Empirical Study of the Benefits of Overparameterization in Learning Latent Variable Models

Rares-Darius Buhai[1], Yoni Halpern[2], Yoon Kim[3], Andrej Risteski[4], David Sontag[1]

[1]MIT, [2]Google, [3]Harvard, [4]CMU

Overparameterization

= training a **larger model** than necessary

Supervised learning: easier optimization, often without sacrificing generalization**.**

→ **practice:** [Zhang et al., 2016] commonly used neural networks are so large that they can learn randomized labels.

→ **theory:** [Allen-Zhu et al., 2018; Allen-Zhu et al., 2019] overparameterized neural networks provably learn and generalize for certain classes of functions.

Overparameterization in unsupervised learning

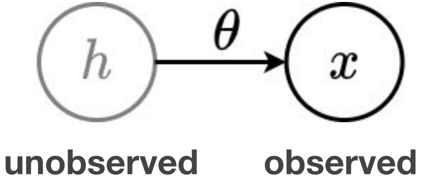
Task: learning **latent variable models**.

Contribution: Empirical study of the benefits of overparameterization in learning latent variable models.

Latent variable models

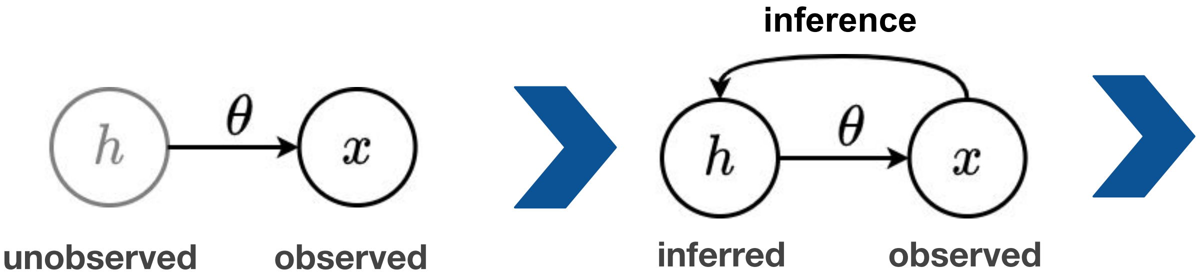
Know $p(x, h; \theta)$.

Task: learn θ.



Maximum likelihood: $\max_\theta p(x; \theta)$. Typically **intractable**.

Iterative algorithms (e.g., EM, variational learning).



$\max_\theta p(x, h; \theta)$

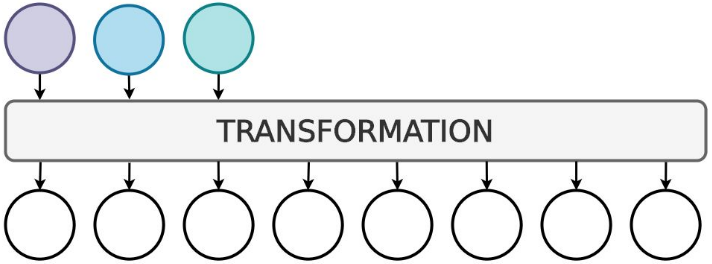
(typically gradient step)

Our setting

Ground truth model.

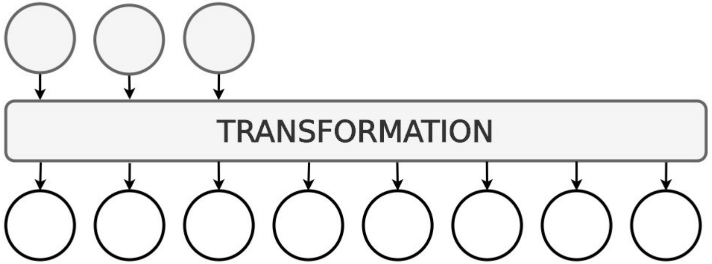
(synthetic setting)

latent variables

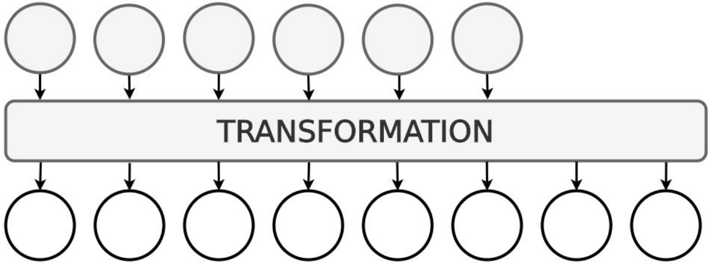


observed variables

Task: learn model from samples.



non-overparameterized



overparameterized

Our question

A ground truth latent variable is **recovered** if there exists a learned latent variable with the same parameters.

How does **overparameterization** affect the **recovery of ground truth latent variables**?

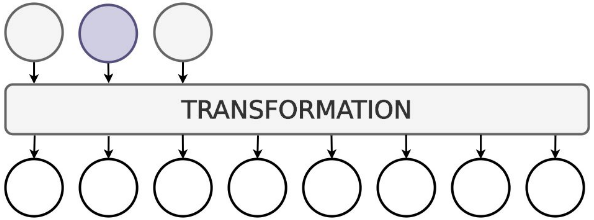
Our finding

With **overparameterization**, the learned model recovers the ground truth latent variables **more often** than without overparameterization.

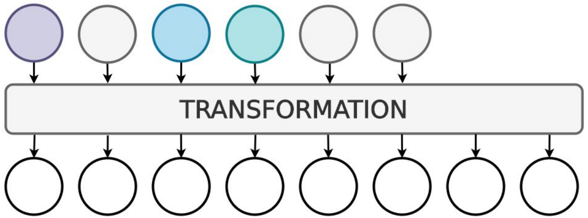
The unmatched learned latent variables are typically redundant.

Demonstration through extensive experiments with:

- **noisy-OR network models**
- sparse coding models
- neural PCFG models

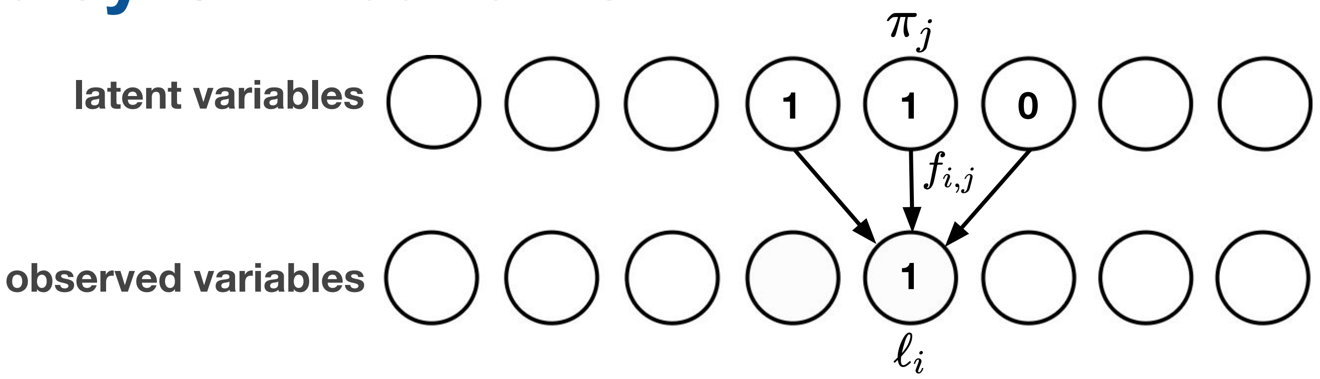


non-overparameterized

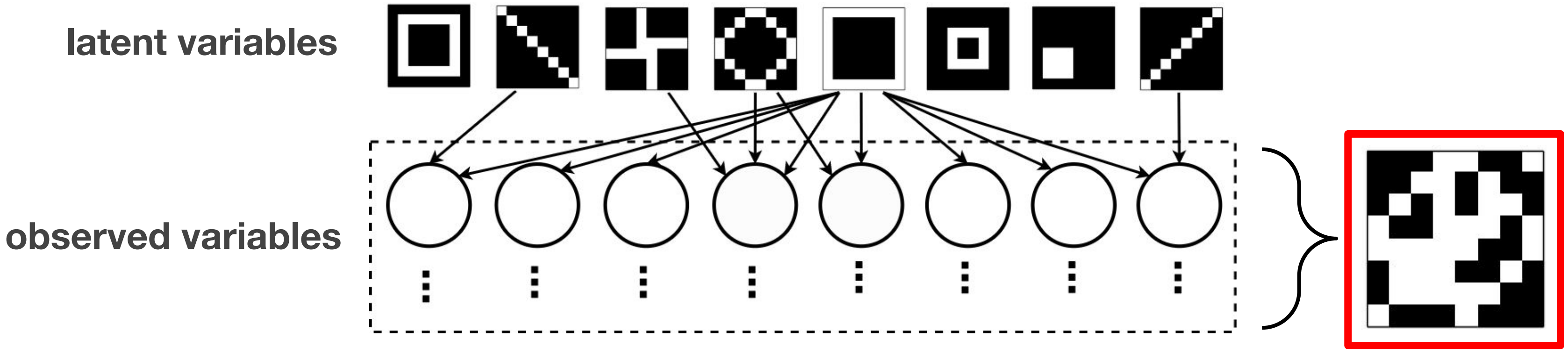


overparameterized

Noisy-OR networks



Example: image model.



Noisy-OR networks

Train using **variational learning**.

noisy-OR network $p(x, h; \theta)$

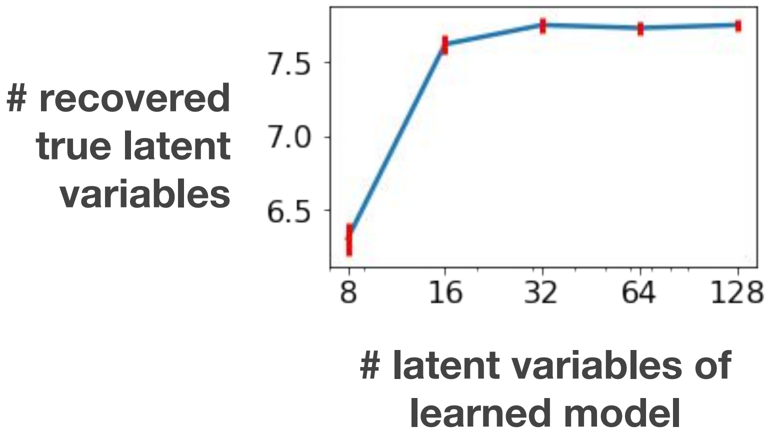
recognition network $q(h|x; \phi)$

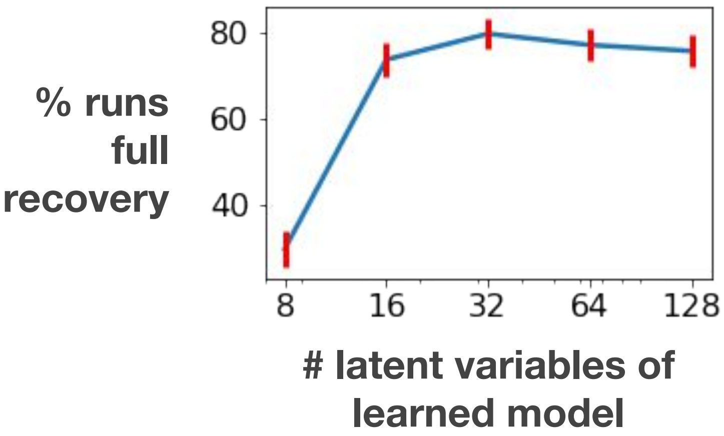
(in our experiments: logistic regression and independent Bernoulli)

Maximize the evidence lower bound (ELBO), alternating between gradient steps w.r.t θ and ϕ.

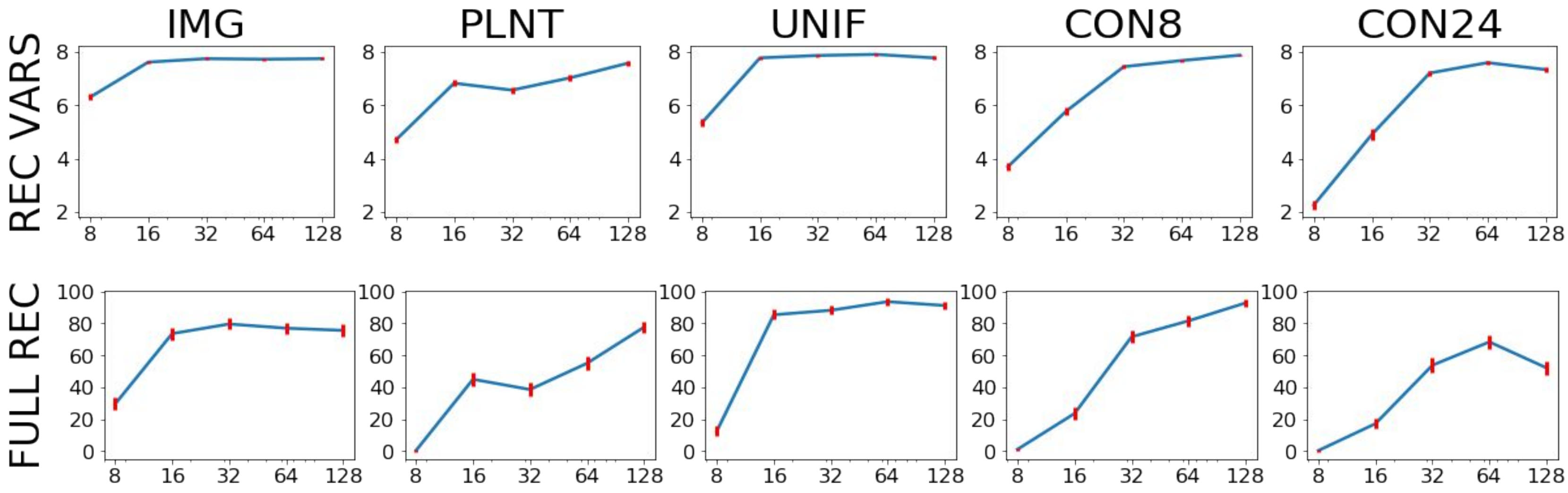
Noisy-OR networks: recovery

Image model.





Noisy-OR networks: recovery



Harm of **extreme overparameterization** is **minor**.

Similar trends for held-out log-likelihood.

Noisy-OR networks: unmatched latent variables

discarded or **duplicates**

Simple **filtering step** to recover ground truth:

- eliminate latent variables with low prior or high failure
- eliminate latent variables that are duplicates

Noisy-OR networks: algorithm variations

Overparameterization remains beneficial:

- **batch size**: 20 → 1000
- **recognition network**: logistic regression → independent Bernoulli

Suggests **benefits are general** when learning latent variable models with iterative algorithms.

Noisy-OR networks: explanation

Hypothesis

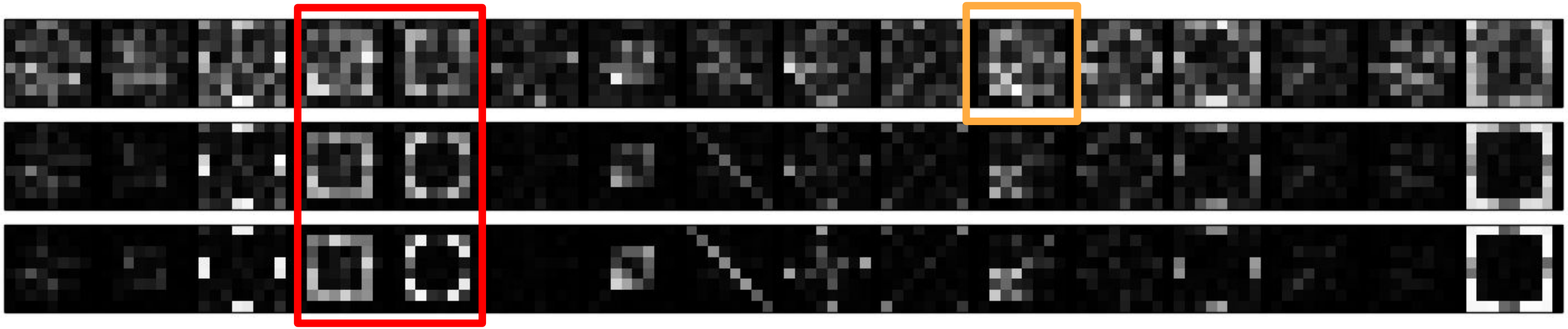
With overparameterization, more latent variables **initialized close to ground truth latent variables**. Then, the benefit is due to a “**warm start**”.

Actual finding

Latent variables do not converge quickly to ground truth latent variables. In the beginning, **undecided**. Throughout, **contentions**.

Noisy-OR networks: optimization stability

State of **latent variables** after 1/9, 2/9, and 3/9 of the first epoch.

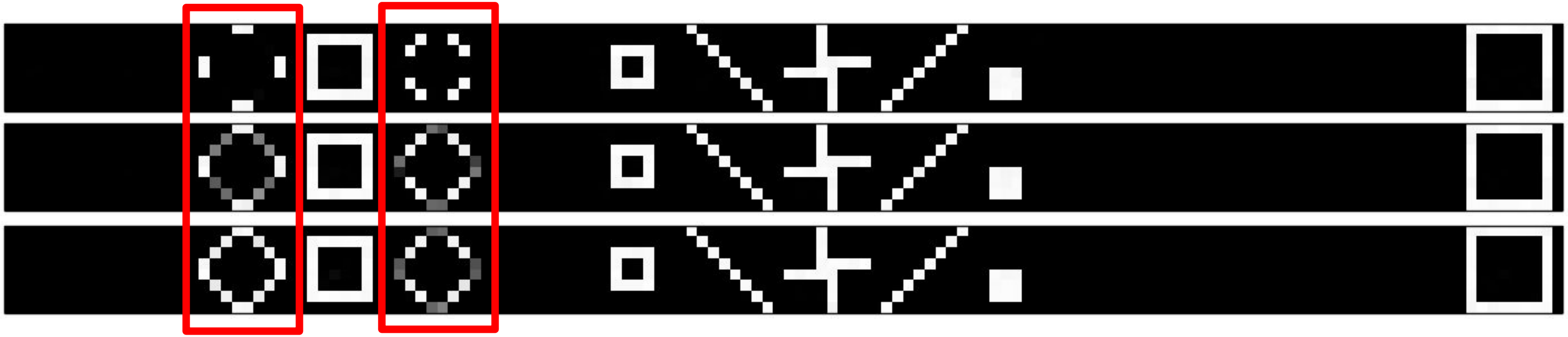


both contend for the same ground truth latent variable

In the beginning, many latent variables are **undecided**.

Noisy-OR networks: optimization stability

State of **latent variables** after 10, 20, and 30 epochs.



both contend for the same ground truth latent variable

Throughout, latent variables often **contend**.

Sparse Coding

Linear model.

Synthetic experiments.

Training with linear alternating minimization algorithm.

→ overparameterization gives better recovery

→ simple filtering step

Neural PCFG

Nonlinear model.

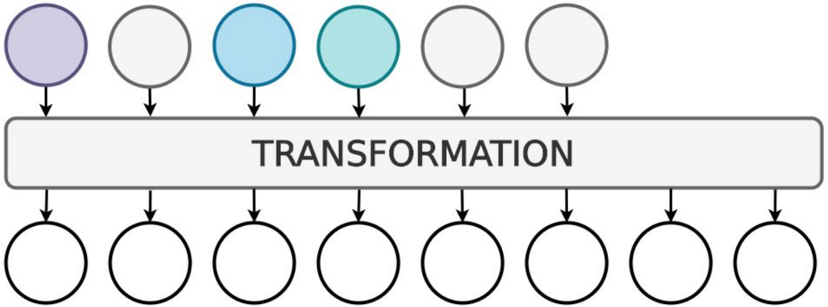
Semi-synthetic experiments with neural network parameterization.

Training with EM and neural network parameterization.

→ overparameterization gives better recovery

(similarity between parse trees)

Discussion



Why is any of this surprising?

Typically, smaller models are more likely to be identifiable.

However, our experiments show that larger models often **make optimization easier** and have an **inductive bias toward ground truth recovery**.

Application

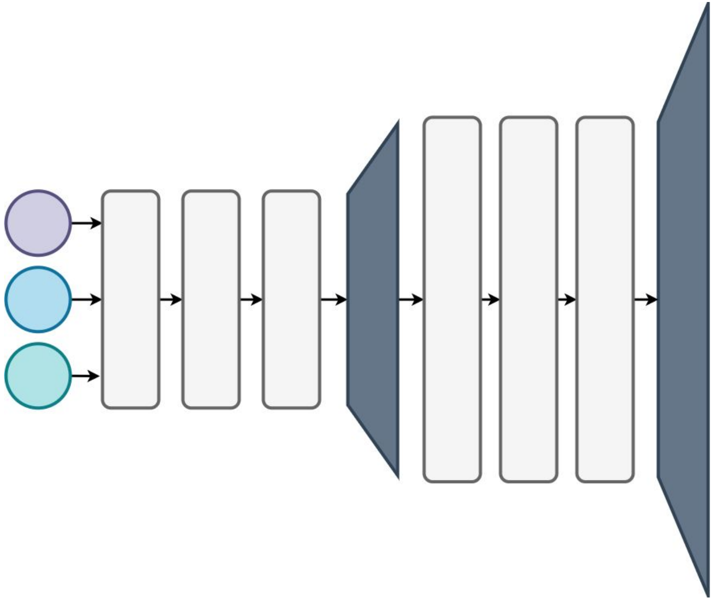
For practice: it is helpful to overparameterize.

For theory: interesting phenomenon, may provide insights into learning and optimization.

Future work

Study **larger and more complex models**, e.g., commonly used deep generative models.

- Understand model identifiability.
- Define overparameterization.
- Define ground truth recovery and design filtering steps.



Thank you!

Our code is available at https://github.com/clinicalml/overparam.