauthorship (2).

01 Michael Jordan	16 Inderjit Dhillon
02 Bernhard Schölkopf	17 Ruslan Salakhutdinov
03 Geoffrey Hinton	18 Tong Zhang
04 Yoshua Bengio	19 Thomas Griffiths
05 Zoubin Ghahramani	20 David Blei
06 Terrence Sejnowski	21 Rémi Munos
07 Peter Dayan	22 Joshua Tenenbaum
08 Alex Smola	23 Lawrence Carin
09 Andrew Ng	24 Eric Xing
10 Francis Bach	25 Richard Zemel
11 Michael Mozer	26 Martin Wainwright
12 Pradeep Ravikumar	27 Yoram Singer
13 Tommi Jaakkola	28 Han Liu
14 Klaus-Robert Müller	29 Satinder Singh
15 Yee Teh	30 Christopher Williams

(glmnet), respectively. For more learned graphs from these paths, see Figure 6 in Appendix G. In GLBI1 and GLBI2, set $\kappa = 10$.

We see that all the learned graphs capture some important coauthorships, such as Pradeep Ravikumar (12) and Inderjit Dhillon (16) in a thick green edge in all the three learned graphs, indicating that they collaborated more often than separately for NIPS. Besides, the most productive author Michael Jordan (01) has coauthored with a lot of other people, but is somewhat unlikely to coauthor with several other productive scholars like Yoshua Bengio (04), Terrence Sejnowski (06), etc., indicating by the red edges between Jordan and those people. Further note the edge widths in the second and third graphs are significantly larger than those in the fourth graph, implying that at a similar sparsity level, GLBI tends to provide an estimator with larger absolute values of entries than that by glmnet. That is because under similar sparsity patterns, GLBI may give *low-biased* estimators.

Acknowledgements

This work is supported in part by National Basic Research Program of China (Nos. 2015CB85600, 2012CB825501), NNSF of China (Nos. 61370004, 11421110001), HKRGC grant 16303817, as well as grants from Tencent AI Lab, Si Family Foundation, Baidu BDI, and Microsoft Research-Asia.

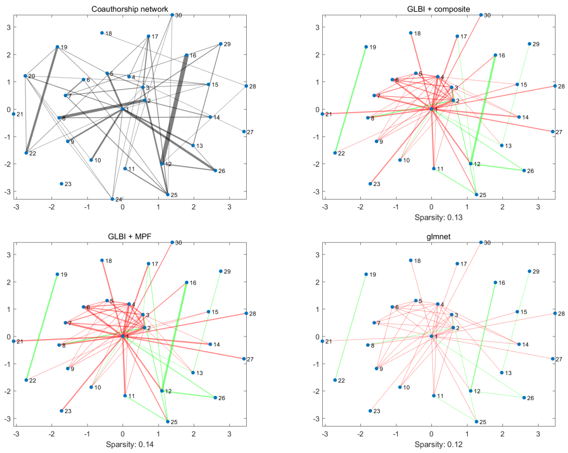

Figure 4: Top left: NIPS coauthorship network, with edge width in proportion to the number of coauthored papers. Top right: a learned graph picked from the path of GLBI1. Bottom left: from GLBI2. Bottom right: from glmnet. Green edges indicate positive conditional dependence of coauthorship – the probability of coauthoring a paper significantly increases the authors' average behavior, while red edges indicating the negative coauthorship. Edge widths show the strength of such a relationship.

References

Beck, Amir and Marc Teboulle (2003). “Mirror descent and nonlinear projected subgradient methods for convex optimization”. In: *Operations Research Letters* 31, pp. 167–175.

Bickel, Peter J., Ya’acov Ritov, and Alexandre B. Tsybakov (2009). “Simultaneous analysis of Lasso and Dantzig selector”. In: *Ann. Statist.* 37.4, pp. 1705–1732.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani (2004). “Least Angle Regression”. In: *Annals of Statistics* 32.2, pp. 407–499.

Huang, Chendi, Xinwei Sun, Jiechao Xiong, and Yuan Yao (2016). “Split LBI: An Iterative Regularization Path with Structural Sparsity”. In: *Advances in Neural Information Processing Systems (NIPS) 29*, pp. 3369–3377.

Negahban, Sahand, Bin Yu, Martin J Wainwright, and Pradeep K. Ravikumar (2009). “A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers”. In: *Advances in Neural Information Processing Systems (NIPS) 22*, pp. 1348–1356.

Nemirovski, Arkadi (2012). “Tutorial: Mirror Descent Algorithms for Large-Scale Deterministic and Stochastic Convex Optimization”. In: *Conference on Learning Theory (COLT)*.

Nemirovski, Arkadi and David Yudin (1983). *Problem complexity and Method Efficiency in Optimization.* Nauka Publishers, Moscow (in Russian), 1978. New York: Wiley.

Osher, Stanley, Feng Ruan, Jiechao Xiong, Yuan Yao, and Wotao Yin (2016). “Sparse recovery via differential inclusions”. In: *Applied and Computational Harmonic Analysis* 41.2, pp. 436–469. ISSN: 1063-5203.

Ravikumar, P., G. Raskutti, M. Wainwright, and B. Yu (2008). “Model selection in Gaussian graphical models: High-dimensional consistency of l1-regularized MLE”. In: *Advances in Neural Information Processing Systems (NIPS)*. Vol. 21.

Ravikumar, Pradeep, Martin J Wainwright, and John D Lafferty (2010). “High-dimensional Ising model selection using l1-regularized logistic regression”. In: *The Annals of Statistics* 38.3, pp. 1287–1319.

Shi, Jianing V., Wotao Yin, and Stanley J. Osher (2013). “A New Regularization Path for Logistic Regression via Linearized Bregman”. In: URL: ftp://ftp.math.ucla.edu/pub/camreport/cam12-67.pdf.

Sohl-Dickstein, Jascha, Peter Battaglino, and Michael DeWeese (2011). “Minimum Probability Flow Learning”. In: ed. by Lise Getoor and Tobias Scheffer. ICML ’11. New York, NY, USA: ACM, pp. 905–912.

Wainwright, Martin J. (2009). “Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using L1-Constrained Quadratic Programming (Lasso)”. In: *Information Theory, IEEE Transactions on* 55.5, pp. 2183–2202.

Xue, Lingzhou, Hui Zou, and Tianxi Cai (2012). “Nonconcave penalized composite conditional likelihood estimation of sparse Ising models”. In: *The Annals of Statistics* 40.3, pp. 1403–1429.

Yin, Wotao, Stanley Osher, Jerome Darbon, and Donald Goldfarb (2008). “Bregman Iterative Algorithms for Compressed Sensing and Related Problems”. In: *SIAM Journal on Imaging Sciences* 1.1, pp. 143–168.

Yuan, Ming and Yi Lin (2007). “Model Selection and Estimation in the Gaussian Graphical Model”. In: *Biometrika* 94, pp. 19–35.

Zhao, Peng and Bin Yu (2006). “On Model Selection Consistency of Lasso”. In: *J. Machine Learning Research* 7, pp. 2541–2567.