Semi-supervised Deep Kernel Learning

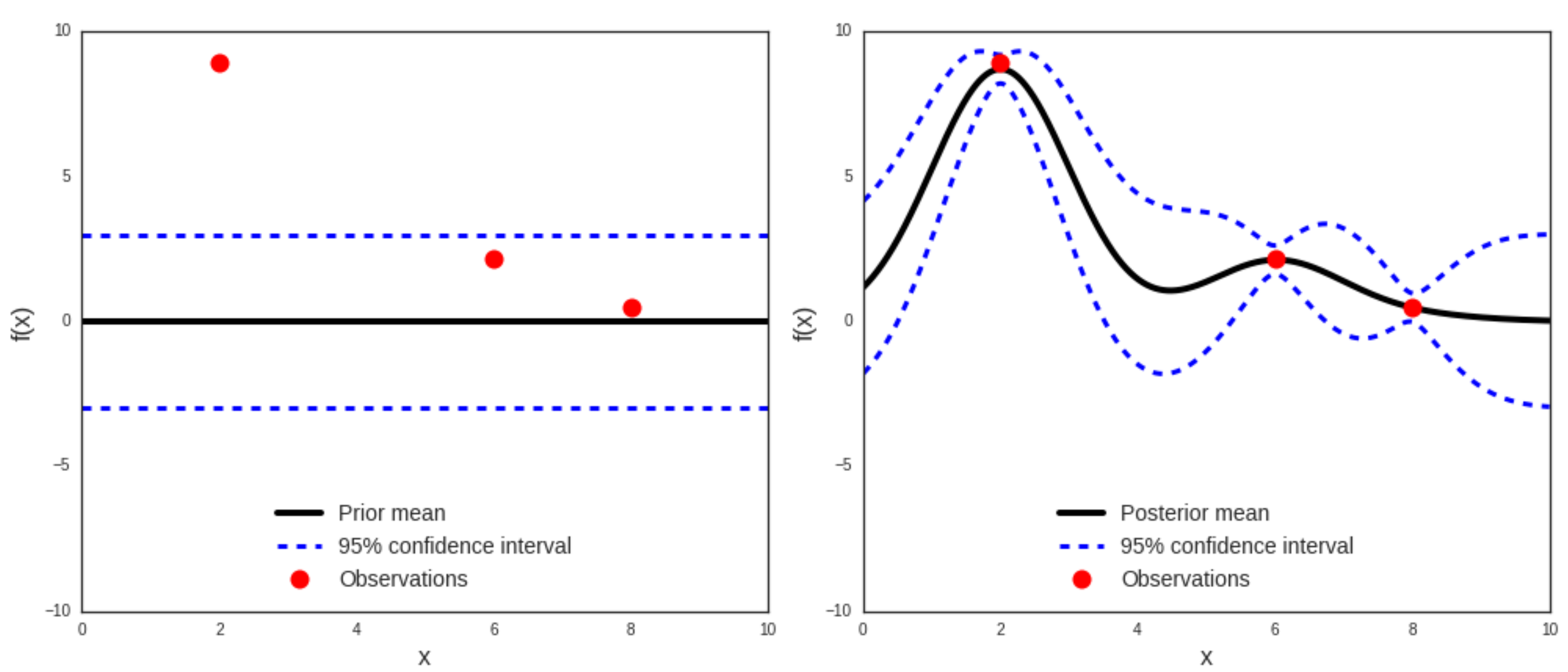
Neal Jean, Michael Xie, Stefano Ermon
{nealjean, xie, ermon}@cs.stanford.edu

Motivation

- Deep neural networks automatically learn good feature representations and have seen great success on a wide variety of supervised learning tasks
- But labeled data is often scarce and/or expensive to obtain – can we learn from unlabeled data?

Approach

- Deep kernel learning (DKL) models combine deep neural networks with Gaussian processes:
  - Neural networks learn adaptive hierarchical feature representations
  - Gaussian processes can quantify the uncertainty in their predictions

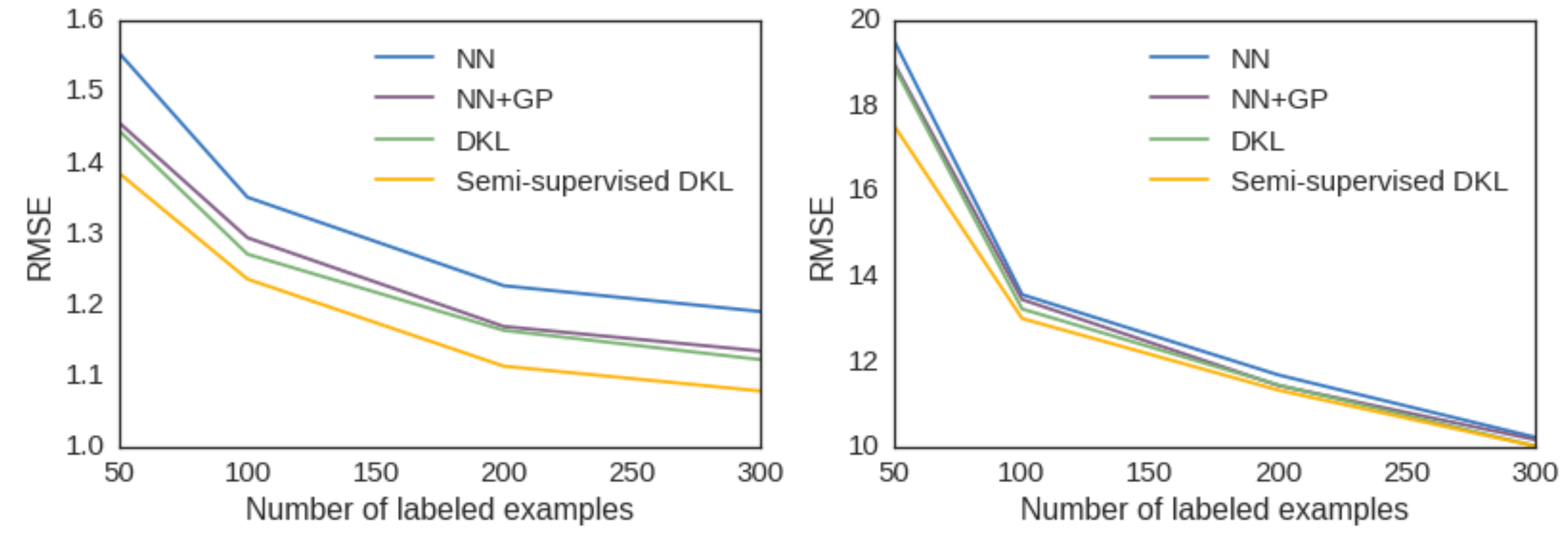


Gaussian process prior (left) and posterior (right)

- Semi-supervised learning can be done by minimizing a compound objective that simultaneously maximizes the log marginal likelihood of labeled data and minimizes the posterior variance of unlabeled data

$$L_{semisup}(\theta,\sigma,\lambda) = \underbrace{-\frac{1}{n}\log p(y \mid X_L,\theta,\sigma,\lambda)}_{\text{labeled}} + \underbrace{\frac{\alpha}{m}\sum_{j:x_j\in X_U}\operatorname{cov}(X_U)_{jj}}_{\text{unlabeled}}$$

UCI Regression Datasets

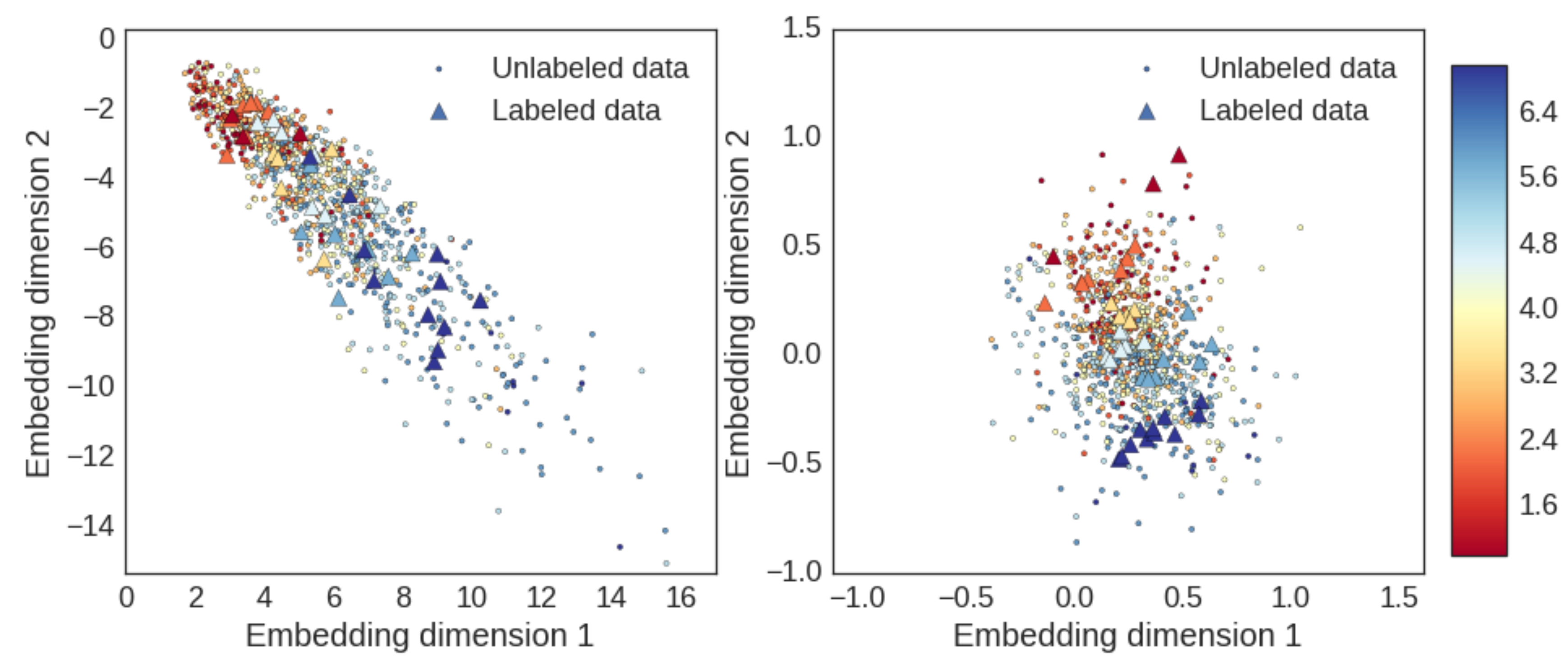


Model comparison on UCI Skillcraft (left) and Ctslice (right) datasets

- Supervised approaches include neural networks (NN), fixed neural networks with a Gaussian process on top (NN + GP), and deep kernel learning models where the NN and GP parameters are learned jointly
- Semi-supervised DKL models that use unlabeled data outperform the supervised models

Intuition

- Neural network learns feature representation in which unlabeled data "look like" labeled examples
- Incorporating abundant unlabeled data has a regularizing effect that prevents overfitting



Two-dimensional embeddings learned by supervised DKL (left) versus embeddings learned by semi-supervised DKL (right)

Poverty Predictions

- **Goal:** Accurately predict poverty and wealth measures from daytime satellite images
- **Challenge:** Ground truth data from surveys is scarce
- This approach allows us to leverage large quantities of unlabeled satellite imagery

	1a	1b	2a	2b	3
Fold	CNN	GP	Fixed CNN+GP	Supervised CNN+GP	Inductive CNN+GP
1	0.762	0.336	0.267	0.263	0.267
2	0.401	0.311	0.219	0.219	0.212
3	0.277	0.341	0.247	0.247	0.244
4	0.636	0.680	0.400	0.429	0.400
5	0.214	0.316	0.262	0.233	0.223
Avg	0.458	0.397	0.279	0.278	**0.269**

Results of 2011 Uganda DHS asset index prediction task

Future work

- Continued experimentation to better understand both the potential and the limitations of the method
- Comparisons with other semi-supervised approaches in different settings
- Explore extension to classification tasks

References

[1] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, "Deep Kernel Learning", *The Journal of Machine Learning Research* (2015).
[2] A. Damianou, N. D. Lawrence, "Deep Gaussian Processes", *The Journal of Machine Learning Research* (2013).
[3] K. Yu, J. Bi, and V. Tresp, "Active Learning via Transductive Experimental Design", *ICML* (2006).
[4] Y. Grandvalet, Y. Bengio, "Semi-supervised Learning by Entropy Minimization", *NIPS* (2004).