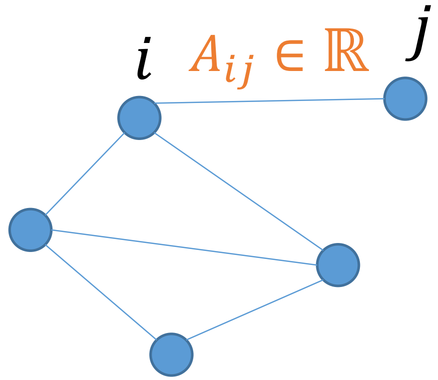
Discrete pairwise graphical model

- **Ising model**:



Edge weight b/t i & j

External field

$$\mathbb{P}[Z = z] \propto \exp\Big(\sum_{1 \le i < j \le n} A_{ij} z_i z_j + \sum_{i \in [n]} \theta_i z_i \Big)$$

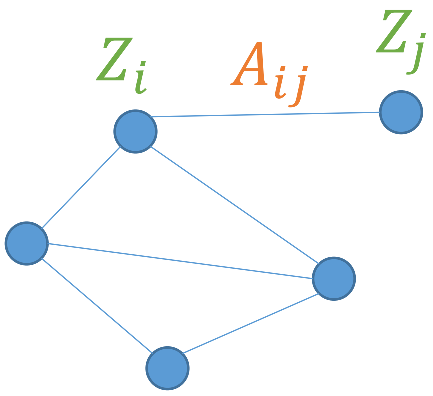
An undirected graph on n nodes → A distribution over $Z \in \{-1,1\}^n$

- **Structure learning:** Given samples, learn the graph structure.

The structure learning problem

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	……	Z_n
Sample 1	[-1	1	-1	-1	-1	1	……	1]
Sample 2	[1	-1	-1	1	-1	-1	……	1]
…				…				
Sample *N*	[1	1	-1	-1	-1	-1	……	-1]

?



Given: *N* samples from an Ising model

Goal: Recover the graph, i.e., find the edge set $\{(i,j): A_{ij} \neq 0\}$

Information-theoretic lower bound

- n = # random variables, degree $\leq d$
- Let $\lambda = \max_{i \in [n]} (\sum_j |A_{ij}| + |\theta_i|)$ Model width
- Let $\eta = \min_{(i,j): A_{ij} \neq 0} |A_{ij}|$ Minimum nonzero edge weight
- Lower bound for sample complexity [Santhanam and Wainwright'12]:

$$\boxed{d\eta \leq \lambda}$$

Information-theoretic lower bound

- n = # random variables, degree $\leq d$
- Let $\lambda = \max_{i \in [n]} (\sum_j |A_{ij}| + |\theta_i|)$ Model width
- Let $\eta = \min_{(i,j): A_{ij} \neq 0} |A_{ij}|$ Minimum nonzero edge weight
- Lower bound for sample complexity [Santhanam and Wainwright'12]:

$$\max\{\frac{\ln(n)}{2\eta \tanh(\eta)}, \frac{d}{8} \ln \frac{n}{8d}, \frac{\exp(\lambda) \ln(nd/4 - 1)}{4\eta d \exp(\eta)}\}$$

$$\boxed{d\eta \leq \lambda}$$

A natural approach to this problem…

- ℓ_1-regularized logistic regression [Ravikumar et al.'10]

- Fact: $\mathbb{P}[Z_i = 1 | Z_{-i} = x] = \sigma(w^* \cdot x')$

A natural approach to this problem…

- ℓ_1-regularized logistic regression [Ravikumar et al.'10]

$[x, 1] \in \{-1,1\}^n$

logistic/sigmoid:

$\sigma(z) = \frac{1}{1+e^{-z}}$

- Fact: $\mathbb{P}[Z_i = 1 | Z_{-i} = x] = \sigma(w^* \cdot x')$

$w^* = 2\left[A_{i1}, \dots, A_{i(i-1)}, A_{i(i+1)}, \dots, A_{in}, \theta_i\right] \in \mathbb{R}^n \longrightarrow \|w^*\|_1 \leq 2\lambda$

A natural approach to this problem…

- ℓ_1-regularized logistic regression [Ravikumar et al.'10]

- Fact: $\mathbb{P}[Z_i = 1 | Z_{-i} = x] = \sigma(w^* \cdot x')$

$[x, 1] \in \{-1,1\}^n$

logistic/sigmoid: $\sigma(z) = \frac{1}{1+e^{-z}}$

$w^* = 2\left[A_{i1}, \dots, A_{i(i-1)}, A_{i(i+1)}, \dots, A_{in}, \theta_i\right] \in \mathbb{R}^n \longrightarrow \|w^*\|_1 \leq 2\lambda$

Estimate A_{ij} to maximize the conditional likelihood ➡ ℓ_1-regularized logistic regression

A natural approach to this problem…

- ℓ_1-regularized logistic regression [Ravikumar et al.’10]
- Sample complexity: $O(d^3 \ln(\mathrm{n}))$

$\boxed{\text{Better than the lower bound}}$

- Incoherence assumption: $\left\| Q^*_{S^cS} (Q^*_{SS})^{-1} \right\|_\infty < 1$

A natural approach to this problem…

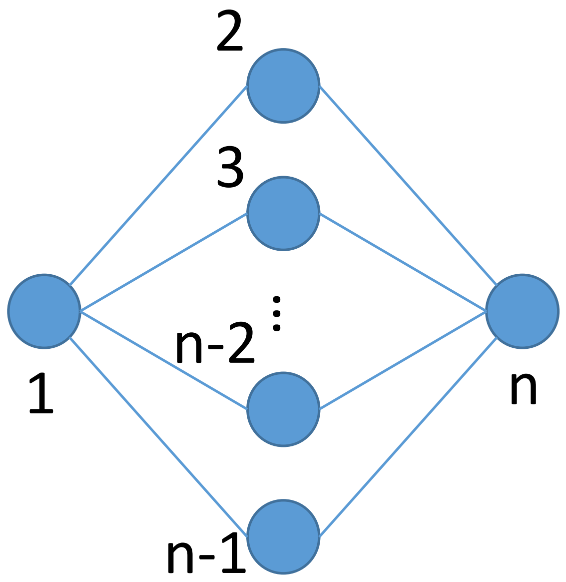
- ℓ_1-regularized logistic regression [Ravikumar et al.'10]
- Sample complexity: $O(d^3 \ln(\mathrm{n}))$

Better than the lower bound

- Incoherence assumption: $\left\| Q^*_{S^cS}(Q^*_{SS})^{-1} \right\|_\infty < 1$

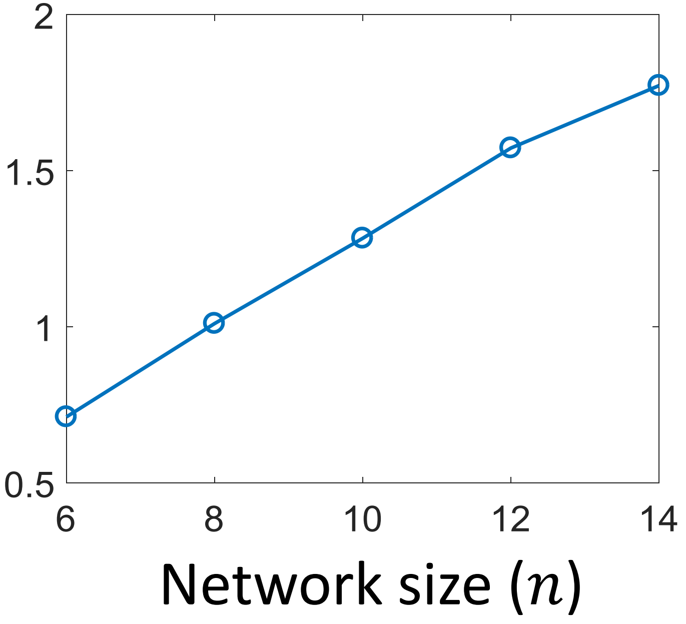
Is this condition necessary?

Is the incoherence condition necessary?

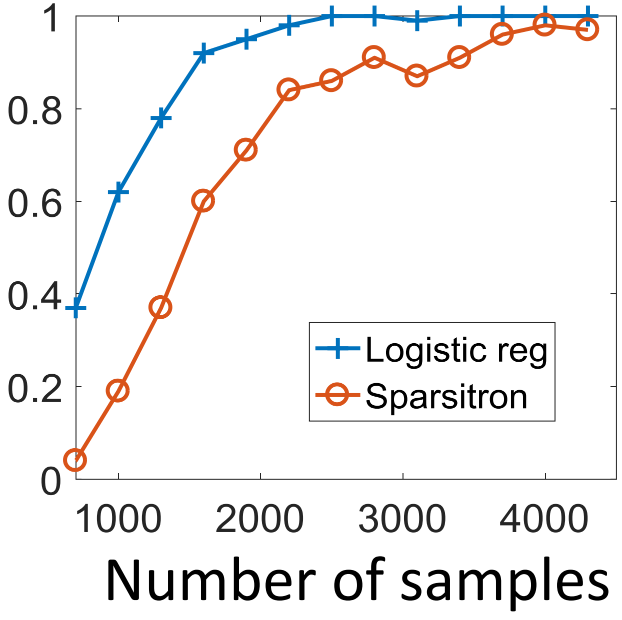


$A_{ij} = 0.2$

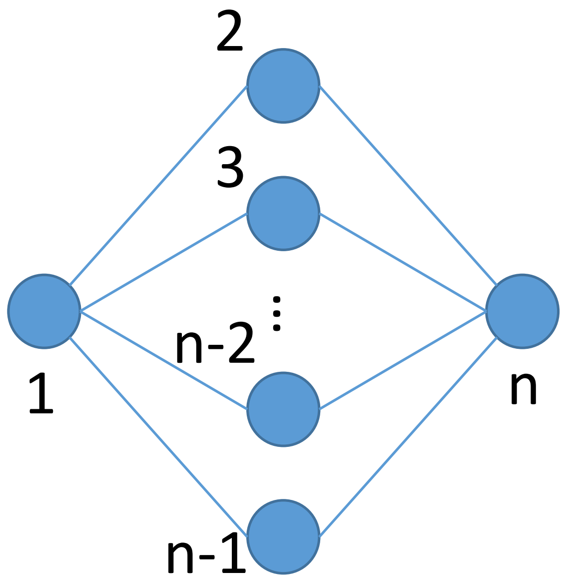
Incoherence condition is violated for $n \geq 10$



Prob of succ in 100 runs for $n = 10$

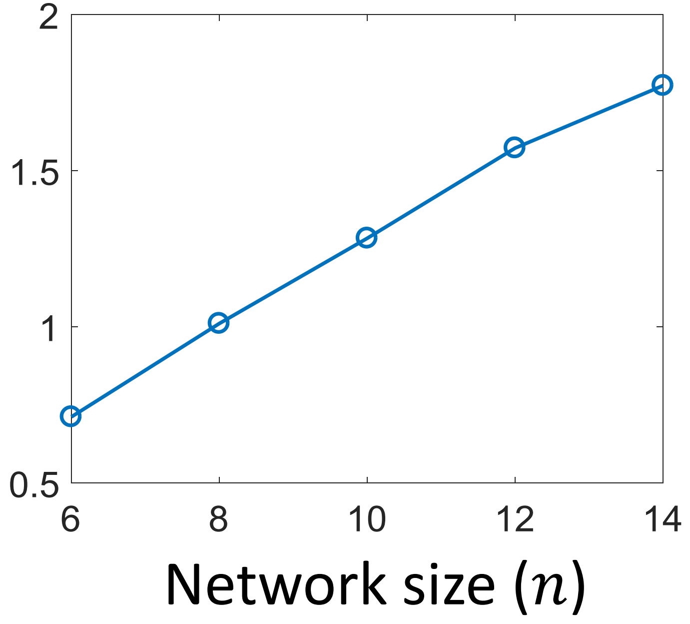


Is the incoherence condition necessary?

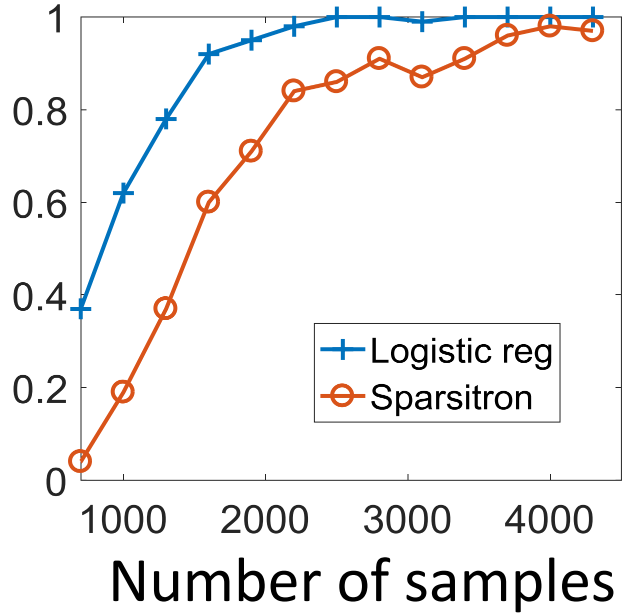


$A_{ij} = 0.2$

Incoherence condition is violated for $n \geq 10$



Prob of succ in 100 runs for $n = 10$



Our contribution: ℓ_1-constrained logistic regression learns all Ising models

Sample complexity comparison

- Recall: n = # variables, degree d, width λ, minimum edge weight η

Algorithm	Sample complexity
ℓ_1-regularized logistic regression [Ravikumar et al.'10]	$O(d^3 \ln(\mathrm{n}))$
Greedy algorithm [Bresler'15]	$O(\exp(\frac{\exp(O(d\lambda))}{\eta^{O(1)}}) \ln n)$
Interaction screening [Vuffray et al.'16]	$O(\max(d, 1/\eta^2)d^3 \exp(6\lambda)\ln(n))$
ℓ_1-regularized logistic regression [Lokhov et al.'18]	$O(\max\left(d, \frac{1}{\eta^2}\right) d^3 \exp(8\lambda)\ln(n))$
Sparsitron [Klivans and Meka'17]	$O(\frac{\lambda^2}{\eta^4} \exp(12\lambda) \ln\left(\frac{n}{\eta}\right))$
ℓ_1-constrained logistic regression [Our work]	$O(\frac{\lambda^2}{\eta^4} \exp(12\lambda) \ln(n))$

Requires incoherence & dependency conditions.

Works for arbitrary Ising models

Sample complexity comparison

- Recall: n = # variables, degree d, width λ, minimum edge weight η

Algorithm	Sample complexity
ℓ_1-regularized logistic regression [Ravikumar et al.'10]	$O(d^3 \ln(n))$
Greedy algorithm [Bresler'15]	$O(\exp(\frac{\exp(O(d\lambda))}{\eta^{O(1)}}) \ln n)$
Interaction screening [Vuffray et al.'16]	$O(\max(d, 1/\eta^2)d^3 \exp(6\lambda)\ln(n))$
ℓ_1-regularized logistic regression [Lokhov et al.'18]	$\boldsymbol{O(\max\left(d, \frac{1}{\eta^2}\right) d^3 \exp(8\lambda)\ln(n))}$
Sparsitron [Klivans and Meka'17]	$O(\frac{\lambda^2}{\eta^4} \exp(12\lambda) \ln\left(\frac{n}{\eta}\right))$
ℓ_1-constrained logistic regression [Our work]	$\boldsymbol{O(\frac{\lambda^2}{\eta^4} \exp(12\lambda) \ln(n))}$

Requires incoherence & dependency conditions.

$d \leq \lambda/\eta$

Works for arbitrary Ising models

Algorithm

- **Input**: $\{z^1, \ldots, z^N\},\ \lambda,\ \eta$
- **For** $i = 1, 2, \ldots, n$
  - $$\boxed{\begin{gathered}\widehat{w} \in \operatorname{argmin}_w \textstyle\sum_{m=1}^{N} \ln\left(1 + e^{-y^m (w \cdot x^m)}\right) / N \\ \text{s.t. } \|w\|_1 \leq 2\lambda\end{gathered}}$$
  - $\hat{A}_{ij} = \widehat{w}_{\tilde{j}}/2 \quad \text{for } j \in [n] \backslash i$
- **Output**: Edges $\{(i,j): |\hat{A}_{ij}| \geq \frac{\eta}{2}\}$

$y^m = z_i^m \in \{-1,1\}$

$x^m = [z_{-i}^m, 1] \in \{-1,1\}^n$

Given enough samples,
$|A_{ij} - \hat{A}_{ij}| < \eta/2$

Main Theorem

Theorem. Given $O(\lambda^2 \exp(12\lambda) \ln(n/\rho)/\epsilon^4)$ samples, then w.p. $\geq 1-\rho$,

$$\max_{i,j} \left|A_{ij} - \hat{A}_{ij}\right| \leq \epsilon\,.$$

$$\eta = \min_{A_{ij}\neq 0} |A_{ij}|$$

Corollary. Let $\epsilon < \eta/2$, i.e., given $O(\lambda^2 \exp(12\lambda) \ln(n/\rho)/\eta^4)$ samples, then w.p. $\geq 1-\rho$, the algorithm recovers the graph structure.

Proof overview

$$\hat{L}(w) \coloneqq \sum_{i=1}^{N} \ln(1 + e^{-y^i (w \cdot x^i)}) \,/N$$

$$\boxed{\begin{aligned}\hat{w} \in \underset{w \in \mathbb{R}^n}{\operatorname{argmin}} \ \hat{L}(w) \\ \text{s.t. } \|w\|_1 \leq 2\lambda\end{aligned}} \xRightarrow{?} \boxed{\left\|\hat{w}_{1:(n-1)} - w^*_{1:(n-1)}\right\|_\infty \leq \epsilon}$$

Proof overview

$$\hat{L}(w) \coloneqq \sum_{i=1}^{N} \ln\left(1 + e^{-y^i (w \cdot x^i)}\right) / N$$

$$\left|\hat{A}_{ij} - A_{ij}\right| \le \epsilon$$

$$\hat{w} \in \underset{w \,\in\, \mathbb{R}^n}{\operatorname{argmin}} \; \hat{L}(w) \quad \text{s.t. } \|w\|_1 \le 2\lambda$$

?

$$\left\|\hat{w}_{1:(n-1)} - w^*_{1:(n-1)}\right\|_\infty \le \epsilon$$

$$w^* = 2\left[A_{i1}, \dots, A_{i(i-1)}, A_{i(i+1)}, \dots, A_{in}, \theta_i\right] \in \mathbb{R}^n$$

Proof outline

- **Step 1:** Given $O(\lambda^2 \ln(n/\rho)/\gamma^2)$ samples, w.p. $\geq 1 - \rho$
$$L(\hat{w}) - L(w^*) \leq \gamma.$$

Proof outline

- **Step 1:** Given $O(\lambda^2 \ln(n/\rho)/\gamma^2)$ samples, w.p. $\geq 1-\rho$
$$L(\widehat{w}) - L(w^*) \leq \gamma.$$
- **Step 2:** For any w,
$$L(w) - L(w^*) \geq \mathrm{E}_X\big(\sigma(w \cdot X) - \sigma(w^* \cdot X)\big)^2.$$

Proof outline

$$\mathrm{E}_X\big(\sigma(\widehat{w}\cdot X)-\sigma(w^*\cdot X)\big)^2 \le \gamma$$

- **Step 1:** Given $O(\lambda^2 \ln(n/\rho)/\gamma^2)$ samples, w.p. $\ge 1-\rho$
$$L(\widehat{w}) - L(w^*) \le \gamma.$$
- **Step 2:** For any w,
$$L(w) - L(w^*) \ge \mathrm{E}_X\big(\sigma(w\cdot X)-\sigma(w^*\cdot X)\big)^2.$$

Proof outline

$$\mathrm{E}_X\big(\sigma(\widehat{w}\cdot X)-\sigma(w^*\cdot X)\big)^2 \le \gamma$$

- **Step 1:** Given $O(\lambda^2 \ln(n/\rho)/\gamma^2)$ samples, w.p. $\ge 1-\rho$

$$L(\widehat{w}) - L(w^*) \le \gamma.$$

- **Step 2:** For any w,

$$L(w) - L(w^*) \ge \mathrm{E}_X\big(\sigma(w\cdot X)-\sigma(w^*\cdot X)\big)^2.$$

- **Step 3** [Lemma 4.3 in Klivans and Meka'17]: If samples ~ Ising model,

$$E_X\big(\sigma(w\cdot X)-\sigma(w^*\cdot X)\big)^2 \le \gamma$$

$$\Big\downarrow \; \gamma = O(\epsilon^2/\exp(6\lambda))$$

$$\left\|w_{1:(n-1)} - w^*_{1:(n-1)}\right\|_\infty \le \epsilon$$

Main Theorem

Theorem. Given $O(\lambda^2 \exp(12\lambda) \ln(n/\rho)/\epsilon^4)$ samples, then w.p. $\geq 1-\rho$,

$$\max_{i,j} \left|A_{ij} - \hat{A}_{ij}\right| \leq \epsilon \,.$$

$$\eta = \min_{A_{ij} \neq 0} |A_{ij}|$$

Corollary. Let $\epsilon < \eta/2$, i.e., given $O(\lambda^2 \exp(12\lambda) \ln(n/\rho)/\eta^4)$ samples, then w.p. $\geq 1-\rho$, our algorithm recovers the graph structure.

Efficiently solving the optimization problem

$$\boxed{\begin{aligned} \hat{w} \in \operatorname*{argmin}_{w \in \mathbb{R}^n} \ \hat{L}(w) \\ \text{s.t. } \|w\|_1 \leq 2\lambda \end{aligned}} \Rightarrow \boxed{\|\hat{w} - w^*\|_\infty \leq \epsilon} \Rightarrow \boxed{|\hat{A}_{ij} - A_{ij}| \leq \epsilon}$$

Efficiently solving the optimization problem

$$\hat{w} \in \underset{w \in \mathbb{R}^n}{\operatorname{argmin}} \ \hat{L}(w) \quad \text{s.t. } \|w\|_1 \leq 2\lambda$$

After $O(\ln(n)/\gamma^2)$ mirror descent steps

$$\hat{L}(\bar{w}) - \hat{L}(\hat{w}) \leq \gamma$$

$$\|\hat{w} - w^*\|_\infty \leq \epsilon$$

$$\|\bar{w} - w^*\|_\infty \leq \epsilon$$

$$|\hat{A}_{ij} - A_{ij}| \leq \epsilon$$

Efficiently solving the optimization problem

$$\hat{w} \in \underset{w \in \mathbb{R}^n}{\operatorname{argmin}} \ \hat{L}(w) \quad \text{s.t. } \|w\|_1 \leq 2\lambda$$

After $O(\ln(n)/\gamma^2)$ mirror descent steps

$$\hat{L}(\overline{w}) - \hat{L}(\hat{w}) \leq \gamma$$

Only need to prove Step 1:

$$L(\overline{w}) - L(w^*) \leq O(\gamma)$$

$$\|\hat{w} - w^*\|_\infty \leq \epsilon$$

$$\|\overline{w} - w^*\|_\infty \leq \epsilon$$

$$|\hat{A}_{ij} - A_{ij}| \leq \epsilon$$

Efficiently solving the optimization problem

$$\hat{w} \in \underset{w \in \mathbb{R}^n}{\operatorname{argmin}} \ \hat{L}(w) \quad \text{s.t.} \ \|w\|_1 \le 2\lambda$$

$$\|\hat{w} - w^*\|_\infty \le \epsilon$$

$$\|\overline{w} - w^*\|_\infty \le \epsilon$$

$$|\hat{A}_{ij} - A_{ij}| \le \epsilon$$

After $O(\ln(n)/\gamma^2)$ mirror descent steps

Time: $\tilde{O}(n)$ per r.v.

$$\hat{L}(\overline{w}) - \hat{L}(\hat{w}) \le \gamma$$

Only need to prove Step 1:

$$L(\overline{w}) - L(w^*) \le O(\gamma)$$

Algorithm (for general alphabet)

- Let k be the alphabet size

Ising model (k=2)

$$\hat{w} \in \underset{w \in \mathbb{R}^n}{\operatorname{argmin}} \ \hat{L}(w) \quad \text{s.t. } \|w\|_1 \le 2\lambda$$

General k

$$\hat{w} \in \underset{w \in \mathbb{R}^{n \times k}}{\operatorname{argmin}} \ \hat{L}(w) \quad \text{s.t. } \|w\|_{2,1} \le 2\lambda\sqrt{k}$$

Group sparsity

Our contribution (for general alphabet)

Algorithm	Sample complexity
Greedy algorithm [Hamilton et al.'17]	$O(\exp(\frac{k^{O(d)}\exp(d^2\lambda)}{\eta^{O(1)}})\ln(nk))$
Sparsitron [Klivans and Meka'17]	$O(\frac{\lambda^2 k^5 \exp(14\lambda)}{\eta^4}\ln\left(\frac{nk}{\eta}\right))$
$\ell_{2,1}$-constrained logistic regression [Our work]	$O(\frac{\lambda^2 k^4 \exp(14\lambda)}{\eta^4}\ln(nk))$

Improves from k^5 to k^4!

Improve the sample complexity from k^5 to k^4

$\ell_{2,1}$-constrained logistic regression [Our work]

$$\|w\|_{2,1} \le 2\lambda\sqrt{k}$$

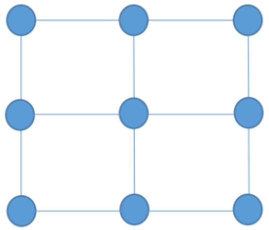
$$O\left(\frac{\lambda^2 k^4 \exp(14\lambda)}{\eta^4}\ln(nk)\right)$$

Sparsitron [Klivans and Meka'17]

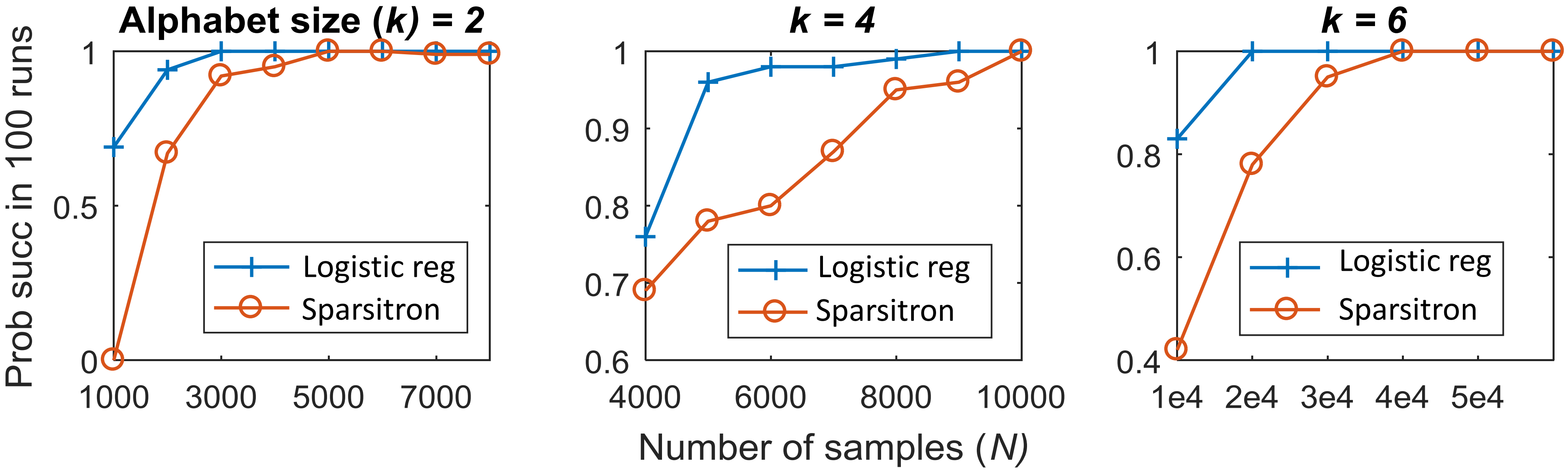
$$\|w\|_1 \le \sqrt{k}\|w\|_{2,1} \le 2\lambda k$$

$$O\left(\frac{\lambda^2 k^5 \exp(14\lambda)}{\eta^4}\ln\left(\frac{nk}{\eta}\right)\right)$$

Experiments



- 3-by-3 grid graph: $n = 9$
- Alphabet sizes: $k = 2, 4, 6$
- Edge weights: random ±0.2
- Run 100 simulations



Sparse logistic regression requires fewer samples for graph recovery.