Graph-Sparse Logistic Regression

Alexander LeNail[1], Ludwig Schmidt[2], Jonathan Li[1], Tobias Ehrenberger[1], Karen Sachs[1], Stefanie Jegelka[2], Ernest Fraenkel[1]

[1]MIT BE, [2]MIT CSAIL

Problem Setup

Variable selection in a linear model:

$$y \;=\; \sigma(X\,\theta^*)$$

- Data matrix $X \in \mathbb{R}^{n \times d}$
- Unknown parameters $\theta^* \in \mathbb{R}^d$
- Binary labels $y \in \{0,1\}^n$
- $\sigma : \mathbb{R} \to \mathbb{R}$ is the logistic function $\sigma(x) = \frac{1}{1+e^{-x}}$

This work: our goal is to select a **graph-sparse** set of variables.

→ **Statistical efficiency:** fewer variables for same error.

→ **Interpretability:** graph-structure in many applications.

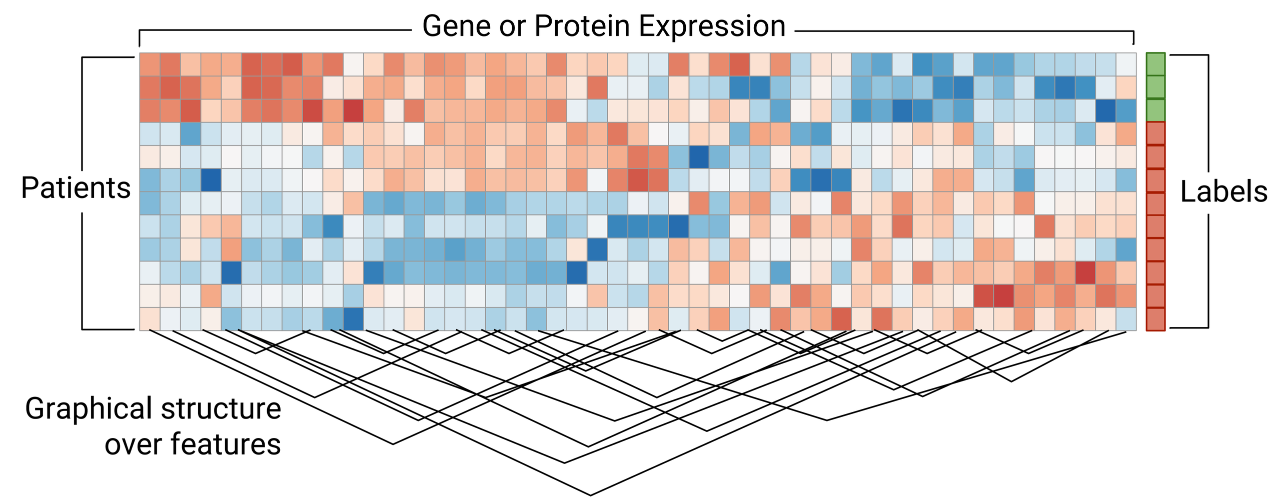
Graph sparsity

- Every variable (parameter index) corresponds to a node.
- Selected variables form a **connected subgraph.**

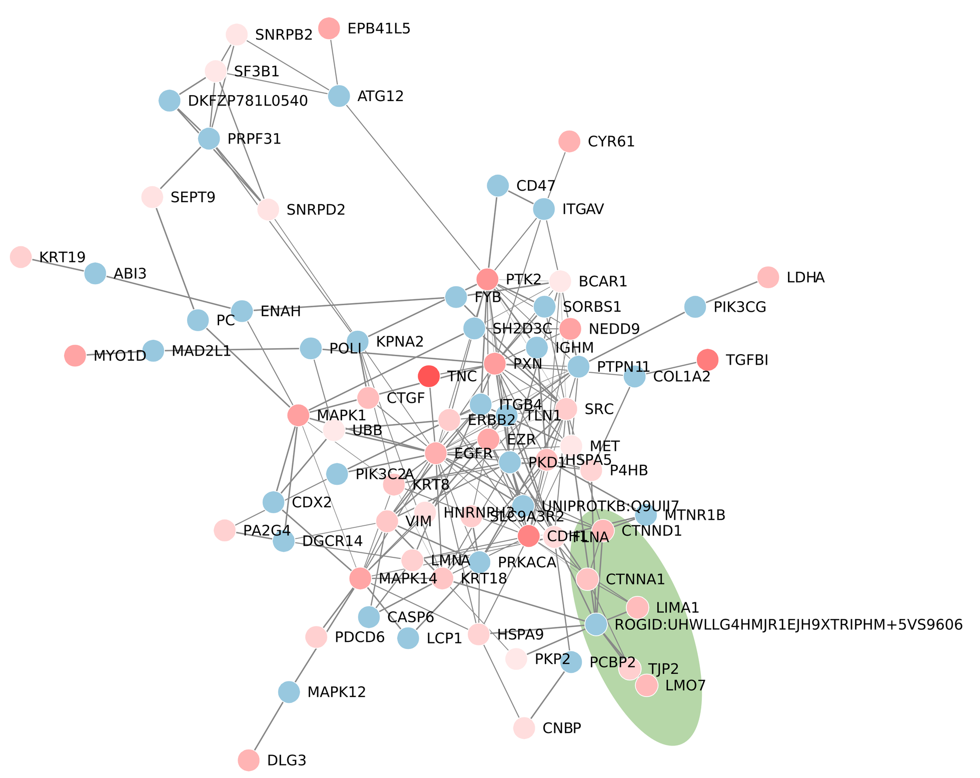
Graph-sparse   Not graph-sparse

Motivation: graph-structed data in biology

Protein expression data with a **protein-protein interation network.**



Graph given by prior knowledge from biology:



Approach

We build on results in **compressive sensing** for graph-sparse data. [Huang, Zhang, Metaxas, 2011], [Hegde, Indyk, Schmidt, 2015]

→ We introduce **Graph-Sparse Logistic Regression** (GSLR).

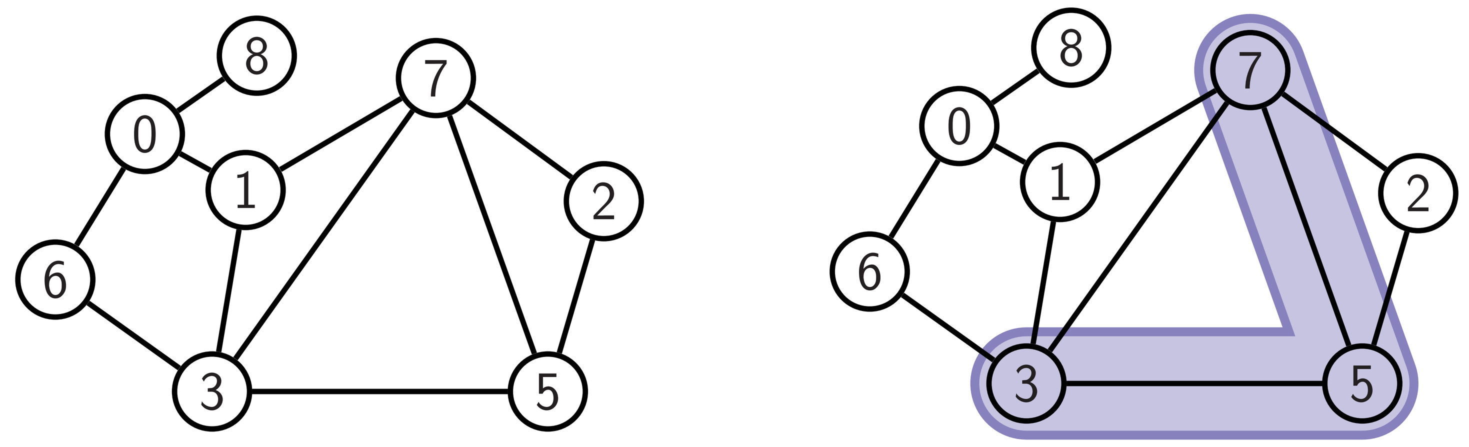
- Gradient descent on logistic loss.
- Efficient projections onto the graph sparse set.

function $\mathrm{GSLR}(X, y, G, s, \eta, k)$
  Let $f(X, y, \theta)$ be the logistic loss.
  $\hat{\theta}^0 \leftarrow 0$
  for $i \leftarrow 0, \ldots, k-1$ **do**
    $\tilde{\theta}^{i+1} \leftarrow \hat{\theta}^i - \eta \cdot \nabla f(X, y, \hat{\theta}^i)$
    $\hat{\theta}^{i+1} \leftarrow P_{G,s}(\tilde{\theta}^{i+1})$ ▷ Graph-sparse projection
  return $\hat{\theta}^k$

Efficient Graph-Sparse Projections

Projection problem: Given $b \in \mathbb{R}^d$ and a graph-sparse set $\mathbb{G}$, find

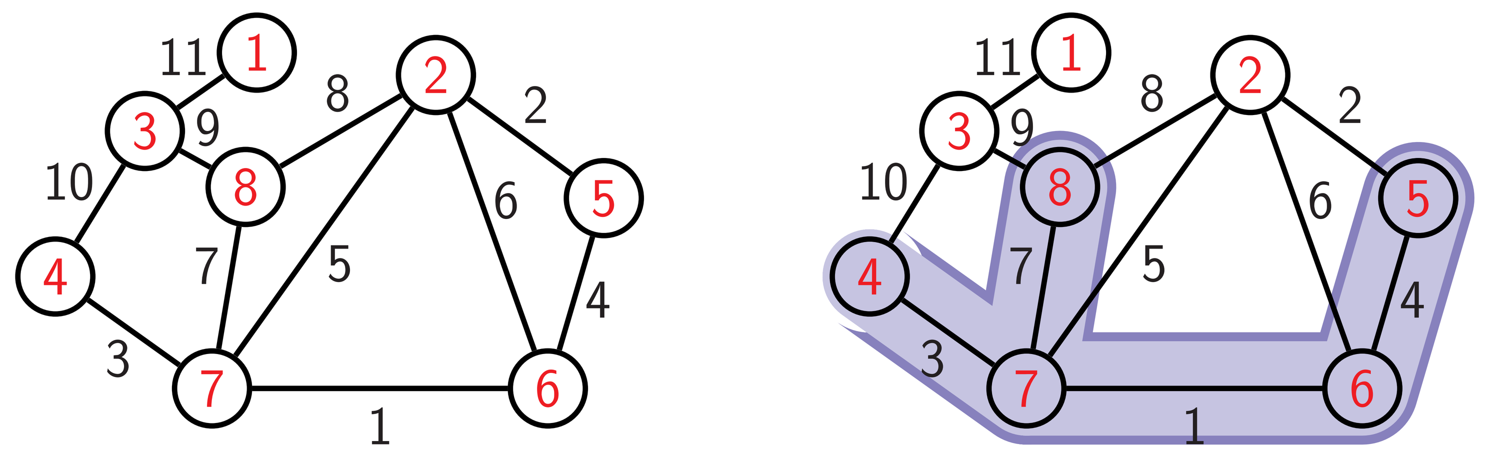
$$\Omega^* \;=\; \operatorname*{arg\,min}_{\Omega \in \mathbb{G}} \|b - b_\Omega\| \,.$$



Input   Output

We solve **approximate** versions of the projection problem via reductions to the **prize-collecting Steiner tree problem** (PCST).

Objective of PCST: Given a graph with edge costs c and node prizes π, find a subtree T minimizing $c(T) + \pi(\overline{T})$.



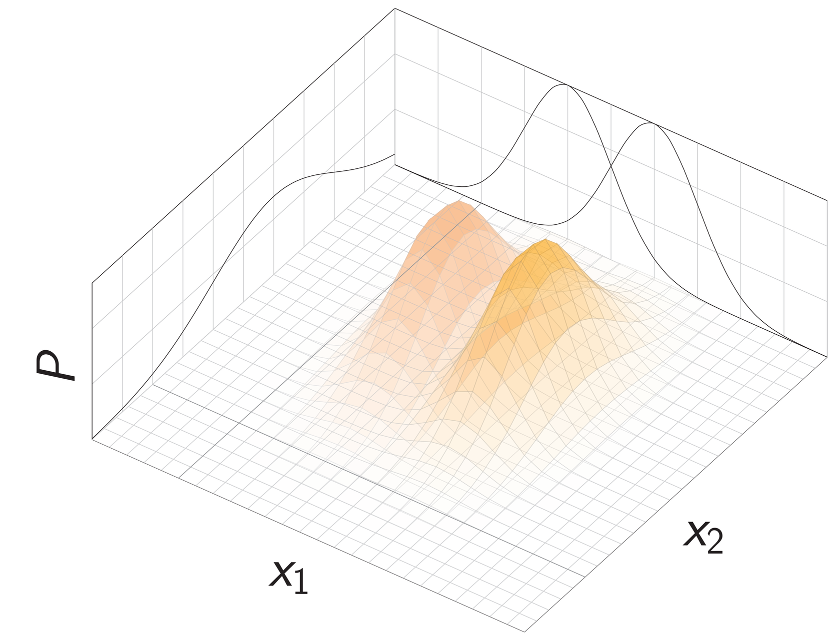
Input   Output

→ **Nearly-linear** time approximate projections.

Experimental Setup

Since we don't have the ground truth subgraphs for the real Ovarian Cancer data, we generate synthetic data by this procedure:

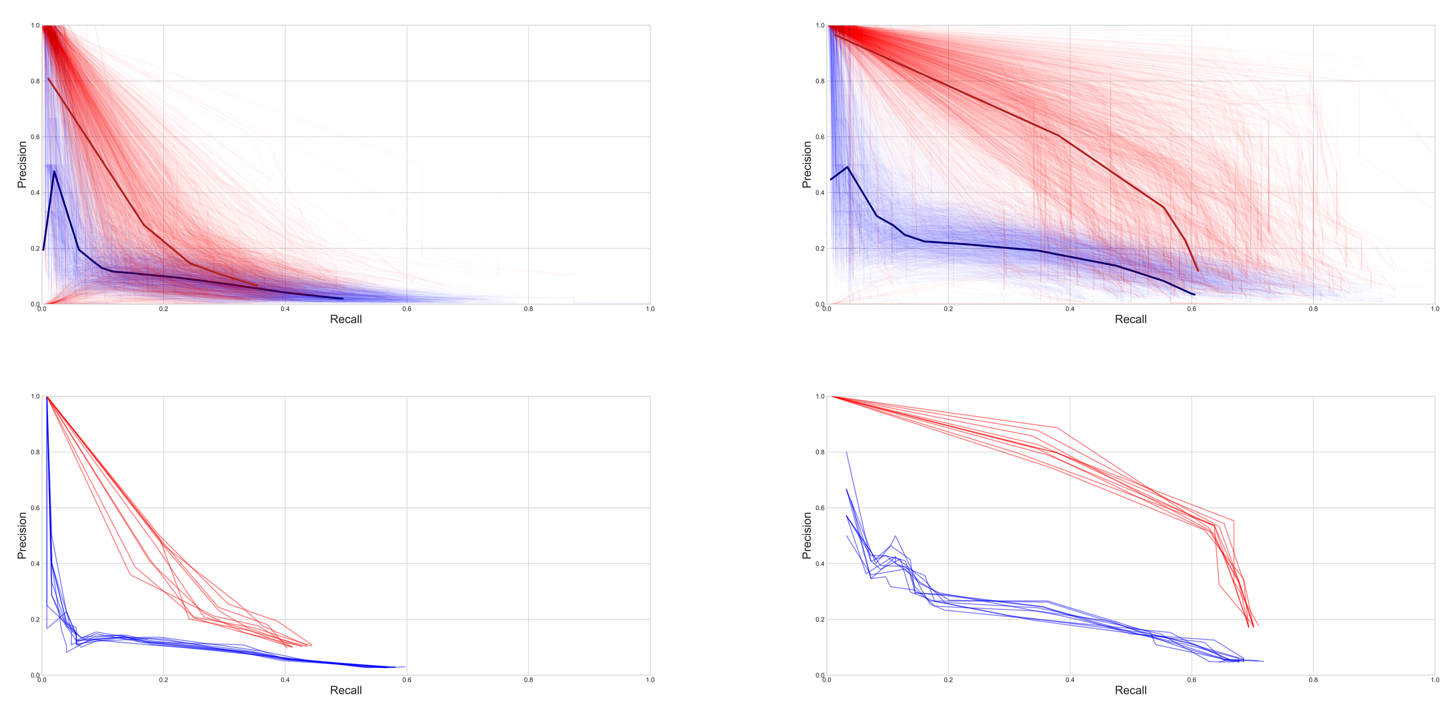
1. Determine μ and Σ from real Ovarian Cancer Proteomics data.
2. Sample from multivariate $\mathcal{N}(\vec{\mu}, \Sigma)$
3. Sample perturbation vector $\vec{x}$:
   scheme 1: $\vec{x_p} = \mathcal{N}(0, \sigma_p^2)$ if $p \in$ KEGG, 0 otherwise
   scheme 2: $\vec{x_p} = \mathcal{N}(\pm\sigma_p, \sigma_p^2)$ if $p \in$ KEGG, 0 otherwise
4. Translate "positive" samples by perturbation vector



Since we know the perturbation vector, we know the ground truth! We can then evaluate algorithms on the feature selection task.

Experimental Results

We benchmark GSLR against the LASSO by how many of the truly "perturbed" features each uses in its support.



We then use our technique on real Ovarian Cancer data, and find that the support chosen by GSLR is qualitatively superior.

Conclusion

Source code and experimental code at https://github.com/fraenkel-lab/gslr

Future Work: Benchmark against related approaches which incorporate the feature graph.