Learning the Structure of Generative Models without Labeled Data

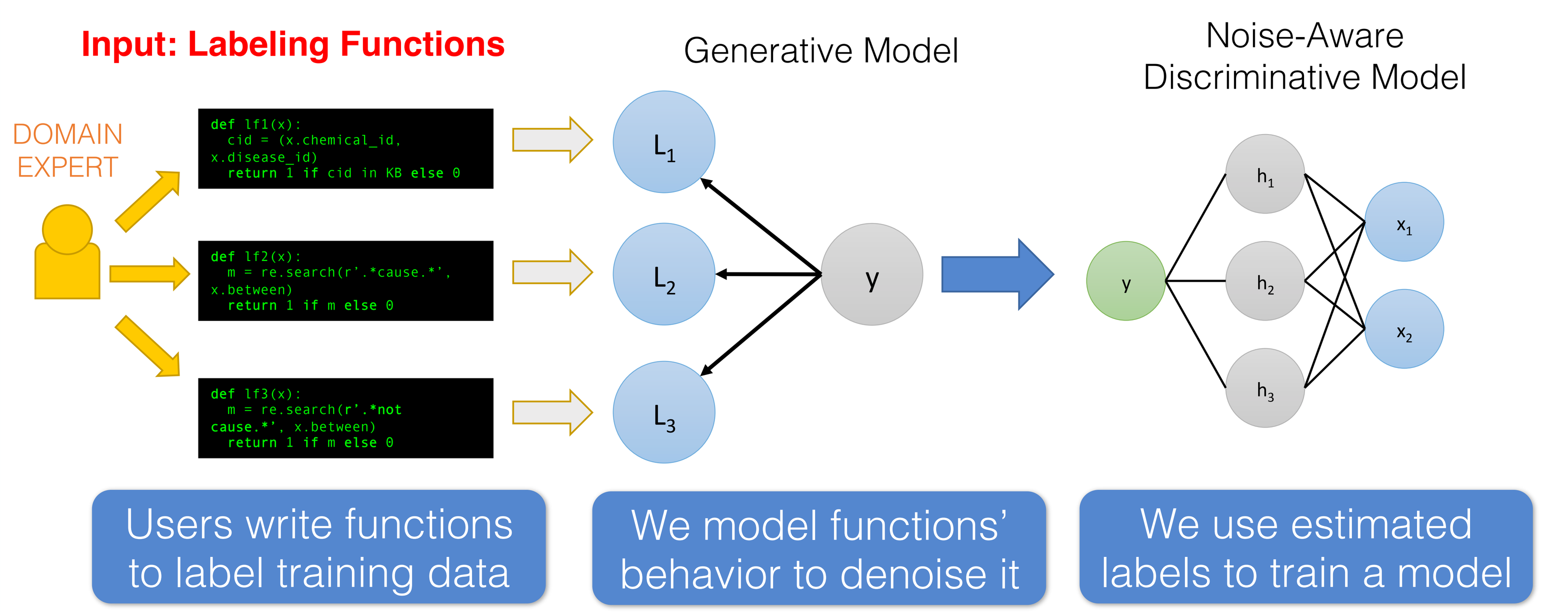
Stephen Bach, Bryan He, Alex Ratner, and Chris Ré

Stanford University

Motivation: Generative Models for Weak Supervision

- Curating training data has become the biggest bottleneck when developing machine learning applications
- Weak supervision sources, such as heuristic rules, distant supervision, and weak classifiers, are much less expensive
- However, weak supervision sources are noisy and conflict
- Idea: encapsulate weak supervision sources as *labeling functions* and model their behavior to denoise their outputs
- By modeling the true class as a latent variable, we can estimate generative parameters *without ground truth*

Learning Pipeline



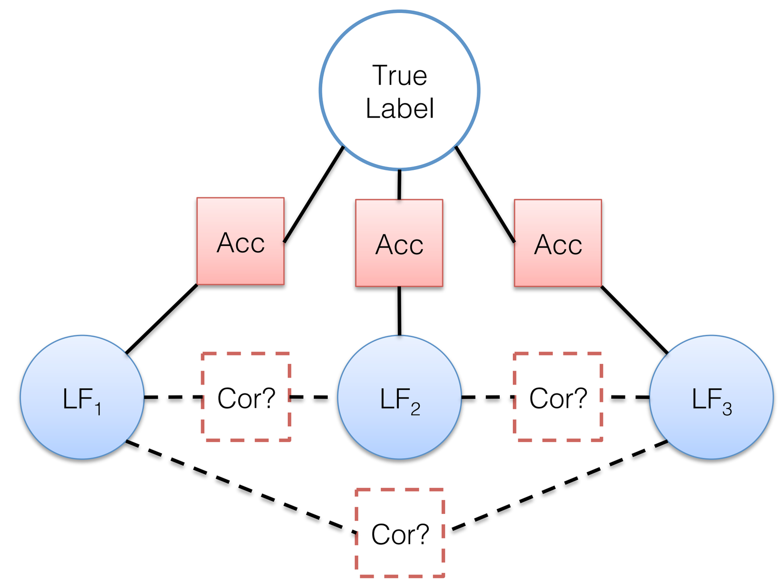
Snorkel: Open source implementation available at `http://snorkel.stanford.edu`

Structure Learning for Generative Models

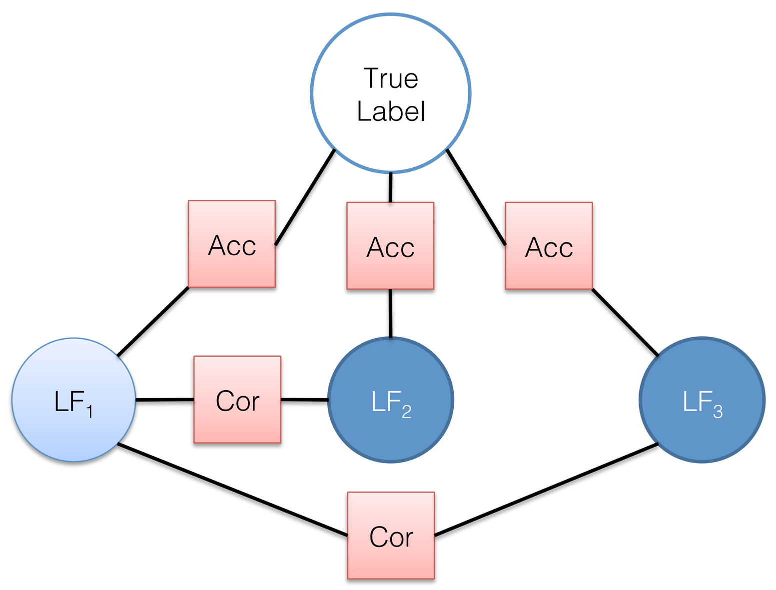
- When domain experts write labeling functions, they often introduce statistical dependencies among them
- Incorrectly modeling dependencies leads to inaccurate estimation of true, latent classes
- Goal is to quickly identify labeling function dependencies

Key

- Latent Variable
- Target Variable
- Conditioning Variable
- Dependency
- Possible Dependency



Our approach: maximize l1-regularized marginal pseudolikelihood

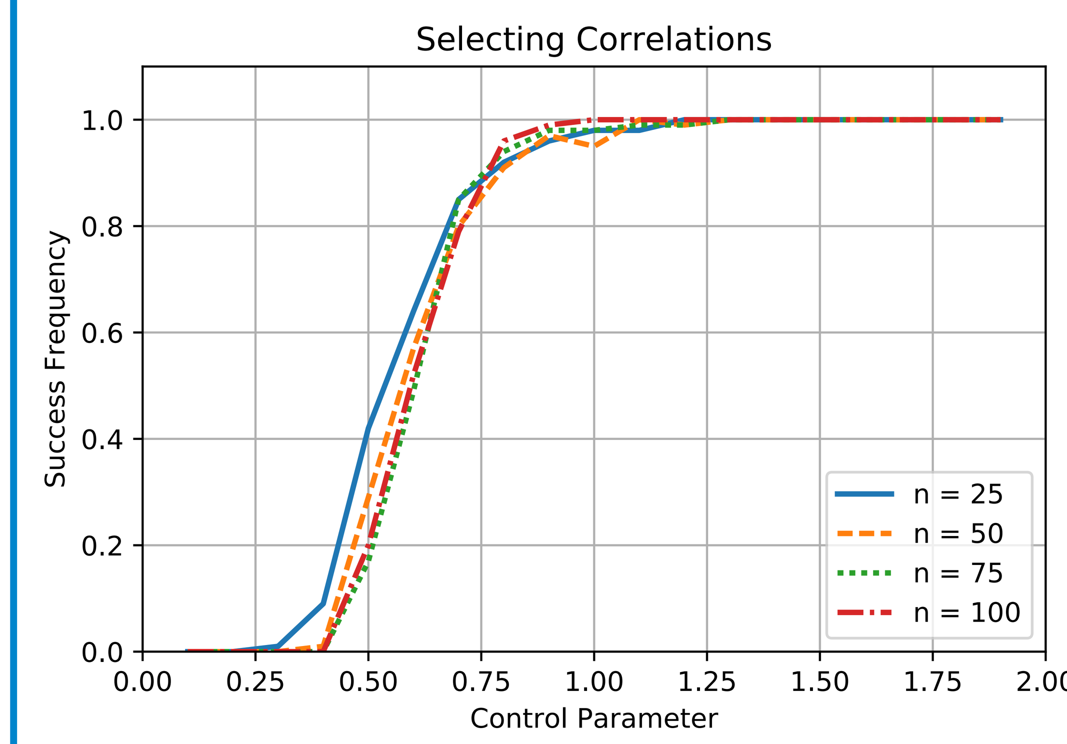


Since the marginal pseudolikelihood objective has only one target variable and one latent variable, efficient to compute gradient exactly

We optimize for each LF and add the dependencies with nonzero weight to the generative model

Empirical Results

Sample Complexity



How is the probability of recovering the true structure affected by problem parameters?

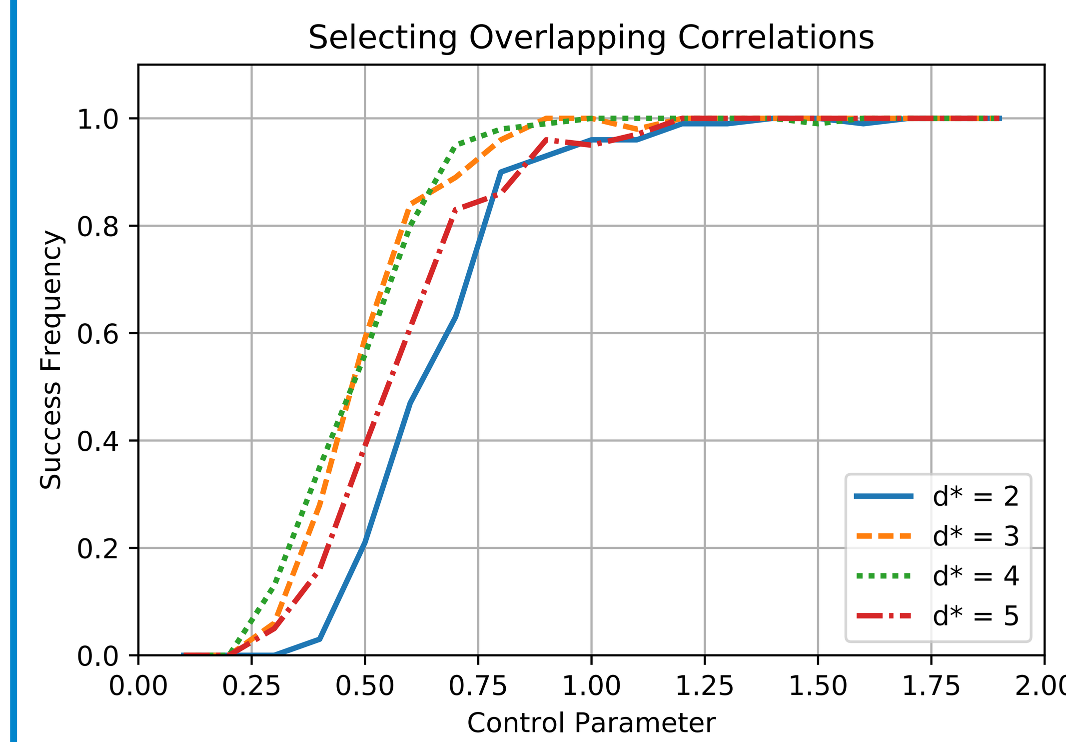
$$m \propto \gamma \cdot d^{\star} \cdot n$$

m : training samples

γ : control parameter

$d^{\star}$: max. degree

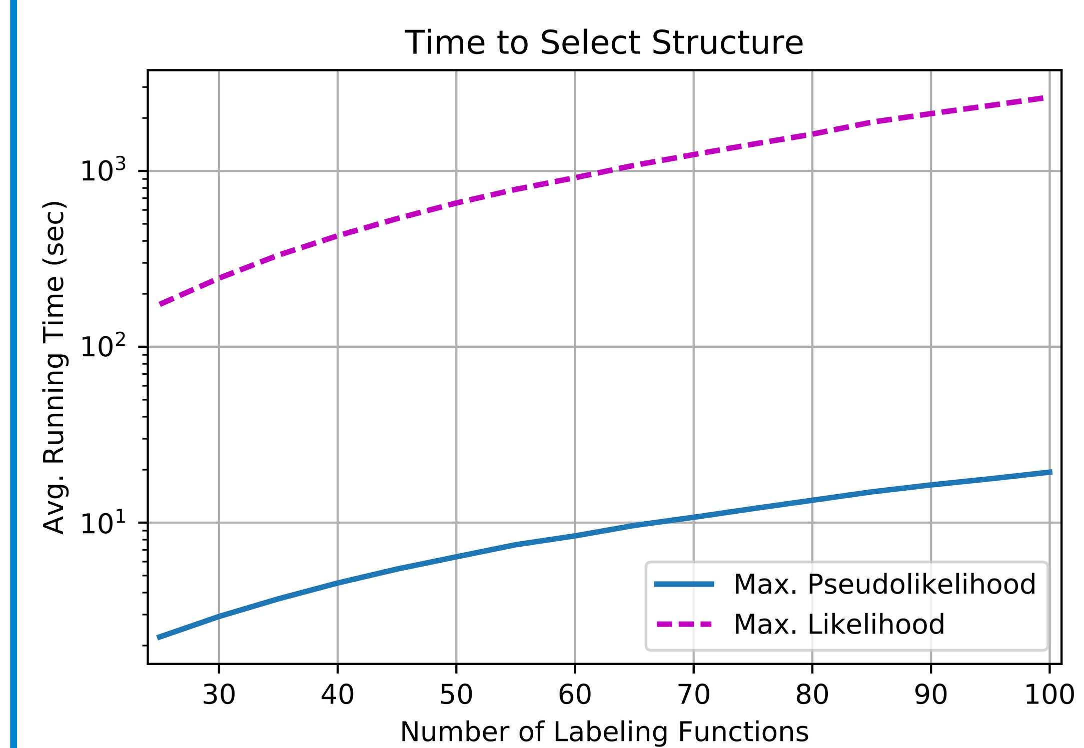
n : labeling functions



Takeaways:

1. Sample complexity better in practice than in theory
2. We observe same sample complexity as in supervised setting (where bounds are also pessimistic)

Speed Up: 100x



Efficient, exact gradient computation leads to two-order-of-magnitude speedup over MLE with Gibbs sampling

Reduction of learning time to seconds enables human-in-the-loop development of labeling functions

Applications

Application	Ind. F1	Struct. F1	F1 Diff	# LF	# Dep.
Disease Tagging	66.3	68.9	+2.6	233	315
Chem-Disease	54.6	55.9	+1.3	33	21
Device Polarity	88.1	88.7	+0.6	12	32

Consistent improvements to information extraction models trained on labels estimated from generative models with learned structure

These experiments used existing labeling functions, demonstrating that modeling structure can even improve the performance of carefully developed weak supervision sources

Example Application: Information Extraction

Task: identify mentions of chemicals causing diseases in scientific literature

ID	Chemical	Disease	Prob.
00	magnesium	Myasthenia gravis	0.84
01	magnesium	quadriplegic	0.73
02	magnesium	paralysis	0.96

TITLE:
Myasthenia gravis presenting as weakness after magnesium administration.
ABSTRACT:
We studied a patient with no prior history of neuromuscular disease who became virtually quadriplegic after parenteral magnesium administration for preeclampsia. The serum magnesium concentration was 3.0 mEq/L, which is usually well tolerated. The magnesium was stopped and she recovered over a few days. While she was weak, 2-Hz repetitive stimulation revealed a decrement without significant facilitation at rapid rates or after exercise, suggesting postsynaptic neuromuscular blockade. After her strength returned, repetitive stimulation was normal, but single fiber EMG revealed increased jitter and blocking. Her acetylcholine receptor antibody level was markedly elevated. Although paralysis after magnesium administration has been described in patients with known myasthenia gravis, it has not previously been reported to be the initial or only manifestation of the disease. Patients who are unusually sensitive to the neuromuscular effects of magnesium should be suspected of having an underlying disorder of neuromuscular transmission.

Example Labeling Functions

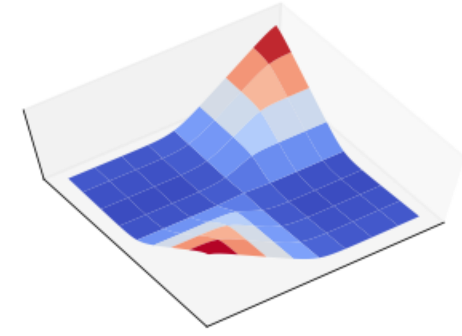
```
def LF_heuristic(x):
  m = re.match('.*caused.*', x.sentence)
  return True if m else None
```

```
def LF_distant_supervision(x):
  in_kb = (x.chemical, x.disease) in ctd
  return True if in_kb else None
```

using Comparative Toxicogenomics Database (`http://ctdbase.org`)

Analysis: Sample Complexity

Challenge: Marginal pseudolikelihood is nonconvex, but previous analyses of l1-regularized parameter estimation for structure learning rely on Lagrangian duality



Assumptions:

1. Feasible set of parameters that contains the true model
2. Over the feasible set, conditioning on a labeling function provides more information than marginalizing it out

Theorem: For pairwise dependencies, such as correlations,

$$m \geq \Omega\left(n \log \frac{n}{\delta}\right)$$

samples are sufficient to recover true dependency structure over n labeling functions with probability at least 1 - δ.

We gratefully acknowledge the support of the Defense Advanced Research Projects Agency (DARPA) SIMPLEX program under No. N66001-15-C-4043, the DARPA D3M program under No. FA8750-17-2-0095, the National Science Foundation (NSF) CAREER Award under No. IIS- 1353606, the Office of Naval Research (ONR) under awards No. N000141210041 and No. N000141310129, a Sloan Research Fellowship, the Moore Foundation, an Okawa Research Grant, Toshiba, and Intel. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, NSF, ONR, or the U.S. government.