# 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

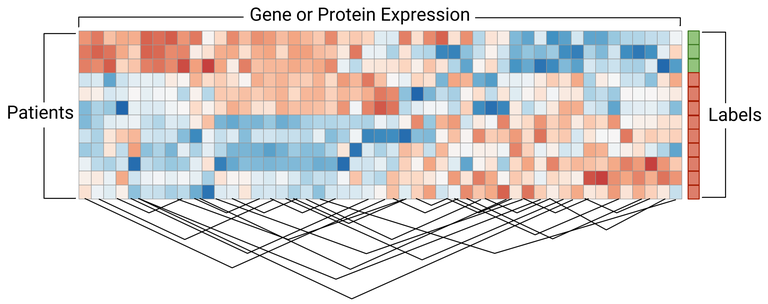
## Introduction

**Scientists** often use machine learning to **learn more about their data**, not to predict it. They want to **interpret** their models.
The simplest way to build an interpretable model is to keep it **small**.
The classic way to accomplish this is with **LASSO** (L1-regularization).

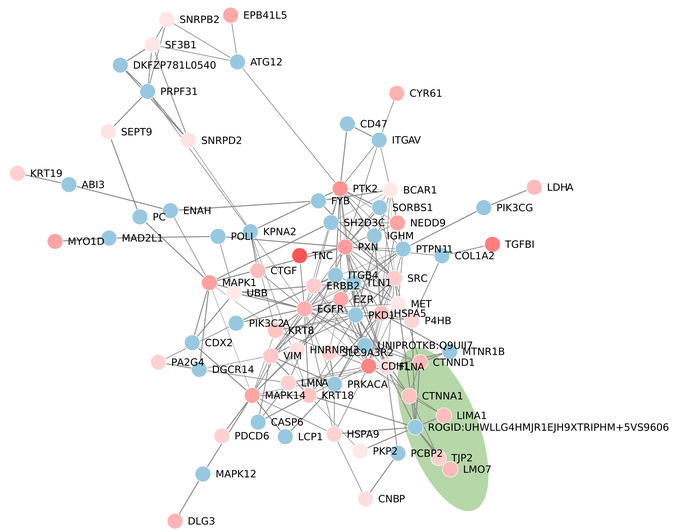
## Problem setup

The LASSO does not always find the "right" model. In some cases, we can leverage side-information to help find the right model.

In our case, side-information is graphical structure among the features.



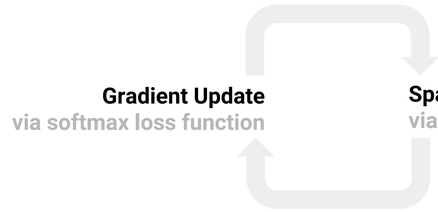

We can re-draw each instance as a labeling of the graph's nodes:




The "right" model is a "graph-sparse" model, i.e. a model with a sparse support which is connected on a graph. The goal is therefore to find the **most predictive connected subgraph**.

## Our approach

Main idea:




[t]

**function** 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** $\hat{\theta}^k$

## Synthetic Data

Since we don't have the ground truth subgraph for our real data, we create synthetic data by sampling from the multivariate gaussian defined by the real data, then translating "positive" examples by a "perturbation" vector. $\textit{negative} = \mathcal{N}(\vec{\mu}, \Sigma)$ $\textit{positive} = \mathcal{N}(\vec{\mu}, \Sigma) + \vec{x}$

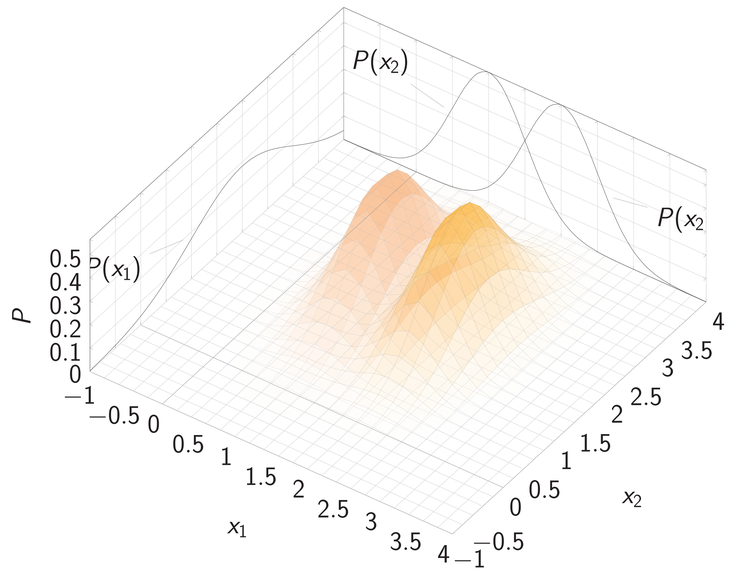
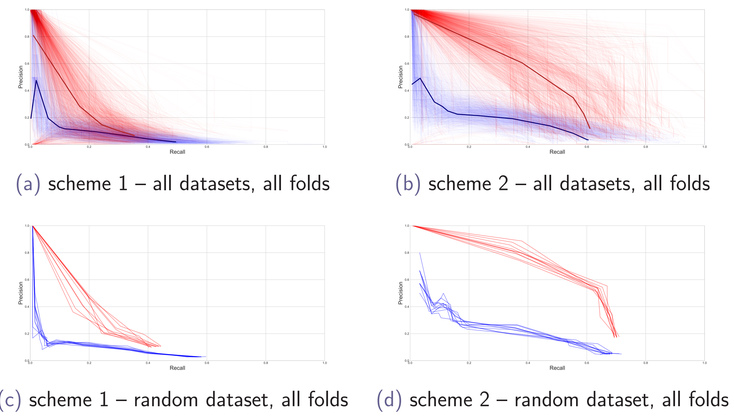



Figure: A low dimensional cartoon of our synthetic data generation strategy. Here, the pink gaussian represents our original data, from which we sample our negative examples. We sample our positive examples from the orange gaussian by first sampling them from the pink gaussian and translating them by the perturbation vector (in this case $< 0, \mu_{x_2} - \mu_{x_2*} >$, with $x_2$ in the pathway and $x_1$ not.

## Experiments

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




(a) scheme 1 – all datasets, all folds (b) scheme 2 – all datasets, all folds

(c) scheme 1 – random dataset, all folds (d) scheme 2 – random dataset, all folds

Figure: L1-Regularized Logistic Regression (blue) versus Graph-Sparse logistic Regression (red). Each trace represents one fold of one dataset, varying sparsity. The bolded trace is the average.

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

## Conclusion