Latent Feature Lasso

Ian E.H. Yen [1] **Wei-Cheng Lee** [2] **Sung-En Chang** [2] **Arun S. Suggala** [1] **Shou-De Lin** [2] **Pradeep Ravikumar** [1]

Abstract

The latent feature model (LFM), proposed in (Griffiths & Ghahramani, 2005), but possibly with earlier origins, is a generalization of a mixture model, where each instance is generated not from a single latent class but from a combination of *latent features*. Thus, each instance has an associated latent binary feature incidence vector indicating the presence or absence of a feature. Due to its combinatorial nature, inference of LFMs is considerably intractable, and accordingly, most of the attention has focused on nonparametric LFMs, with priors such as the Indian Buffet Process (IBP) on infinite binary matrices. Recent efforts to tackle this complexity either still have computational complexity that is exponential, or sample complexity that is high-order polynomial w.r.t. the number of latent features. In this paper, we address this outstanding problem of tractable estimation of LFMs via a novel atomic-norm regularization, which gives an algorithm with polynomial run-time and sample complexity without impractical assumptions on the data distribution.

1. Introduction

Latent variable models are widely used in unsupervised learning, in part because they provide compact and interpretable representations of the distribution over the observed data. The most common and simplest such latent variable model is a mixture model, which associates each observed object with a *latent class*. However, in many real-world applications, observations are better described by a combination of *latent features* than a single latent class. Accordingly, admixture or mixed membership models have been proposed (Airoldi et al., 2014), that in the simplest settings, assign each object to a convex combination of latent classes. For instance, a document object could be modeled as a convex combination of topic objects. There are many settings however where a convex combination might be too restrictive, and the objects are better modeled as simply a collection of latent classes. An example is web image, which can often described by multiple tags rather than a single class, or even by a convex combination of tag objects. Another example is the model of user, who might have multiple interests in the context of a recommendation system, or be involved in multiple communities in a social network. With such settings in mind, (Griffiths & Ghahramani, 2005) proposed a latent feature model (LFM), where each observed object can be represented by a binary vector that indicates the presence or absence of each of a collection of *latent features*. Their proposed model extended earlier models with a similar flavor for specialized settings, such as (Ueda & Saito, 2003) for bag of words models for text. The latent feature model can also be connected to sparse PCA models (d'Aspremont et al., 2007; Jolliffe et al., 2003) by considering a pointwise product of the binary feature incidence vector with another real-valued vector. As (Griffiths & Ghahramani, 2005) showed, LFM handily outperforms clustering as an efficient and interpretable data representation, particularly in settings where the object can be naturally represented as a collection of latent features or parts.

However, the estimation (inference) of an LFM from data is difficult, due to the combinatorial nature of the binary feature incidence vectors. Indeed, with N samples, and K latent features, the number of possible binary matrices consisting of the N binary feature incidence vectors is 2^{NK}. And not in the least, the log-likelihood of LFM is not a concave function of its parameters.

Given that the finite feature case seems intractable, right from the outset, attention has focused on the nonparametric infinite feature case, where a prior known as the *Indian Buffet Process (IBP)* has been proposed for the infinite binary matrices consisting of the feature incidence vectors given infinite set of latent features (Griffiths & Ghahramani, 2011). While the IBP prior provides useful structure, inference remains a difficult problem, and in practice, one often relies on local search methods (Broderick et al., 2013) to find an estimate of parameters, or employ Markov Chain Monte Carlo (MCMC) (Doshi-Velez & Ghahramani, 2009)

[1]Carnegie Mellon University, U.S.A. [2]National Taiwan University, Taiwan. Correspondence to: Ian E.H. Yen <eyan@cs.cmu.edu>.

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

or variational methods (Doshi-Velez et al., 2009) to obtain an approximate posterior distribution. However, none of these approaches can provide guarantees on the quality of solution in polynomial time.

Note that both in the mixture model, as well as the admixture model cases, the parametric variants have been hugely popular alongside or perhaps even more so than the nonparametric variants e.g. clustering procedures based on finite number of clusters, or topic models with a finite number of topics. This is in part because the parametric variants have a lower model complexity, which might be desired under certain settings, and also have simpler inference procedures. However, in the LFM case, the parametric variant has received very little attention, which might suggest the relatively lesser popularity for LFMs when compared to mixture or admixture/topic models.

Accordingly, in this paper, we consider the question of computationally tractable estimation of parametric LFMs. In the nonparametric setting with an IBP prior, (Tung & Smola, 2014) have proposed the use of spectral methods, which bypasses the problem of non-concave log-likelihood by estimating the *moments* derived from the model, and then recovers parameters by solving a system of equations. Their spectral methods based procedure produces consistent estimates of LFMs in polynomial time, however with a sample complexity that has a high-order (more than six-order) polynomial dependency on the number of latent features and the occurrence probability of each feature. Moreover, the application of spectral methods requires knowledge of the distribution, which results in non-robustness to model mis-specification in practice. Under a noiseless setting, (Slawski et al., 2013) leveraged identifiability conditions under which the solution is unique, to propose an algorithm for a parametric LFM. Their algorithm is guaranteed to recover the parameters in the noiseless setting, but with the caveat that it has a computational complexity that is *exponential* in the number of latent features.

We note that even under the assumption of an nonparametric LFM, specifically an *Indian Buffet Process with Linear Gaussian Observations*, deriving its MAP point estimate under low-variance asymptotics following the approach of *MAD-Bayes Asymptotics* (Broderick et al., 2013) yields an objective similar to that of a parametric LFM with an additional term that is linear in the number of latent features. Thus, developing computationally tractable approaches for parametric LFMs would be broadly useful.

In this work, we propose the *Latent Feature Lasso*, a novel convex estimation procedure for the estimation of a Latent Feature Model using atomic-norm regularization. We construct a greedy algorithm with strong optimization guarantees for the estimator by relating each greedy step to a MAX-CUT like problem. We also provide a risk bound for the estimator under general data distribution settings, which trades off between risk and sparsity, and has a sample complexity linear in the number of components and dimension. Under the noiseless setting, we also show that Latent Feature Lasso estimator recovers the parameters of LFM under an identifiability condition similar to (Slawski et al., 2013).

2. Problem Setup

A Latent Feature Model represents data as a combination of *latent features*. Let $x \in \mathbb{R}^D$ be an observed random vector that is generated as:

$$x = W^T z + e,$$

where $z \in \{0,1\}^K$ is a latent binary feature incidence vector that denotes the presence or absence of K features, $W \in \mathbb{R}^{K \times D}$ is an unknown matrix of K latent features of dimension D, and $e \in \mathbb{R}^D$ is an unknown noise vector. We say that the model is biased when $E[e|z] = E[x|z] - W^T z \neq 0$, and which we allow in our analysis. Suppose we observe N samples of the random vector x. It will be useful in the sequel to collate the various vectors corresponding to the N samples into matrices. We collate the observations into a matrix $X \in \mathbb{R}^{N \times D}$, the N latent incidence vectors into a matrix $Z \in \{0,1\}^{N \times K}$, and the noise vectors into an $N \times D$ matrix ϵ. We thus obtained the vectorized form of the model as $X = ZW + \epsilon$.

Most existing works on LFM make two strong assumptions. The first is that the model has zero bias $E[e|z] = 0$ (Tung & Smola, 2014; Broderick et al., 2013; Griffiths & Ghahramani, 2011; Slawski et al., 2013; Zoubin, 2013; Doshi-Velez & Ghahramani, 2009; Doshi-Velez et al., 2009; Hayashi & Fujimaki, 2013). The second common but strong class of assumptions is distributional (Hayashi & Fujimaki, 2013; Tung & Smola, 2014):

$$p(x|z) = N(W^T z, \sigma^2 I)\ ,\ p(z) = Bern(\pi),$$

where $Bern(\pi)$ denotes the distribution of K independent Bernoulli with $z_k \sim Bern(\pi_k)$. In the Nonparametric Bayesian setting (Griffiths & Ghahramani, 2011; Zoubin, 2013; Doshi-Velez et al., 2009; Doshi-Velez & Ghahramani, 2009; Broderick et al., 2013), one replaces $Bern(\pi)$ with an *Indian Buffet Process* IBP(α) over the $N \times K^+$ binary incidence matrix $Z \in \{0,1\}^{N \times K^+}$ where K^+ can be inferred from data instead of being specified a-priori. We note that both classes of assumptions need not hold in practice: the zero bias assumption $E[x|z] = W^T z$ is stringent given the linearity of the model, while the Bernoulli and IBP distributional assumptions are also restrictive, in part since they assume independence between the presence of two features z_{ik} and $z_{ik'}$. Our method and analyses do not impose either of these assumptions.

It is useful to contrast the different estimation goals ranging over the LFM literature. In the Bayesian approach line of work (Griffiths & Ghahramani, 2011; Broderick et al., 2013; Zoubin, 2013; Doshi-Velez & Ghahramani, 2009; Hayashi & Fujimaki, 2013), the goal is to infer the posterior distribution $P(Z, W|X)$ given X. The line of work using Spectral Methods (Tung & Smola, 2014) on the other hand aim to estimate $p(z)$, $p(x|z)$ in turn by estimating parameters (π, W). In some other work (Slawski et al., 2013), they aim to estimate W, leaving the distribution of z unmodeled. In this paper, we focus on the more realistic setting where we make *no assumption* on $p(x)$ except that of boundedness, and aim to find an LFM W^* that minimizes the risk

$$r(W) := E[\min_{z\in\{0,1\}^K} \frac{1}{2}\|x - W^T z\|^2]. \quad (1)$$

where the expectation is over the random observation x.

3. Latent Feature Lasso

We first consider the non-convex formulation that was also previously studied in (Broderick et al., 2013) as asymptotics of the MAP estimator of IBP Linear-Gaussian model:

$$\min_{K\in\mathbb{N}, Z\in\{0,1\}^{N\times K}, W\in\mathbb{R}^{K\times D}} \frac{1}{2N}\|X - ZW\|_F^2 + \lambda K. \quad (2)$$

The estimation problem in (Slawski et al., 2013) could also be cast in the above form with $\lambda = 0$ and K treated as a fixed hyper-parameter, while (Broderick et al., 2013) treats K as a variable and controls it through λ. (2) is a combinatorial optimization of $N \times K + 1$ integer variables. In the following we develop a tight convex approximation to (2) with ℓ_2 regularization on W, by introducing a type of atomic norm (Chandrasekaran et al., 2012).

For a fixed K, Z, consider the minimization over W of the ℓ_2 regularized version of (2)

$$\min_{W\in\mathbb{R}^{K\times D}} \frac{1}{2N}\|X - ZW\|_F^2 + \frac{\tau}{2}\|W\|_F^2, \quad (3)$$

which is a convex minimization problem. Applying Lagrangian duality to (3) results in the following dual form

$$\max_{A\in\mathbb{R}^{N\times D}} \left\{ \frac{-1}{2N^2\tau} tr(AA^T M) - \frac{1}{N}\sum_{i=1}^N L^*(x_i, -A_{i,:}) \right\}. \quad (4)$$

where $M := ZZ^T$, $A \in \mathbb{R}^{N\times D}$ are dual variables that satisfy $W^* = \frac{1}{N} Z^* A^*$ at the optimum of (3) and (4), and $L^*(x, \alpha) = \langle x, \alpha\rangle + \frac{1}{2}\|\alpha\|^2$ is the convex conjugate of square loss $L(x, \xi) = \frac{1}{2}\|x - \xi\|^2$ w.r.t. its second argument.

Let $G(M, A)$ denote the objective in (4) for any fixed M, and let $g(M) = \max_A G(M, A)$ denote the optimal value of the objective when optimized over A. The objective in (2) for a fixed K could thus be simply reformulated as a minimization of this dual-derived objective $g(M)$. It can be seen that $g(M)$ is a convex function w.r.t. M since it is the maximum of linear functions of M. The key caveat however is the combinatorial structure on M since it has the form $M = ZZ^T$, $Z \in \{0,1\}^{N\times K}$. We address this caveat by introducing the following atomic norm

$$\|M\|_{\mathcal{S}} := \min_{c\geq 0} \sum_{a\in\mathcal{S}}^K c_a \; s.t. \; M = \sum_{a\in\mathcal{S}} c_a a. \quad (5)$$

with $\mathcal{S} := \{zz^T | z \in \{0,1\}^N\}$. Note $\|M\|_{\mathcal{S}} = \sum_{a\in\mathcal{S}} c_a = K$ when c_a in (5) are constrained at integer value $\{0, 1\}$, and it serves a convex approximation to K similar to the ℓ_1-norm used in *Lasso* for the approximation of cardinality. This results in the following *Latent Feature Lasso* estimator

$$\min_M \; \{g(M) + \lambda\|M\|_{\mathcal{S}}\}. \quad (6)$$

Algorithm 1 A Greedy Algorithm for Latent Feature Lasso

0: $\mathcal{A} = \emptyset$, $c = 0$.
for $t = 1...T$ **do**
1: Find a greedy atom zz^T by solving (8).
2: Add zz^T to an active set $\mathcal{A}$.
3: Minimize (7) w.r.t. coordinates in $\mathcal{A}$ via updates (9).
4: Eliminate $\{z_j z_j^T | c_j = 0\}$ from $\mathcal{A}$.
end for.

4. Algorithm

The estimator (6) seems intractable at first sight in part since the atomic norm involves a set $\mathcal{S}$ of 2^N atoms. In this section, we study a variant of approximate greedy coordinate descent method for tractably solving problem (6). We begin by rewriting the optimization problem (6) as an ℓ_1-regularized problem with $\bar{K} = 2^N - 1$ coordinates, by expanding the matrix M in terms of the $\bar{K}$ atoms underlying the atomic norm $\|.\|_{\mathcal{S}}$:

$$\min_{c\in\mathbb{R}_+^{\bar{K}}} \underbrace{\left\{ \underbrace{g\left(\sum_{j=1}^{\bar{K}} c_j z_j z_j^T\right)}_{f(c)} + \lambda\|c\|_1 \right\}}_{F(c)} \quad (7)$$

where $\{z_j\}_{j=1}^{\bar{K}}$ enumerates all possible $\{0,1\}^N$ patterns except the 0 vector. Our overall algorithm is depicted in Algorithm 1. In each iteration, it finds

$$\begin{aligned} j^* &:= arg\max_j \; -\nabla_j f(c) \\ &= arg\max_j \langle -\nabla g(M), z_j z_j^T\rangle \end{aligned} \quad (8)$$

approximately with a constant approximation ratio via a reduction to a MAX-CUT-like problem (see Section 4.1). An active set $\mathcal{A}$ is maintained to contain all atoms $z_j z_j^T$ with non-zero coefficients c_j and the atom returned by the greedy search (8). Then we minimize (7) over coordinates in $\mathcal{A}$ by a sequence of proximal updates:

$$c^{r+1} \leftarrow \left[c^r - \frac{\nabla f(c^r) + \lambda}{\gamma |\mathcal{A}|} \right]_+, \; r = 1...T_2 \tag{9}$$

where γ is the Lipschitz-continuous constant of the coordinate-wise gradient $\nabla_{c_j} f(c)$.

Computing cooordinate-wise gradients. By Danskin's Theorem, the gradient of function $f(c)$ takes the form

$$\nabla_{c_j} f(c) = z_j A^* A^{*T} z_j / (2N^2 \tau), \tag{10}$$

which in turn requires finding the maximizer A^* of (4).

Computing A^*. By taking advantage of the strong duality between (4) and (3), the maximizer A^* can be found by finding the minimizer W^* of

$$\min_W \; \frac{1}{2N} \|X - Z_{\mathcal{A}} W\|_F^2 + \sum_{k \in \mathcal{A}} \frac{\tau}{2c_k} \|W_{k,:}\|^2 \tag{11}$$

and computing $A^* = (X - Z_{\mathcal{A}} W^*)$, where $Z_{\mathcal{A}}$ denotes $N \times |\mathcal{A}|$ matrix of columns taking from the active atom basis $\{z_k\}_{k \in \mathcal{A}}$.

Computing W^*. There is a closed-form solution W^* to (11) of the form

$$W^* = (Z_{\mathcal{A}}^T Z_{\mathcal{A}} + N\tau \text{diag}^{-1}(c_{\mathcal{A}}))^{-1} Z_{\mathcal{A}}^T X. \tag{12}$$

An efficient way of computing (12) is to maintain $Z_{\mathcal{A}}^T Z_{\mathcal{A}}$ and $Z_{\mathcal{A}}^T X$ whenever the active set of atoms $\mathcal{A}$ changes. This has a cost of $O(NDK_{\mathcal{A}})$ for a bound $K_{\mathcal{A}}$ on the active size, which however is almost neglectable compared to the other costs when amortized over iterations. Then the evaluation of (12) would cost only $O(K_{\mathcal{A}}^3 + K_{\mathcal{A}}^2 D)$ for each evaluation of different c. Similarly the matrix computation of (10) can be made more efficient as $\nabla_c f(c) \propto$

$$\text{diag}((Z_{\mathcal{A}}^T X - Z_{\mathcal{A}}^T Z_{\mathcal{A}} W^*)(Z_{\mathcal{A}}^T X - Z_{\mathcal{A}}^T Z_{\mathcal{A}} W^*)^T)$$

can be computed in $O(K^2 D + K^3)$ via the maintenance of $Z_{\mathcal{A}}^T Z_{\mathcal{A}}$, $Z_{\mathcal{A}}^T X$.

The output of Algorithm 1 is the coefficient vector c, and with the resulting latent feature matrix $W(c)$ given by (12). Since the solution could contain many atoms of small weight c_k. In practice, we perform a rounding procedure that ranks atoms according to the score $\{c_k \|W_{k,:}\|^2\}_{k \in \mathcal{A}}$ and then pick top K atoms as the output Z^*, and solve a simple least-squares problem to obtain the corresponding W^*.

4.1. Greedy Atom Generation

A key step to the greedy algorithm (Algorithm 1) is to find the direction (8) of steepest descent, which however is a *convex maximization* problem with *binary constraints* that in general cannot be exactly solved in polynomial time. Fortunately in this section, we show that (8) is equivalent to a MAX-CUT-like *Boolean Quadratic Maximization* problem that has efficient Semidefinite relaxation with constant approximation guarantee. Furthermore, the resulting Semidefinite Programming (SDP) problem is of special structure that allows iterative method of complexity linear to the matrix size (Boumal et al., 2016; Wang & Kolter, 2016).

In particular, let C=$\nabla g(M)$=$A^* A^{*T}/(2\tau N)$ the maximization problem

$$\max_{z \in \{0,1\}^N} \langle C, zz^T \rangle \tag{13}$$

can be reduced to an optimization problem over variables taking values in $\{-1, 1\}$ via the transformation $y = 2z - 1$, which results in the problem

$$\max_{y \in \{-1,1\}^N} \frac{1}{4} \left(\langle C, yy^T \rangle + 2\langle C, \mathbf{1}y^T \rangle + \langle C, \mathbf{1}\mathbf{1}^T \rangle \right). \tag{14}$$

where $\mathbf{1}$ denotes N-dimensional vector of all 1s. By introducing a dummy variable y_0, (14) can be expressed as

$$\max_{(y_0;y) \in \{-1,1\}^{N+1}} \frac{1}{4} \begin{bmatrix} y_0 \\ y \end{bmatrix}^T \begin{bmatrix} \mathbf{1}^T C \mathbf{1} & \mathbf{1}^T C \\ C\mathbf{1} & C \end{bmatrix} \begin{bmatrix} y_0 \\ y \end{bmatrix}. \tag{15}$$

Note that one can ensure finding a solution with $y_0 = 1$ by flipping signs of the solution vector to (15), since this does not change the quadratic form objective value. Denote the quadratic form matrix in (15) be $\hat{C}$. Problem of form (15) is a MAXCUT-like Boolean Quadratic problem, for which there is SDP relaxation of the form

$$\begin{aligned} \max_{Y \in \mathbb{S}^N} \quad & \langle \hat{C}, Y \rangle \\ s.t. \quad & Y \succeq 0, \; diag(Y) = \mathbf{1} \end{aligned} \tag{16}$$

rounding from which guarantees a solution $\hat{y}$ to (15) satisfying

$$\overline{h} - h(\hat{y}) \leq \rho(\overline{h} - \underline{h}) \tag{17}$$

for $\rho = 2/5$ (Nesterov et al., 1997), where $h(y)$ denotes the objective function of (15) and $\overline{h}$, $\underline{h}$ denote the maximum, minimum value achievable by some $y \in \{-1, 1\}^{N+1}$ respectively. Note this result holds for any symmetric matrix $\hat{C}$. Since our problem has a positive-semidefinite matrix $\hat{C}$, $\underline{h} = 0$ and thus

$$-\nabla_{\hat{j}} f(c) = h(\hat{y}) \geq \mu \overline{h} = \mu(-\nabla_{j^*} f(c)) \tag{18}$$

for $\mu = 1 - \rho = 3/5$, where $\hat{j}$ is coordinate selected by rounding from a solution of (16) and j^* is the exact maximizer of (8).

Finally, it is noteworthy that, although solving a general SDP is computationally expensive, SDP of the form (16) has been shown to allow much faster solver that has linear cost w.r.t. the matrix size $nnz(\hat{C})$ (Boumal et al., 2016; Wang & Kolter, 2016). In our implementation we adopt the method of (Wang & Kolter, 2016) due to its strong empirical performance.

5. Analysis

5.1. Convergence Analysis

The aim of this section is to show the convergence of Algorithm 1 under the approximation of greedy atom generation. In particular, we show the multiplicative approximation error incurred in the step (8) only contributes an additive approximation error proportional to λ, as stated in the following theorem.

Theorem 1. *The greedy algorithm proposed (Algorithm 1) satisfies*

$$F(c^t) - F(c^*) \leq \frac{2\gamma\|c^*\|_1^2}{\mu^2}\frac{1}{t} + \underbrace{\frac{2(1-\mu)}{\mu}\lambda\|c^*\|_1}_{\Delta(\lambda)},$$

where c^* is any reference solution, $\mu = 3/5$ is the approximation ratio given by (18) and γ is the Lipschitz-continuous constant of coordinate-wise gradient $\nabla_j f(c),\ \forall j \in [K]$.

The theorem thus shows that the iterates converge sublinearly to within statistical precision λ of any reference solution c^* scaled in main by its ℓ_1 norm $\|c^*\|_1$. In the following theorem, we show that, with the additional assumption that $F(c)$ is strongly convex over a restricted support set $\mathcal{A}^*$, one can get a bound in terms of the ℓ_0**-norm** of a reference solution c^* with support A^*.

Theorem 2. *Let $\mathcal{A}^* \in [\bar{K}]$ be a support set and $c^* := arg\min_{c:supp(c)=\mathcal{A}^*} F(c^*)$. Suppose $F(c)$ is strongly convex on $\mathcal{A}^*$ with parameter β. The solution given by Algorithm 1 satisfies*

$$F(c^T) - F(c^*) \leq \frac{4\gamma\|c^*\|_0}{\beta\mu^2}\left(\frac{1}{T}\right) + \frac{2(1-\mu)\lambda}{\mu}\sqrt{\frac{2\|c^*\|_0}{\beta}}.$$

Let $\bar{K} = 2^N$ be the size of the atomic set. Any target latent structure Z^*W^* can be expressed as $\mathbf{Z}D(c^*)\tilde{W}^*$ where $\mathbf{Z}$ is an $N \times \bar{K}$ dictionary matrix, $D(c^*)$ is a $\bar{K} \times \bar{K}$ diagonal matrix of diagonal elements $D_{kk} = \sqrt{c_k^*}$ with $c_k^* = 1$ for columns corresponding to Z^* and $c_k^* = 0$ for the others, and $\tilde{W}^*$ is W^* padded with 0 on rows in $\{k \mid c_k = 0\}$. Then since $\|c^*\|_1 = \|c^*\|_0 = K^*$, Theorem 2 shows that our algorithm has an iteration complexity of $O(K/\epsilon)$ to achieve ϵ error, with an additional error term proportional to $\lambda\sqrt{K}$ due to the approximation made in (18).

5.2. Risk Analysis

In this section, we investigate the performance of the output from Algorithm 1 in terms of the population risk $r(\cdot)$ defined in (1). Given coefficients c with support $\mathcal{A}$ obtained from algorithm (1) for T iterations, we construct the weight matrix by $\hat{W} = \text{diag}(\sqrt{c_\mathcal{A}})W^*$ with $W^*(c_\mathcal{A}) = \frac{1}{N}Z_\mathcal{A}^T A^*$, where A^* is the maximizer of (4) as a function of c. It can be seen that $\hat{W}$ satisfies

$$F(c) = \frac{1}{2N}\|X - Z_\mathcal{A}\hat{W}\|_F^2 + \frac{\tau}{2}\|\hat{W}\|_F^2 + \lambda\|c_\mathcal{A}\|_1. \quad (19)$$

The following theorem gives a risk bound for $\hat{W}$. Without loss of generality, we assume x is bounded and scaled such that $\|x\| \leq 1$.

Theorem 3. *Let $\hat{W} = diag(\sqrt{c_\mathcal{A}})W^*(c_\mathcal{A})$ be the weight matrix obtained from T iterations of Algorithm 1, and $\bar{W}$ be the minimizer of the population risk (1) with K components and $\|\bar{W}\|_F \leq R$. We then have the following bound on population risk: $r(\hat{W}) \leq r(\bar{W}) + \epsilon$ with probability $1 - \rho$ for*

$$T \geq \frac{4\gamma}{\mu^2\beta}\left(\frac{K}{\epsilon}\right) \text{ and } N = \Omega\left(\frac{DK}{\epsilon^3}\log\left(\frac{RK}{\epsilon\rho}\right)\right),$$

with λ, τ chosen appropriately as functions of N.

Note the output of Algorithm 1 has number of components $\hat{K}$ bounded by number of iterations T. Therefore, Theorem (3) gives us a trade-off between risk and sparsity—one can guarantee to achieve ϵ-suboptimal risk compared to the optimal solution of size K, via $O(K/\epsilon)$ components and $\tilde{O}(DK/\epsilon^3)$ samples. Notice the result (3) is obtained without any distributional assumption on $p(x)$ and $p(z)$ except that of boundedness. Comparatively, the theoretical result obtained from Spectral Method (Tung & Smola, 2014) requires the knowledge/assumption of the distribution $p(x|z)$, $p(z)$, which is sensitive to model misspecification in practice.

5.3. Identifiability

It is noteworthy that the true parameters (Z^*, W^*) might not be identifiable. In particular, it is possible to have $(Z, W) \neq (Z^*, W^*)$ with $ZW = Z^*W^*$, in which case it is impossible to recover the true parameters (Z^*, W^*) from $\Theta^* = Z^*W^*$. The following theorem introduces conditions that ensure uniqueness of the factorization $\Theta^* = Z^*W^*$.

Theorem 4. *Let $\Theta^* = Z^*W^*$ be of rank K. If*

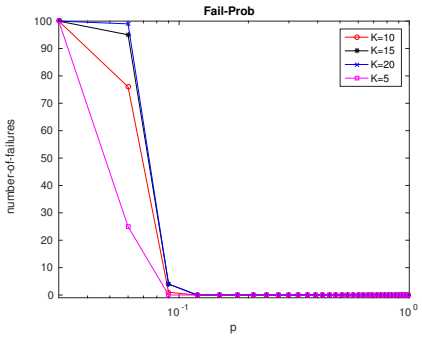

Figure 1: The frequency of failures of condition in Theorem 4 out of 100 trials, for a spectrum of i.i.d. Bernoulli parameter p and different K. We use algorithm proposed in (Slawski et al., 2013) to check the condition efficiently.

1. *Z^*:$N \times K$ and W^*:$K \times D$ are both of rank K.*
2. *$span(Z^*) \cap \{0,1\}^N \setminus \{0\} = \{Z^*_{:,j}\}_{j=1}^K$.*

Then for any rank-K matrices Z:$N \times K$ and W:$K \times D$, $ZW = \Theta^*$ implies $\{Z_{:,j}\}_{j=1}^K = \{Z^*_{:,j}\}_{j=1}^K$ and $\{W_{j,:}\}_{j=1}^K = \{W^*_{j,:}\}_{j=1}^K$.

The conditions in Theorem 4 are similar to that discussed in (Slawski et al., 2013), where an additional affine constraint on W is considered. For random binary matrix of binary value $\{-1,+1\}$ instead of $\{0,1\}$, the conditions are known to hold with high probability when entries are i.i.d. $Bernoulli(0.5)$ (Tao & Vu, 2007; Kahn et al., 1995). Here we also conduct numerical experiments for matrices of i.i.d. $Bernoulli(p)$ with a wide range of p. Results in Figure 1 shows that the probability with which such condition fails is almost 0 when $p \geq 0.1$, while it increases when p becomes smaller than 0.1.

5.4. Parameter Recovery without Noise

Let the true parameters be (Z^*, W^*) with $\|W^*\|_F^2 = R$. We can find some $\tau(R)$ such that the estimator (6) is equivalent to solving the following problem:

$$\begin{aligned} \min_{c \in \mathbb{R}_+^{\bar{K}}, W \in \mathbb{R}^{\bar{K} \times D}} \quad & \frac{1}{2N}\|X - \mathbf{Z}\text{diag}(c)W\|_F^2 + \lambda\|c\|_1 \\ s.t. \quad & \|W\|_F^2 \leq R. \end{aligned} \tag{20}$$

where $\text{diag}(c)$ is a diagonal matrix with $\text{diag}_{kk}(c) = \sqrt{c_k}$. In the noiseless setting ($\epsilon = 0$), one can find a feasible solution to the following problem

$$\begin{aligned} \min_{c \in \mathbb{R}_+^{\bar{K}}, W \in \mathbb{R}^{\bar{K} \times D}} \quad & \|c\|_1 \\ s.t. \quad & \mathbf{Z}D(c)W = X, \ \ \|W\|_F^2 \leq R, \end{aligned} \tag{21}$$

which is equivalent to problem (20) with any $\lambda \leq \bar{\lambda}$ for some $\bar{\lambda} > 0$. One can thus choose an arbitrarily small $\lambda \leq \bar{\lambda}$ and solve (20) to obtain a solution (c, W) of (21), which satisfies the following theorem.

Theorem 5. *Let (c, W) be a solution to (21), and (Z_S, W_S) be columns of $\mathbf{Z}$ and rows of W corresponding to the set of non-zero indexes S of c respectively. Suppose the identifiability condition in Theorem 4 holds and W_S has full row-rank. Then*

$$\{Z_{:,j}\}_{j \in S} = \{Z^*_{:,j}\}_{j=1}^K \ , \ \{W_{j,:}\}_{j \in S} = \{W^*_{j,:}\}_{j=1}^K$$

Note since we can choose an arbitrarily small $\lambda \leq \bar{\lambda}$ to find a solution of (21). The approximation error due to approximate atom generation can be reduced to arbitrarily small.

5.5. Parameter Recovery under Noise

In the noisy setting, parameter recovery is more tricky. When the model is unbiased (i.e. $E[x|z] = (W^*)^T z$) , by appealing to well-known results in high-dimensional estimation (Negahban et al., 2009), we can achieve a bound on the ℓ_2 norm of the error $\hat{c} - c^*$, where c^* is coefficient vector corresponding to the ground-truth parameter (W^*, Z^*).

We defer the resulting Theorem 8 to Section 7.7 in the Appendix. The theorem bounds the ℓ_2 error $\|\hat{c} - c^*\|_2$ as $(1/\kappa_n)\sqrt{K^*}\,\rho_n$, where ρ_n is a term capturing the noise-level, κ_n is a term capturing the restricted strong convexity of the objective, and defined in detail in Section 7.7.

However, extending this bound on $\|c - c^*\|$ to derive bounds on $\|Z - Z^*\|$ and $\|W - W^*\|$ is a delicate matter that we defer to future work, in part due to the exponential size $\bar{K} = 2^N$ of the atomic set, and since the size of Z grows with N. In particular, in the following theorem, we show that it is in general not possible to estimate Z^* accurately even with a large number of samples.

Theorem 6. *Let $K = D = 1$. Consider the following noise model: $X = Z^*W^* + E$, where $Z^* \in \{0,1\}^N$, $W^* \in \mathbb{R}$ and $\forall i \in [N]$, E_i are i.i.d random variables with $E_i \sim \mathcal{N}(0,1)$. Moreover, suppose we know the true parameter $W^* = 1$. Then the Latent Feature Model estimator for Z^* given by:*

$$\hat{Z} = \underset{Z \in \{0,1\}^N}{argmin} \quad \frac{1}{2N}\|X - ZW^*\|^2 \tag{22}$$

satisfies the following:

$$\mathbb{E}(\|\hat{Z} - Z^*\|_2^2) \geq cN,$$

for some positive constant c.

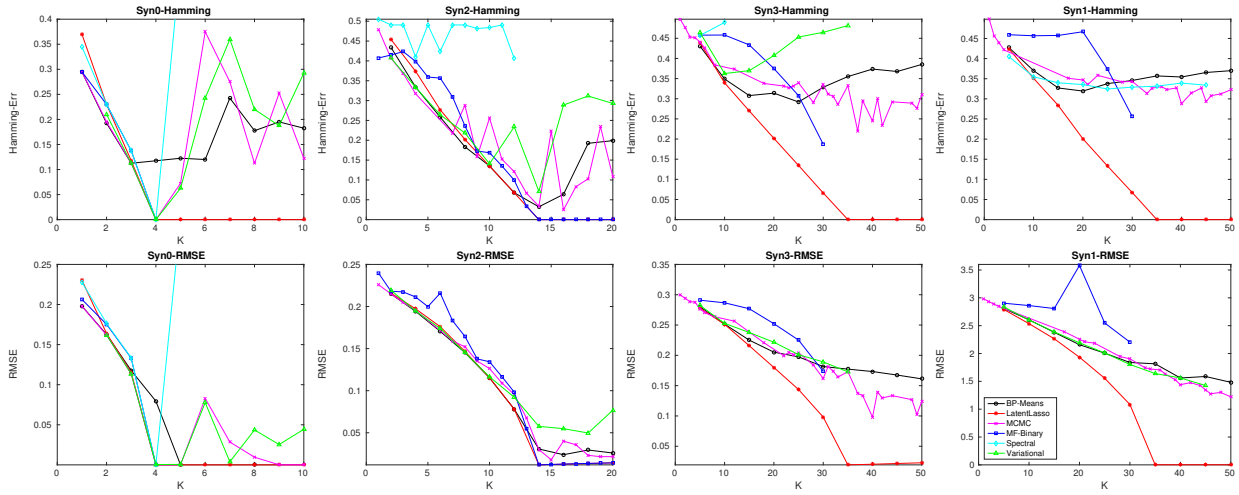

Figure 2: From left to right, each column are results for Syn0 (K=4), Syn2 (K=14), Syn3 (K=35) and Syn1 (K=35) respectively. The first row shows the Hamming loss between the ground-truth binary assignment matrix Z^* and the recovered ones $\hat{Z}$. The second row shows RMSE between $\Theta^* = Z^*W^*$ and the estimated $\hat{\Theta} = \hat{Z}\hat{W}$.

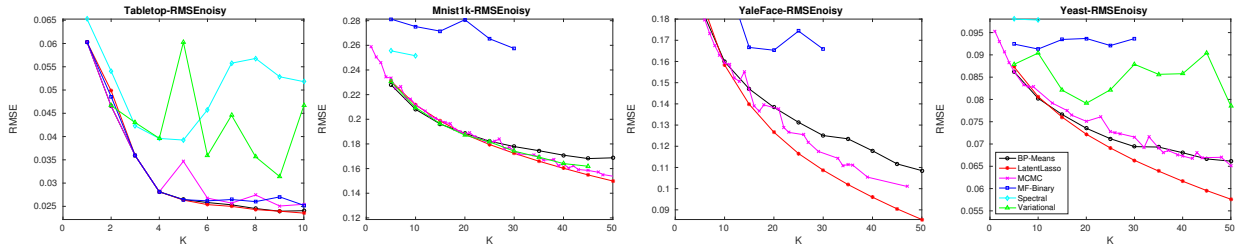

Figure 3: From left to right are results for Tabletop, Mnist1k, YaleFace and Yeast, where Spectral Method does not appear in the plots for YaleFace and Yeast due to a much higher RMSE, and Variational method reports a runtime error when running on the YaleFace data set.

Table 1: Data statistics.

Dataset	N	D	K	σ	$nnz(W^*_{k,:})$
Syn0	100	196	4	0	≤ 8
Syn1	1000	1000	35	0.01	1000
Syn2	1000	900	14	0.1	49
Syn3	1000	900	35	0.1	36
Tabletop	100	8560	4	n/a	n/a
Mnist1k	1000	777	n/a	n/a	n/a
YaleFace	165	2842	n/a	n/a	n/a
Yeast	1500	104	n/a	n/a	n/a

6. Experiments

In this section, we compare our proposed method with other state-of-the-art approaches on both synthetic and real data sets. The dataset statistics are listed in Table 1. For the synthetic data experiments, we used a benchmark simulated dataset *Syn0* that was also used in (Broderick et al., 2013; Tung & Smola, 2014). But since this has only a small number of latent features ($K = 4$), to make the task more challenging, we created additional synthetic datasets (which we denote Syn1, Syn2, Syn3) with more latent features. Figure 4 shows example of our synthetic data, where we reshape dimension D into an image and pick a contiguous region. Each pixel $W(k, j)$ in the region is set as $N(0, \sigma^2)$, while pixels not in the region are set to 0. In the examples of Figure 4, the region has size $nnz(W(k, :))$=36. Note the problem becomes harder when the region size $nnz(W(k, :))$, number of features K, or noise level σ becomes larger. For real data, we use a benchmark *Tabletop* data set constructed by (Griffiths & Ghahramani, 2005), where there is a ground-truth number of features $K = 4$ for the 4 objects on the table. We also take two standard multi-label (multiclass) classification data sets *Yeast* and *Mnist1k* from the LIBSVM repository [1], and one Face data set *YaleFace* from the Yale Face database [2].

Given the estimated factorization (Z, W), we use the fol-

[1] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

[2] http://vision.ucsd.edu/content/yale-face-database

lowing 3 evaluation metrics to compare different algorithms:

- Hamming-Error: $\min_{\mathcal{S}:|\mathcal{S}|=K} \frac{\|Z_{:,\mathcal{S}}-Z^*\|_F^2}{NK}$.
- RMSE: $\frac{\|Z^*W^*-ZW\|_F}{\sqrt{ND}}$.
- RMSEnoisy: $\frac{\|X-ZW\|_F}{\sqrt{ND}}$.

where the first two can only be applied when the ground truth Z^* are W^* are given. For real data, we can only evaluate the noisy version of RMSE, which can be interpreted as trying to find a best approximation to the observation X via a factorization with binary components.

The methods in comparison are listed as follows: **(a) MCMC:** An accelerated version of the Collapsed Gibbs sampler for the Indian Buffet Process (IBP) model (Doshi-Velez & Ghahramani, 2009). We adopted the implementation published by [3]. We ran it with 25 random restarts and recorded the best results for each K. **(b) Variational:** A Variational approximate inference method for IBP proposed in (Doshi-Velez et al., 2009). We used implementation published by the author [4]. **(c) MF-Binary:** A Matrix Factorization with the Binary Components model (Slawski et al., 2013), which has recovery guarantees in the noiseless case but has a $O(K2^K)$ complexity and thus cannot scale to $K > 30$ on our machine. We use the implementation published by the author [5]. **(d) BP-Means:** A local search method that optimizes a MAD-Bayes Latent Feature objective function (Broderick et al., 2013). We used code provided by the author [6]. We ran it with 100 random restarts and recorded the best result. **(e) Spectral:** Spectral Method for IBP Linear Gaussian model proposed in (Tung & Smola, 2014). We used code from the author. The implementation has a memory requirement that restricts its use to $K < 14$. **(f) LatentLasso:** The proposed Latent Feature Lasso method (Algorithm 1).

The results are shown in Figure 2 and 3. On synthetic data, we observe that, when the number of features K is small (e.g. Syn0), most of methods perform reasonably well. However, when the number of features becomes slightly larger (i.e. $K = 35$ in Syn1, Syn3), most of algorithms lose their ability of recovering the hidden structure, and when they fail to do so, they can hardly find a good approximation to $\Theta^* = Z^*W^*$ even using a much larger number of components up to 50. We found the proposed *LatentLasso* method turns out to be the only method that can still recover the desired hidden structure on the Syn1 and Syn3 data sets, which gives 0 RMSE and Hamming Error. On Syn2 ($K = 14$) data set, *MF-Binary* and *LatentLasso* are the only two methods that achieve 0 RMSE and Hamming-Error. However, MF-Binary has a complexity growing exponential with K, which results in its failure on Syn1 and Syn3 due to a running time more than one day when $K > 30$. The proposed LatentLasso algorithm actually runs significantly faster than other methods in our experiments. For example, on the Syn1 dataset (N=1000, D=1000, K=35), the runtime of LatentLasso is 398s, while MCMC, Variational, MF-Binary and BP-Means all take more than 10000s to obtain their best results reported in the Figures. We provide a comparison of the time complexities of all compared methods in Section 7.1 in the Appendix. Our overall lower time complexity is also corroborated empirically by our experiments. We also observe that *LatentLasso* is the only method that has RMSE and Hamming error monotonically decreasing with K. On Syn0 and Tabletop which have ground-truth $K = 4$, we found most of algorithms could become unstable when trying to use a number of components K larger than the ground truth. Among all algorithms, *Spectral*, *Variational* methods are the most unstable, while *BP-Means* and *MCMC* are more stable possibly due to the large number of random re-trials employed in their procedures.

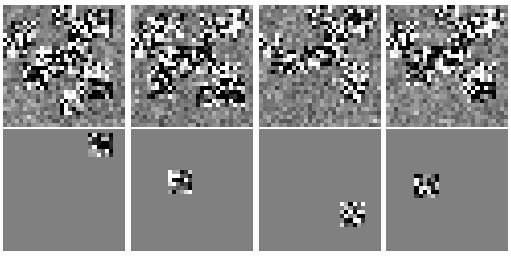

Figure 4: Sample images from the synthetic data we created (i.e. Syn1, Syn2, Syn3). The first row shows observations $X_{i,:}$, and the second row shows latent features $W_{k,:}$.

On real data sets, the LFM model assumption might not hold, and might serve at best as an approximation to the ground-truth. Even in such cases, we found that our *LatentLasso* method finds a better approximation than other existing approaches, especially when using a larger number of components K. We conjecture that for local search methods, the performance breakdown for larger K is possibly due to an exponentially increased number of local optimums, which makes strategies such as random restarts less effective for methods such as BP-Means and MCMC. On the other hand, the Spectral Method simply has a sample complexity bound with a high-order polynomial dependency on K, which makes the estimation error increase dramatically as K becomes larger.

[3] https://github.com/davidandrzej/PyIBP
[4] http://mloss.org/software/view/185/
[5] https://sites.google.com/site/slawskimartin/code
[6] https://github.com/tbroderick/bp-means

Acknowledgements P.R. acknowledges the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1447574, and DMS-1264033, and NIH via R01 GM117594-01. S.D. Lin acknowledges the support of AOARD and MOST via 104-2628-E-002-015-MY3.

References

Airoldi, Edoardo M, Blei, David, Erosheva, Elena A, and Fienberg, Stephen E. *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC, 2014.

Boumal, Nicolas, Voroninski, Vlad, and Bandeira, Afonso. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pp. 2757–2765, 2016.

Broderick, Tamara, Kulis, Brian, and Jordan, Michael I. Mad-bayes: Map-based asymptotic derivations from bayes. In *ICML (3)*, pp. 226–234, 2013.

Chandrasekaran, Venkat, Recht, Benjamin, Parrilo, Pablo A, and Willsky, Alan S. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

d'Aspremont, Alexandre, El Ghaoui, Laurent, Jordan, Michael I, and Lanckriet, Gert RG. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

Doshi-Velez, Finale and Ghahramani, Zoubin. Accelerated sampling for the indian buffet process. In *Proceedings of the 26th annual international conference on machine learning*, pp. 273–280. ACM, 2009.

Doshi-Velez, Finale, Miller, Kurt T, Van Gael, Jurgen, Teh, Yee Whye, and Unit, Gatsby. Variational inference for the indian buffet process. In *Proceedings of the Intl. Conf. on Artificial Intelligence and Statistics*, volume 12, pp. 137–144, 2009.

Griffiths, Thomas L and Ghahramani, Zoubin. Infinite latent feature models and the indian buffet process. In *NIPS*, volume 18, pp. 475–482, 2005.

Griffiths, Thomas L and Ghahramani, Zoubin. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(Apr):1185–1224, 2011.

Hayashi, Kohei and Fujimaki, Ryohei. Factorized asymptotic bayesian inference for latent feature models. In *Advances in Neural Information Processing Systems*, pp. 1214–1222, 2013.

Jolliffe, Ian T, Trendafilov, Nickolay T, and Uddin, Mudassir. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003.

Kahn, Jeff, Komlós, János, and Szemerédi, Endre. On the probability that a random$\pm$1-matrix is singular. *Journal of the American Mathematical Society*, 8(1):223–240, 1995.

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

Nesterov, Yurii et al. *Quality of semidefinite relaxation for nonconvex quadratic optimization*. Université Catholique de Louvain. Center for Operations Research and Econometrics [CORE], 1997.

Shaban, Amirreza, Farajtabar, Mehrdad, Xie, Bo, Song, Le, and Boots, Byron. Learning latent variable models by improving spectral solutions with exterior point method. In *UAI*, pp. 792–801, 2015.

Slawski, Martin, Hein, Matthias, and Lutsik, Pavlo. Matrix factorization with binary components. In *Advances in Neural Information Processing Systems*, pp. 3210–3218, 2013.

Tao, Terence and Vu, Van. On the singularity probability of random bernoulli matrices. *Journal of the American Mathematical Society*, 20(3):603–628, 2007.

Tung, Hsiao-Yu and Smola, Alexander J. Spectral methods for indian buffet process inference. In *Advances in Neural Information Processing Systems*, pp. 1484–1492, 2014.

Ueda, Naonori and Saito, Kazumi. Parametric mixture models for multi-labeled text. *Advances in neural information processing systems*, pp. 737–744, 2003.

Wang, Po-Wei and Kolter, J Zico. The mixing method for maxcut-sdp problem. *NIPS LHDS Workshop.*, 2016.

Wu, Lingfei, Yen, Ian EH, Chen, Jie, and Yan, Rui. Revisiting random binning features: Fast convergence and strong parallelizability. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1265–1274. ACM, 2016.

Zhao, Han and Poupart, Pascal. A sober look at spectral learning. *arXiv preprint arXiv:1406.4631*, 2014.

Zoubin, Ghahramani. Scaling the indian buffet process via submodular maximization. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1013–1021, 2013.

7. Appendix

7.1. Comparison on Time Complexity

The proposed LatentLasso algorithm runs significantly faster than other methods in our experiments. For example, on the Syn1 dataset (N=1000, D=1000, K=35), the runtime of LatentLasso is 398s, while MCMC, Variational, MF-Binary and BP-Means all take more than 10000s to obtain their best results reported in the Figures (and the implementation of Spectral Method we obtained from the authors has memory requirement that restricts $K < 14$). On the real data sets, we report only up to K=50 because most of the compared methods already took one day to train.

Table 2: Comparison of Time Complexity. (T denotes number of iterations)

MCMC	Variational	MF-Binary
$(NK^2D)T$	$(NK^2D)T$	$(NK)2^K$
BP-Means	Spectral	LatentLasso
$(NK^3D)T$	$ND + K^5log(K)$	$(ND + K^2D)T$

The complexity of each algorithm can be summarized in Table 2. The reason for the smaller runtime of LatentLasso is due to the decoupling of factor ND from the factor related to K, where the factor $O(ND)$ comes from the cost of solving a MAX-CUT-like problem using the method of (Boumal et al., 2016) or (Wang & Kolter, 2016), while the factor $O(K^2D)$ comes from the cost of solving a least-square problem given by (11) with the maintenance cost of Z^TZ amortized.

7.2. Proof for Theorem 1

Let $L(M)$ be a smooth function such that $\nabla L(M)$ is Lipschitz-continuous with parameter β, that is,

$$L(M') - L(M) - \langle \nabla L(M), M' - M \rangle \leq \frac{\beta}{2}\|M' - M\|_F^2.$$

Then

$$\nabla_j f(c) = z_j^T \nabla L(M) z_j$$

is Lipschitz-continuous with parameter γ, which is of order $O(1)$ when loss function $L(.)$ is an empirical average normalized by ND.

Let $\mathcal{A}$ be the active set before adding $\hat{j}$. Consider the descent amount produced by minimizing $F(c)$ w.r.t. the $c_{\hat{j}}$ given that $0 \in \partial_j F(c)$ for all $j \in \mathcal{A}$ due to the subproblem solved in the previous iteration. Let $j = \hat{j}$, for any η_j we have

$$\begin{aligned} F(c + \eta_j e_j) - F(c) &\leq \nabla_j f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \\ &\leq \mu \nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \end{aligned}$$

Minimize w.r.t η_j gives

$$\begin{aligned} &\min_{\eta_j} \; F(c + \eta_j e_j) - F(c) \\ &\leq \min_{\eta_j} \; \mu \nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \\ &= \min_{\eta_k : k \notin \mathcal{A}} \; \sum_{k \notin \mathcal{A}} \left(\mu \nabla_k f(c)\eta_k + \lambda|\eta_k| \right) + \frac{\gamma}{2}\left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 \\ &\leq \min_{\eta_k : k \notin \mathcal{A}} \; \mu \sum_{k \notin \mathcal{A}} \left(\nabla_k f(c)\eta_k + \lambda|\eta_k| \right) + \frac{\gamma}{2}\left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 \\ &+ (1-\mu)\lambda \sum_{k \notin \mathcal{A}} |\eta_k| \end{aligned}$$

where the last equality is justified later in Lemma 1. For $k \in \mathcal{A}$, we have

$$0 = \min_{\eta_k : k \in \mathcal{A}} \; \mu \sum_{k \in \mathcal{A}} \left(\nabla_k f(c)\eta_k + \lambda|c_k + \eta_k| - \lambda|c_k| \right)$$

Combining cases for $k \notin \mathcal{A}$ and $k \in \mathcal{A}$, we can obtain a global estimate of descent amount compared to some optimal solution x^* as follows

$$\begin{aligned} &\min_{\eta_{\hat{j}}} \; F(c + \eta_{\hat{j}} e_{\hat{j}}) - F(c) \\ &\leq \min_{\eta} \; \mu \Big(\langle \nabla f(c), \eta \rangle + \lambda\|c + \eta\|_1 - \lambda\|c\|_1 \Big) \\ &+ \frac{\gamma}{2}\left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 + (1-\mu)\lambda \sum_{k \notin \mathcal{A}} |\eta_k| \\ &\leq \min_{\eta} \; \mu \Big(F(c + \eta) - F(c) \Big) + \frac{\gamma}{2}\left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 \\ &+ (1-\mu)\lambda \sum_{k \notin \mathcal{A}} |\eta_k| \\ &\leq \min_{\alpha \in [0,1]} \; \mu \Big(F(c + \alpha(c^* - c)) - F(c) \Big) + \frac{\alpha\gamma}{2}\|c^*\|_1^2 \\ &+ \alpha(1-\mu)\lambda\|c^*\|_1 \\ &\leq \min_{\alpha \in [0,1]} \; -\alpha\mu \Big(F(c) - F(c^*) \Big) + \frac{\alpha^2\gamma}{2}\|c^*\|_1^2 \\ &+ \alpha(1-\mu)\lambda\|c^*\|_1. \end{aligned}$$

It means we can always choose an α small enough to guarantee descent if

$$F(c) - F(c^*) > \frac{(1-\mu)}{\mu}\lambda\|c^*\|_1. \tag{23}$$

In addition, for

$$F(c) - F(c^*) \geq \frac{2(1-\mu)}{\mu}\lambda\|c^*\|_1, \tag{24}$$

we have

$$\min_{\eta_{\hat{j}}} F(c+\eta_{\hat{j}}e_{\hat{j}}) - F(c)$$
$$\leq \min_{\alpha\in[0,1]} -\frac{\alpha\mu}{2}\Big(F(c)-F(c^*)\Big) + \frac{\alpha^2\gamma}{2}\|c^*\|_1^2.$$

Minimizing w.r.t. to α gives the convergence guarantee

$$F(c^t) - F(c^*) \leq \frac{2\gamma\|c^*\|_1^2}{\mu^2}\frac{1}{t}.$$

for any iterate with $F(c^t) - F(c^*) \geq \frac{2(1-\mu)}{\mu}\lambda\|c^*\|_1$.

Lemma 1.

$$\min_{\eta_j} \mu\nabla_{j^*}f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \tag{25}$$
$$= \min_{\eta_k:k\notin\mathcal{A}} \sum_{k\notin\mathcal{A}}\Big(\mu\nabla_k f(c)\eta_k + \lambda|\eta_k|\Big) + \frac{\gamma}{2}\left(\sum_{k\notin\mathcal{A}}|\eta_k|\right)^2 \tag{26}$$

Proof. The minimization (34) is equivalent to

$$\min_{\eta_k:k\notin\mathcal{A}} \quad \sum_{k\notin\mathcal{A}}\Big(\mu\nabla_k f(c)\eta_k\Big)$$
$$s.t. \quad \left(\sum_{k\notin\mathcal{A}}|\eta_k|\right)^2 \leq C_1$$
$$\sum_{k\notin\mathcal{A}}|\eta_k| \leq C_2$$

and therefore is equivalent to

$$\min_{\eta_k:k\notin\mathcal{A}} \quad \mu\sum_{k\notin\mathcal{A}}\nabla_k f(c)\eta_k$$
$$s.t. \quad \sum_{k\notin\mathcal{A}}|\eta_k| \leq \min\{\sqrt{C_1}, C_2\}$$

which is a linear objective subject to a convex set and thus always has solution that lies on the corner point with only one non-zero coordinate η_{j^*}, which then gives the same minimum as (33). □

7.3. Proof of Theorem 2

Lemma 2. *Let $\mathcal{A}^* \in [\bar{K}]$ be a support set and $c^* := arg\min_{c:supp(c)=\mathcal{A}^*} F(c^*)$. Suppose $F(c)$ is strongly convex on $\mathcal{A}^*$ with parameter β. We have*

$$\|c^*\|_1 \leq \sqrt{\frac{2\|c^*\|_0(F(0)-F(c^*))}{\beta}}. \tag{27}$$

Proof. Since $supp(c^*) = \mathcal{A}^*$, and c^* is optimal when restricted on the support, we have $\langle, c^*\rangle = 0$ for some $\in \partial F(c^*)$. And since $F(c)$ is strongly convex on the support $\mathcal{A}^*$ with parameter β, we have

$$F(0) - F(c^*) = F(0) - F(c^*) - \langle, 0 - c^*\rangle$$
$$\geq \frac{\beta}{2}\|c^* - 0\|_2^2,$$

which gives us

$$\|c^*\|_2^2 \leq \frac{2(F(0)-F(c^*))}{\beta}.$$

Combining above with the fact for any c, $\|c\|_1^2 \leq \|c\|_0\|c\|_2^2$, we obtain the result. □

Since $F(0) - F(c^*) \leq \frac{1}{2N}\sum_{i=1}^N y_i^2 \leq 1$, from Theorem (1) and (27), we have

$$F(c^T) - F(c^*) \leq \frac{4\gamma\|c^*\|_0}{\beta\mu^2}\left(\frac{1}{T}\right) + \frac{2(1-\mu)\lambda}{\mu}\sqrt{\frac{2\|c^*\|_0}{\beta}}. \tag{28}$$

for any $c^* := arg\min_{c:supp(c)=\mathcal{A}^*} F(c)$.

7.4. Proof of Theorem 3

Before delving into the analysis of the *Latent Feature Lasso* method, we first investigate what one can achieve in terms of the risk defined in (1) if the *combinatorial version of objective* is solved. Let

$$f(x;W) := \min_{z\in\{0,1\}^K} \frac{1}{2}\|x - W^T z\|^2.$$

Suppose we can obtain solution $\hat{W}$ to the following empirical risk minimization problem:

$$\hat{W} := \underset{W\in\mathbb{R}^{K\times D}:\|W\|_F\leq R}{argmin} \quad \frac{1}{N}\sum_{i=1}^N f(x_i;W). \tag{29}$$

Then the following theorem holds.

Theorem 7. *Let W^* be the minimizer of risk (1) and $\hat{W}$ be the empirical risk minimizer (29). Then*

$$E[f(x;\hat{W})] - E[f(x;W^*)]$$
$$\leq \frac{3}{N} + \sqrt{\frac{DK\log(4R^2KN)}{2N} + \frac{1}{2N}\log(\frac{1}{\rho})}$$

with probability $1-\rho$.

Proof Sketch. Let $E_N[f(x,W)]$ denote the empirical risk. We have

$$E[f(x;\hat{W})] - E[f(x;W^*)]$$
$$\leq 2\left(\sup_{W\in\mathbb{R}^{K\times D}:\|W\|_F\leq R}|E[f(x;W)] - E_N[f(x;W)]|\right) \tag{30}$$

from error decomposition and $E_N[f(x,\hat{W})] \leq E_N[f(x,W^*)]$. Then by introducing a δ-net $\mathcal{N}(\delta)$ with covering number $|\mathcal{N}(\delta)| = \left(\frac{4R}{\delta}\right)^{DK}$, we have $\|\tilde{W} - W\|_F \leq \delta$ for some $\tilde{W} \in \mathcal{N}(\delta)$ and

$$P\left(\sup_{\tilde{W}\in\mathcal{N}(\delta)}\left|E[f(x;\tilde{W})] - E_N[f(x;\tilde{W})]\right| \leq \epsilon\right) \geq 1 - \left(\frac{4R}{\delta}\right)^{DK}\exp(-2N\epsilon^2). \tag{31}$$

Then since

$$\begin{aligned}
&2(f(x,\tilde{W}) - f(x,W)) \leq \|x - \tilde{W}^T z^*\|^2 - \|x - W^T z^*\|^2 \\
&= z^{*T}(W - \tilde{W})x + \langle \tilde{W}\tilde{W}^T - WW^T, z^* z^{*T}\rangle \\
&\leq \|z^*\|_2\|W - \tilde{W}\|_F + 2R\|\tilde{W} - W\|_F\|z^*\|_2^2 \\
&\leq 3RK\|\tilde{W} - W\|_F,
\end{aligned}$$

we have

$$\begin{aligned}
&\sup_{W:\|W\|_F\leq R}\left|E[f(x;W)] - E_N[f(x;W)]\right| \\
&\leq (3RK\delta) + \sup_{\tilde{W}\in\mathcal{N}(\delta)}\left|E[f(x;\tilde{W})] - E_N[f(x;\tilde{W})]\right| \\
&\leq 3RK\delta + \sqrt{\frac{DK}{2N}\log(\frac{4R}{\delta}) + \frac{1}{2N}\log(\frac{1}{\rho})}
\end{aligned} \tag{32}$$

with probability $1-\rho$. Choosing $\delta = 1/(RKN)$ yields the result. □

Now we establish the proof of Theorem (3) for bounding risk of the *Latent Feature Lasso* estimator.

Proof. Let $Z^* \in arg\min_{Z\in\{0,1\}^{NK}} \frac{1}{N}\|X - ZW^*\|_F^2$ and $\mathcal{S}^*$ be the set of column index of $\mathbf{Z}$ with the same 0-1 patterns to columns in Z^*. Let c^* be indicator vector with $c_k^* = 1, k \in \mathcal{S}^*$ and $c_k^* = 0, k \notin \mathcal{S}^*$. We have

$$F(\bar{c}) \leq F(c^*) \leq E_N[f(x;W^*)] + \frac{\tau}{2}\|W^*\|_F^2 + \lambda\|c^*\|_1 \tag{33}$$

where $\bar{c} \in \underset{c:supp(c)=\mathcal{S}^*}{argmin} F(c)$. Then let (c,W) with $supp(c) = \hat{\mathcal{S}}$ be the output obtained from running T iterations of the greedy algorithm, we have

$$\begin{aligned}
&E_N[f(x,D_cW)] + \frac{\tau}{2}\|W\|_F^2 + \lambda\|c\|_1 \\
&= \frac{1}{2N}\sum_{i=1}^N \min_{z\in\{0,1\}^{\|c\|_0}} \|x_i - W^T D_c^T z\|^2 + \frac{\tau}{2}\|W\|_F^2 + \lambda\|c\|_1 \\
&\leq F(c)
\end{aligned} \tag{34}$$

Combining (33), (34) and (28), we obtain a bound on the *bias* and *optimization error* of the Latent Feature Lasso estimator

$$\begin{aligned}
&E_N[f(x,D_cW)] \leq F(c) \leq E_N[f(x;W^*)] \\
&+ \underbrace{\frac{\tau}{2}\|W^*\|^2 + \lambda K}_{\text{regularize bias}} + \underbrace{\frac{2\gamma K}{\beta}\left(\frac{1}{T}\right) + \sqrt{\frac{2(1-\mu)K}{\mu\beta}}\lambda}_{\text{optimization error}}
\end{aligned} \tag{35}$$

To bound the estimation error, notice that the matrix $\hat{W} := D_cW$ is $\hat{K} \times D$ with $\hat{K} \leq T$. Furthermore, the descent condition $F(c) \leq F(0)$ guarantees that

$$\frac{\tau}{2}\|W\|_F^2 + \lambda\|c\|_1 \leq \frac{1}{N}\|X - 0\|^2 \leq 1$$

and thus $\|W\|_F^2 \leq 1/\tau$, $\|c\|_1 \leq 1/\lambda$.

Let $\mathcal{W}(T,\lambda,\tau) := \{\hat{W} \in (\mathbb{R}^{T\times D}) \mid \|\hat{W}\|_F \leq \sqrt{1/(\lambda\tau)}\}$. We have

$$\begin{aligned}
&\sup_{(c,W)\in\mathcal{W}(T,\lambda,\tau)} E[f(x;\hat{W})] - E_N[f(x,\hat{W})] \\
&\leq \sqrt{\frac{DT\log(4TN/(\tau\lambda))}{2N} + \frac{1}{2N}\log(\frac{1}{\rho})}
\end{aligned}$$

with probability $1-\rho$ through the same argument as in the case of combinatorial objective (32). Combining the above *estimation error* with the *bias* and *optimization error* in (35), we have

$$\begin{aligned}
&E[f(x;W)] - E[f(x;W^*)] \\
&\leq \frac{\tau}{2}R^2 + \lambda K + \frac{2\gamma K}{\beta T} + \sqrt{\frac{2(1-\mu)K}{\mu\beta}}\lambda \\
&+ \sqrt{\frac{DT\log(4TN/(\tau\lambda))}{2N} + \frac{1}{2N}\log(\frac{1}{\rho})}
\end{aligned}$$

Choosing $T = \frac{2\gamma K}{\beta}(\frac{1}{\epsilon})$, $\lambda = \tau = \frac{1}{\sqrt{N}}$ and $N \gtrsim \frac{DT}{\epsilon^2} = \frac{DK}{\epsilon^3}$ gives the result. □

7.5. Proof of Theorem 4

Proof. Since W^* is of rank K, we have $span(\Theta^*) = span(Z^*)$. Therefore, from condition 2,

$$span(\Theta^*) \cap \{0,1\}^N \setminus \{0\} = \{Z^*_{:,j}\}_{j=1}^K. \tag{36}$$

For any $(Z,W) : ZW = \Theta^*$, we have $Z \in span(\Theta^*)$ since $Z = \Theta^* V\Sigma^{-1}U^T$ where $U\Sigma V^T$ is the SVD of W with $\Sigma : K\times K$. Then by (36) we know that $Z = Z^*$. Then it follows $W = W^*$ since the linear system $\Theta^* = Z^*W$ has unique solution for W. □

7.6. Proof of Theorem 5

Proof. The solution of (21) satisfies

$$Z_S W_S = X = Z^* W^*.$$

Since W_S has full row-rank, we have $rank(Z_S) = rank(X) = rank(Z) = K$ by condition 1 in Theorem 4. Then let $W_S = U\Sigma V^T$ be the SVD of W_S with $\Sigma : |S| \times |S|$, we have

$$Z_S = XV\Sigma^{-1}U^T = Z^*W^*V\Sigma^{-1}U^T \in span(Z^*).$$

Then by condition 2 in Theorem 4, the columns of Z_S can only be in $\{Z^*_{:,j}\}_{j=1}^K$, which implies Z_S equal to Z^* up to a permutation. Then we know $|S| = K$ and by Theorem 4 W_S also equals W^* up to a permutation. □

7.7. ℓ_2 error bounds on the coefficient vector $\hat{c}$

Theorem 8. *Let c^* be the true underlying vector, with support S and sparsity K^*. Let $\hat{c}$ be the minimizer of $F(c)$, defined in Equation (7). Define the* noise-level *term*

$$\rho_n := \max_{z\in\{0,1\}^N} \frac{1}{2N^2\tau}\|A^{*T}z\|_2^2,$$

where $A^* = (I-P)\epsilon + (I-P)(Z_S W^*)$ where

$$P = Z_S(Z_S^T Z_S + N\tau I)^{-1}Z_S^T.$$

Let κ_n be the restricted strong convexity *term defined as :*

$$\kappa_n := \inf_{\Delta\in\mathcal{C}} \left\{ f(c^*+\Delta) - f(c^*) - \langle \nabla f(c^*), \Delta\rangle \right\},$$

where $\mathcal{C} = \{c | \|c_{S^c}\|_1 \le 3\|c_S\|_1\}$. Then, if the regularization parameter is set as $\lambda \ge \rho_n$, we have the following bound on the norm of the error $\hat{c} - c^*$:

$$\|\hat{c} - c^*\|_2 \le \frac{\rho_n\sqrt{K^*}}{\kappa_n}.$$

Proof. By an application of Theorem1 of (Negahban et al., 2009), for $\lambda \ge \|\nabla f(c^*)\|_\infty$, we have the following bound on the ℓ_2 norm of $\hat{c} - c^*$:

$$\|\hat{c} - c^*\|_2 \le \frac{\sqrt{K^*}\lambda}{\kappa_n},$$

where $\|\nabla f(c^*)\|_\infty$ is given by:

$$\|\nabla f(c^*)\|_\infty = \max_{z\in\{0,1\}^N} \frac{1}{2N^2\tau}\|A^{*T}z\|_2^2,$$

where A^* is defined as:

$$\begin{aligned} A^* &= (I - Z_S(Z_S^T Z_S + N\tau I)^{-1}Z_S^T)X \\ &= (I - Z_S(Z_S^T Z_S + N\tau I)^{-1}Z_S^T)(Z_S W^* + \epsilon) \end{aligned} \tag{37}$$

Given $P = Z_S(Z_S^T Z_S + N\tau I)^{-1}Z_S^T$, it can be seen that A^* can be rewritten as :

$$A^* = (I-P)\epsilon + (I-P)(Z_S W^*).$$

□

7.8. Proof of Theorem 6

Proof. Note that the optimization problem in Equation (22) can be rewritten as:

$$\begin{aligned} &\underset{Z\in\{0,1\}^N}{argmin}\ \tfrac{1}{2N}\|E + (Z^* - Z)\|_2^2 \\ &= \tfrac{1}{2N}\sum_{i=1}^N \underset{Z_i\in\{0,1\}}{argmin}\ (E_i + (Z_i^* - Z_i))^2 \end{aligned} \tag{38}$$

So, we have the following closed form expression for $\hat{Z}$:

$$\hat{Z}_i = \begin{cases} 1 & \text{if } Z_i^* + E_i \ge 0.5 \\ 0 & \text{o.w} \end{cases}.$$

We now compute the probability that $Z_i^* \ne \hat{Z}_i$:

$$\begin{aligned} \mathbb{P}(Z_i^* \ne \hat{Z}_i) &= \mathbb{P}(E_i \ge 0.5) * \mathbb{P}(Z_i^* = 0) \\ &\quad + \mathbb{P}(E_i \le -0.5) * \mathbb{P}(Z_i^* = 1) \\ &\ge \min\{\mathbb{P}(E_i \ge 0.5), \mathbb{P}(E_i \le -0.5)\} \ge c, \end{aligned} \tag{39}$$

for some positive constant c. We now use the fact that $\mathbb{E}((Z_i^* - \hat{Z}_i)^2) = \mathbb{P}(Z_i^* \ne \hat{Z}_i)$ to complete the proof of the Lemma. □