l_1 Regularized Logistic Regression (for Ising Model)

Reviewer: Abhirup Mallik

Ising Model

- Undirected graph $G = (V, E)$. $V = \{1, 2, \cdots, p\}$.
- $X = (X_1, \cdots, X_p)$, where X_s corresponds to vertex $s \in V$.
- $X_s \in \{-1, 1\}$ for each $s \in V$. $\phi_{st}(x_s, x_t) = \theta^*_{st} x_s x_t$
- $P_{\theta^*}(x) = \frac{1}{Z(\theta^*)} \exp(\sum_{(s,t)\in E} \theta^*_{st} x_s x_t)$
- θ^* is $\binom{p}{2}$ dimensional vector.
- $Z(\theta^*)$ is normalizing factor.
- Edge sign vector: $E^* := sign(\theta^*_{st})$ if $(s, t) \in E$, 0 ow.

l_1 Regularized Logistic Regression

- $\mathcal{X}_1^n = \{x^{(1)}, x^{(2)}, \cdots, x^{(n)}\}$, iid samples.
- $x^{(i)} \in \{-1, 1\}^p$
- $\mathcal{P}_{\theta^*}(x_r^{(i)} | x_{\setminus r}^{(i)}) = \frac{\exp(2x_r \sum_{t \in V \setminus r} \theta_{rt}^* x_t)}{\exp(2x_r \sum_{t \in V \setminus r} \theta_{rt}^* x_t) + 1}$
- $\min_{\theta_{\setminus r} \in \mathbb{R}^{p-1}} \{ l(\theta; \mathcal{X}_1^n) + \lambda_{(n,p,d)} || \theta_{\setminus r} ||_1 \}$
- $l(\theta; \mathcal{X}_1^n) := -\frac{1}{n} \sum_{i=1}^n \log \mathcal{P}_\theta(x_r^{(i)} | x_{\setminus r}^{(i)})$
- $\theta_{\setminus r}^* := \{\theta_{ru}^*; u \in V \setminus r\}$

Connection with Logistic regression

$$\min_{\theta_{\setminus r} \in \mathbb{R}^{p-1}} \{ \frac{1}{n} \sum_{i=1}^{n} f(\theta; x^{(i)}) - \sum_{u \in V \setminus r} \theta_{ru} \hat{\mu}_{ru} + \lambda_n ||\theta_{\setminus r}||_1 \}$$

- $f(\theta; x) := \log\{\exp(\sum_{t \in V \setminus r} \theta_{rt} x_t) + \exp(-\sum_{t \in V \setminus r} \theta_{rt} x_t)\}$
- $\hat{\mu}_{ru} := \frac{1}{n} \sum_{i=1}^{n} x_r^{(i)} x_u^{(i)}$
- $\mathcal{N}_{\pm}(r) := \{sign(\theta^*_{rt}) t | t \in \mathcal{N}(r)\}$
- $\mathcal{N}(r) := \{t \in V | (r, t) \in E\}$
- $\hat{\mathcal{N}}_{\pm}(r) := \{sign(\hat{\theta}_{ru}) u | u \in V \setminus r, \hat{\theta}_{su} \neq 0\}$
- Objective fn not strictly convex, but $\hat{\theta}^n_{\setminus r}$ is unique.

List of Assumptions

- Fisher Information matrix: $Q_r^* := E_{\theta^*}\{\Delta^2 \log \mathcal{P}_{\theta^*}[X_r|X_{\backslash r}]\}$
- $Q_r^* := E_{\theta^*}[\eta(X;\theta^*)X_{\backslash r}X_{\backslash r}^T]$
- $\eta(u;\theta) := \frac{4\exp(2u_r \sum_{t\in V\backslash r}\theta_{rt}u_t)}{(\exp(2u_r \sum_{t\in V\backslash r}\theta_{rt}u_t)+1)^2}$
- $S := \{(r,t)|t \in \mathcal{N}(r)\}$ (r is understood)
- $Q_{SS}^* := Q^*[S] \in \mathbb{R}^{d\times d}$
- Dependency: $\Lambda_{\min}(Q_{SS}^*) \geq C_{\min} > 0$ and, $\Lambda_{\max}(E_{\theta^*}[X_{\backslash r}X_{\backslash r}^T]) \leq D_{\max}$
- Incoherence: $|||Q_{S^cS}^*(Q_{SS}^*)^{-1}|||_\infty \leq 1-\alpha$

Main Result

- $\lambda_n \geq \frac{16(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$
- $L, K > 0$ independent of $(n.p, d)$, such that $n > Ld^3 \log p$
- wp at least $1 - 2\exp(-K\lambda_n^2 n)$, the following holds:
- For each node $r \in V$, the l_1 regularized logistic regression given $\mathcal{X}_1^n$, has a unique solution, uniquely specifies $\hat{\mathcal{N}}_{\pm}(r)$.
- $\hat{\mathcal{N}}_{\pm}(r)$ correctly excludes all edges not in true Neighborhood. Moreover, it correctly includes all edges (r, t), for which $|\theta^*_{r,t}| \geq \frac{10}{C_{min}}\sqrt{d}\lambda_n$

Consistency

- Consider $\{E^*_{p(n)}\}$ and parameters $\{\theta^*_{(n,p,d)}\}$.
- Dependency and Incoherence assumption holds element.
- $(n, p(n), d(n))$ satisfies the conditions above.
- $\{\lambda_n\}$ satisfies conditions above and $\lambda_n^2 n \to \infty$
- $\min_{(r,t)\in E_n^*} |\theta^*_{(n,p,d)}(r,t)| \geq \frac{10}{C_{\min}}\sqrt{d}\lambda$ for large n.
- Then $P[\hat{E}_{p(n)} = E^*_{p(n)}] \to 1$ as $n \to \infty$

Proof Approach

1. Sample Fisher Information matrix:
   $Q^n := \hat{E}[\eta(X, \theta^*) X_{\backslash r} X_{\backslash r}^T] = \frac{1}{n} \sum_{i=1}^{n} \eta(x^{(i)}; \theta^*) x_{\backslash r}^{(i)} (x_{\backslash r}^{(i)})^T$
2. Show that under Dependency and Incoherence on sample Fisher Information matrix, the growth condition on (n, p, d) and choice of λ_n are sufficient to ensure the recovery with high probability.
3. Under the specified growth condition, with incoherence and dependence assumptions on the population Fisher Information Matrix Q^* guarantees that similar results hold for sample version Q^n.

Primal Dual Witness for Graph Recovery

- primal dual pair: $(\hat{\theta}, \hat{z})$ satisfies zero sub gradient condition: $\Delta l(\hat{\theta}) + \lambda_n \hat{z} = 0$
- $\hat{z} \in \mathbb{R}^{p-1}$ must satisfy $\hat{z}_{rt} = sign(\hat{\theta}_{rt})$ if $\hat{\theta}_i \neq 0$ and $|\hat{z}_{rt}| \leq 1$ otherwise.
- We want that this primal dual pair to correctly specify the signed neighborhood of node r :
- $sign(\hat{z}_{rt}) = sign(\theta^*_{rt})\ \forall (r,t) \in S := \{(r,t) \in E\}$
- $\hat{\theta}_{ru} = 0$ for all $(r,u) \in S^C := E \setminus S$

Uniqueness of the Optimal solution

- Suppose that there exist an optimal primal solution $\hat{\theta}$ with associated optimal dual vector $\hat{z}$ such that $||\hat{z}_{S^c}||_\infty < 1$. Then any optimal primal solution $\tilde{\theta}$ must have $\tilde{\theta}_{S^c} = 0$. Moreover, if the Hessian sub-matrix $[\Delta^2 l(\hat{\theta})]_{SS}$ is strictly positive definite, then $\hat{\theta}$ is the unique optimal solution.

Construction of PDW $(\hat{\theta}, \hat{z})$

1. $\hat{\theta}_s = \arg\min_{(\theta,0)\in\mathbb{R}^{p-1}}\{l(\theta;\mathcal{X}_1^n) + \lambda_n||\theta_S||_1\}$
2. SET $\hat{z}_S = sign(\hat{\theta}_S)$
3. SET $\hat{\theta}_{S^C} = 0$
4. Get $\hat{z}_{S^C}$ from zero sub gradient condition.
5. Show with the stated (n, p, d) the remaining conditions are satisfied with high probability.

Proof part One: Sample Fisher Matrix

- "Good Event":
  $\mathcal{M}(\mathcal{X}_1^n) := \{\mathcal{X}_1^n \in \{-1, +1\}^{n \times p} | Q^n \text{satisfies A1 and A2}\}$
- If the event $\mathcal{M}(\mathcal{X}_1^n)$ holds, the sample seize satisfies $n > Ld^2 \log(p)$, and the regularization parameter is chosen such that $\lambda_n \geq \frac{16(2-\alpha)}{\alpha} \frac{\log p}{n}$. Then wp at least $1 - 2\exp(-K\lambda_n^2 n) \to 1$ the following holds:
- For each $r \in V$, the l_1-regularized logistic regression has a unique solution, and so uniquely specifies $\hat{\mathcal{N}}_{\pm}(r)$
- For each $r \in V$, the estimated signed neighborhood vector $\hat{\mathcal{N}}_{\pm}(r)$ correctly excludes all edges not in the true neighborhood and correctly includes all edges with $|\theta_{rt}| \geq \frac{10}{C_{\min}} \sqrt{d} \lambda_n$

Sample Fisher Matrix (Cont'd)

- $\Delta l(\hat{\theta}; \mathcal{X}_1^n) - \Delta l(\theta^*; \mathcal{X}_1^n) = W^n - \lambda_n \hat{z}$
- $W^n := -\Delta l(\theta^*; \mathcal{X}_1^n) =$
  $-\frac{1}{n}\sum_{i=1}^n x_{\setminus r}^{(i)} \{x_r^{(i)} - \frac{\exp(\sum_{t\in V\setminus r}\theta_{rt}^* x_t^{(i)}) - \exp(-\sum_{t\in V\setminus r}\theta_{rt}^* x_t^{(i)})}{\exp(\sum_{t\in V\setminus r}\theta_{rt}^* x_t^{(i)}) + \exp(-\sum_{t\in V\setminus r}\theta_{rt}^* x_t^{(i)})}\}$
- Co-ordinate wise mean value theorem
- $\Delta^2 l(\theta^*; \mathcal{X}_1^n)[\hat{\theta} - \theta^*] = W^n - \lambda_n \hat{z} + R^n$
- $R_j^n = [\Delta^2 l(\bar{\theta}^{(j)}; \mathcal{X}_1^n) - \Delta^2 l(\theta^*; \mathcal{X}_1^n)]_j^T (\hat{\theta} - \theta^*)$
- $\bar{\theta}^{(j)}$ is a parameter vector on the line between θ^* and $\hat{\theta}$, and $[.]_j^T$ is j'th row.

Sample Fisher Matrix (Cont'd)

- $P(\frac{2-\alpha}{\lambda_n}||W^n||_\infty \geq \frac{\alpha}{4}) \leq 2\exp(-\frac{\alpha^2\lambda_n^2}{128(2-\alpha)^2}n + \log(p))$
- Converges to zero at rate $\exp(-c\lambda_n^2 n)$ as long as $\lambda_n \geq \frac{16(2-\alpha)}{\alpha}\sqrt{\frac{\log p}{n}}$
- If $\lambda_n d \leq \frac{C_{\min}^2}{10D_{\max}}$ and $||W^n||_\infty \leq \frac{\lambda_n}{4}$, then,
- $||\hat{\theta}_S - \theta_S||_2 \leq \frac{5}{C_{\min}}\sqrt{d}\lambda_n$
- If $\lambda_n d \leq \frac{C_{\min}^2}{10D_{\max}}\frac{\alpha}{2-\alpha}$ and $||W^n||_\infty \leq \frac{\lambda_n}{4}$, then,
- $\frac{||R^n||_\infty}{\lambda_n} \leq \frac{25D_{\max}}{C_{\min}^2}\lambda_n d \leq \frac{\alpha}{4(2-\alpha)}$

Sample Fisher Matrix (Cont'd)

- Choose $\lambda_n = 16\frac{2-\alpha}{\alpha}\sqrt{\frac{\log p}{n}}$.
- By previous results, $||W^n||_\infty \leq \lambda/4$ with probability $\rightarrow 1$
- We need n to find upper bound of $\lambda_n d$
- Take $n > \frac{100^2 D_{\max}^2}{C_{\min}^4}\frac{(2-\alpha)^4}{\alpha^4} d^2 \log p$

$$\begin{aligned}
\lambda_n d &= 16\frac{2-\alpha}{\alpha}\sqrt{\frac{\log p}{n}} d \\
&\leq \frac{16 C_{\min}^2}{100 D_{\max}}\frac{\alpha}{2-\alpha} \\
&< \frac{C_{min}^2}{10 D_{\max}}
\end{aligned}$$

Hence, all conditions of previous slide are satisfied.

Sample Fisher Matrix (Cont'd)

- $Q^n_{S^CS}[\hat{\theta} - \theta^*] = W^n_{S^C} - \lambda_n \hat{z} S^C + R^n_{S^C}$
- $Q^n_{SS}[\hat{\theta} - \theta^*] = W^n_S - \lambda_n \hat{z}_S + R^n_S$
- $Q^n_{S^CS}(Q^n_{SS})^{-1}[W^n_S - \lambda_n \hat{z}_S + R^n_S] = W^n_{S^C} - \lambda_n \hat{z} S^C + R^n_{S^C}$
- $\lambda_n \hat{z}_{S^C} = [W^n_{S^C} - R^n_{S^C}] - Q^n_{S^CS}(Q^n_{SS})^{-1}[W^n_S - R^n_S] + \lambda_n Q^n_{S^CS}(Q^n_{SS})^{-1}\hat{z}_S$

$$
\begin{aligned}
||\hat{z}_{S^C}||_\infty &\leq |||Q^n_{S^CS}(Q^n_{SS})^{-1}|||_\infty \left[\frac{||W^n_S||_\infty}{\lambda_n} + \frac{||R^n_S||_\infty}{\lambda_n} + 1\right] \\
&\quad + \frac{||R^n_{S^C}||_\infty}{\lambda_n} + \frac{||W^n_{S^C}||_\infty}{\lambda_n} \\
&\leq (1-\alpha) + (2-\alpha)\left[\frac{||R^n_{S^C}||_\infty}{\lambda_n} + \frac{||W^n_{S^C}||_\infty}{\lambda_n}\right]
\end{aligned}
$$

Sample Fisher Matrix (Cont'd)

- We use the bounds on the rest of the terms.
- $||\hat{z}_{S^C}||_\infty \leq (1-\alpha) + \alpha/4 + \alpha/4 = 1 - \alpha/2$ (wp $\to 1$)
- Sign recovery: $||\theta_S - \theta_S^*||_\infty \leq \frac{\theta^*_{\text{min}}}{2}$

$$
\begin{aligned}
\frac{2}{\theta^*_{\text{min}}}||\theta_S - \theta_S^*||_\infty &\leq \frac{2}{\theta^*_{\text{min}}}||\theta_S - \theta_S^*||_2 \\
&\leq \frac{2}{\theta^*_{\text{min}}}\frac{5}{C_{\text{min}}}\sqrt{d}\lambda_n \\
&\leq 1 \ (\text{for } \theta^*_{\text{min}} > \frac{10}{C_{\text{min}}}\sqrt{d}\lambda_n)
\end{aligned}
$$

Uniform Convergence of Sample Information Matrix

- Lemma5: Suppose that dependence condition holds for the population matrix Q^* and $E_{\theta^*}[XX^T]$. For any $\delta > 0$ and some fixed constants A and B,

$$P\left[\Lambda_{\max}\left[\frac{1}{n}\sum_{i=1}^{n} x_{\setminus r}^{(i)}(x_{\setminus r}^{(i)})^T\right] \geq D_{\max} + \delta\right] \leq 2\exp(-A\frac{\delta^2 n}{d^2} + B\log(d))$$

$$P[\Lambda_{\min}(Q_{SS}^n) \leq C_{\min} - \delta] \leq 2\exp(-A\frac{\delta^2 n}{d^2} + B\log(d))$$

Uniform Convergence of Sample Information Matrix

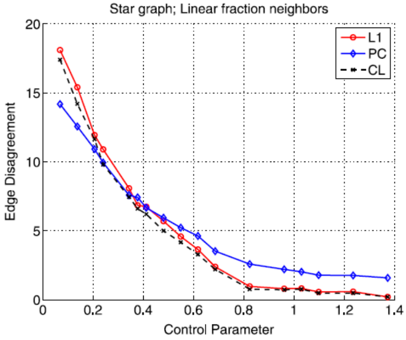
- Lemma6: If the population covariance satisfies a mutual incoherence condition with parameter $\alpha \in (0, 1]$, as in assumption, then the sample matrix satisfies an analogous version, with high probability in the sense that:

$$P[|||Q^n_{S^cS}(Q^n_{SS})^{-1}|||_\infty \geq 1 - \frac{\alpha}{2}] \leq \exp(-K\frac{n}{d^3} + \log(p))$$

Proof Idea

- $Q^n(\theta) - Q(\theta)$ can be written as an iid sum of the form $Z_{jk} = \frac{1}{n}\sum_{i=1}^{n} Z_{jk}^{(i)}$, where each $Z_{jk}^{(i)}$ is zero mean and bounded. By Azuma-Hoeffding bound,
- $P[(Z_{jk})^2 \geq \epsilon^2] = P\left[|\frac{1}{n}\sum_{i=1}^{n} Z_{jk}^{(i)}| \geq \epsilon\right] \leq 2\exp(-\frac{\epsilon^2 n}{32})$
- $\Lambda_{\min}(Q_{SS}^n) \geq C_{\min} - |||Q_{SS} - Q_{SS}^n|||_2$
- $|||Q_{SS} - Q_{SS}^n|||_2 \leq (\sum_{j=1}^{d}\sum_{k=1}^{d}(Z_{jk})^2)^{1/2}$

Simulation



Control parameter $\beta(n, p, d) = n/[10d\log(p)]$, Edge disagreement: $E[|\{(s,t)|\hat{E}_{st} \neq E^*_{st}\}|]$

Reference

High Dimensional Ising Model Selection Using l_1 Regularized Logistic Regression, by Pradeep Ravikumar, Martin Wainwright and John Lafferty. Annals of Statistics, 2010, (38) (1287-1319)