Why Does Unsupervised Pre-training Help Deep Discriminant Learning?

Dumitru Erhan*, Yoshua Bengio*, Aaron Courville*,
Pierre-Antoine Manzagol*, Pascal Vincent* & Samy Bengio+

*University of Montreal +Google

The Generative and Discriminative Learning Interface workshop @ NIPS'09 Whistler (BC), December 12th, 2009

Why Does Unsupervised Pre-training Help Deep Discriminant Learning?

Dumitru Erhan[*], Yoshua Bengio[*], Aaron Courville[*],
Pierre-Antoine Manzagol[*], Pascal Vincent[*] & Samy Bengio[+]

[*]University of Montreal [+]Google

The Generative and Discriminative Learning Interface workshop @ NIPS'09 Whistler (BC), December 12th, 2009

Unsupervised pre-training

- Why deep? Brains, ideas, efficiency, statistical strengths.
- < 2006, fully-connected deep networks not popular.
- > 2006, Hinton et al.: use unsupervised pre-training with Restricted Boltzman Machines for initialization.
- It works: vision, NLP, speech, etc.
- Crucial ingredient is unsupervised initialization: RBMs, auto-encoders, even kernel PCAs (Cho & Saul @ NIPS '09).
- Widely applied, but well-understood?

Why does it work so well?

- **Plan**:
   i. propose explanatory hypotheses
   ii. observe the effects of pre-training
   iii. infer its role & level of agreement with our hypotheses.
- **Regularization** hypothesis:
   - Unsupervised component constrains the network to model P(x)
   - P(x) representations good for P(y|x).
- **Optimization** hypothesis:
   - Unsupervised initialization near better local minimum of P(y|x)
   - Reach lower local minimum not achievable by random initialization.

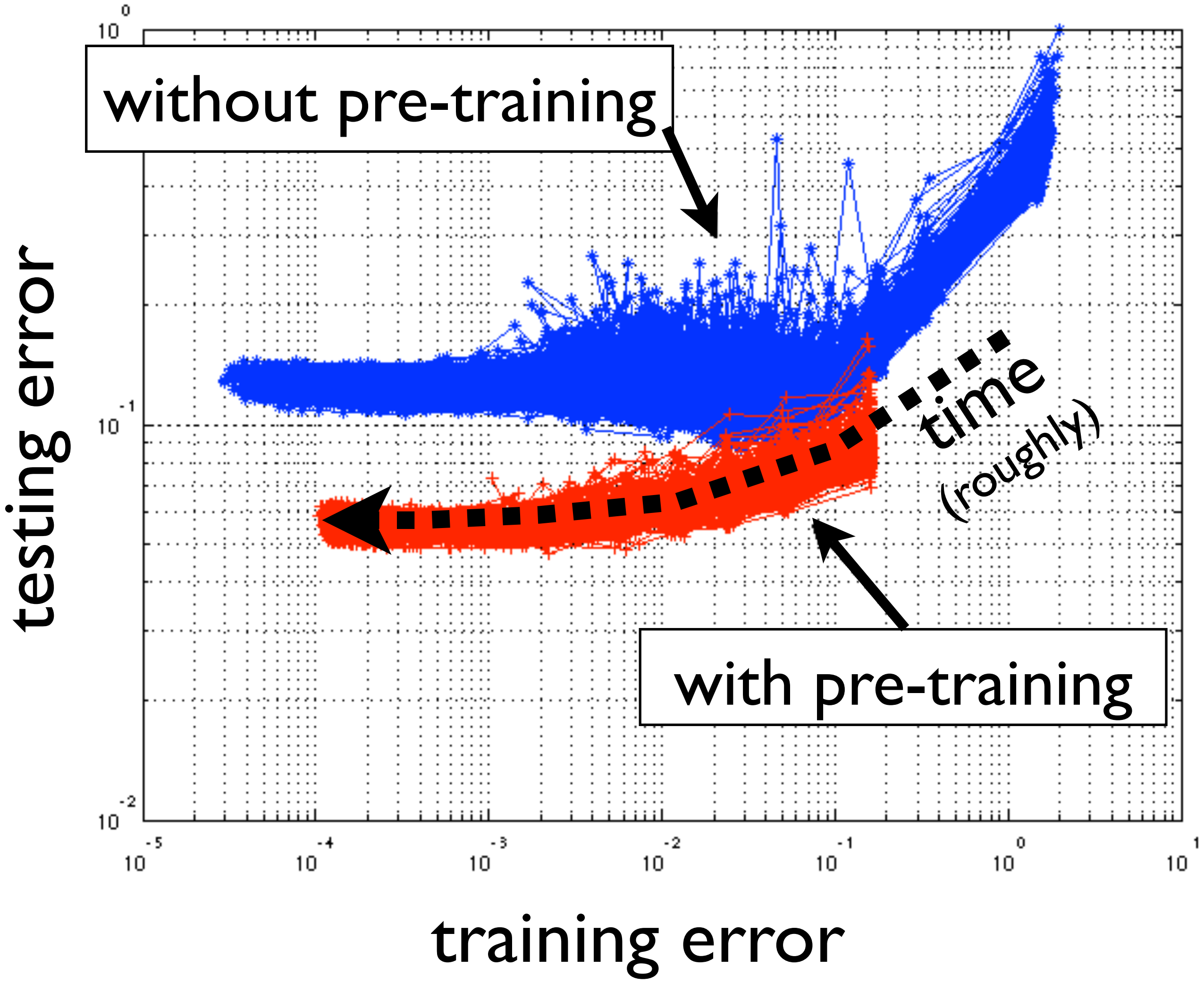
Errors over time

- Pre-training = better generalization for the same training error

- Worse training error, even at the end

- A **regularization** interpretation fits well.

50 3-layer networks
same hyper-parameters, different seeds

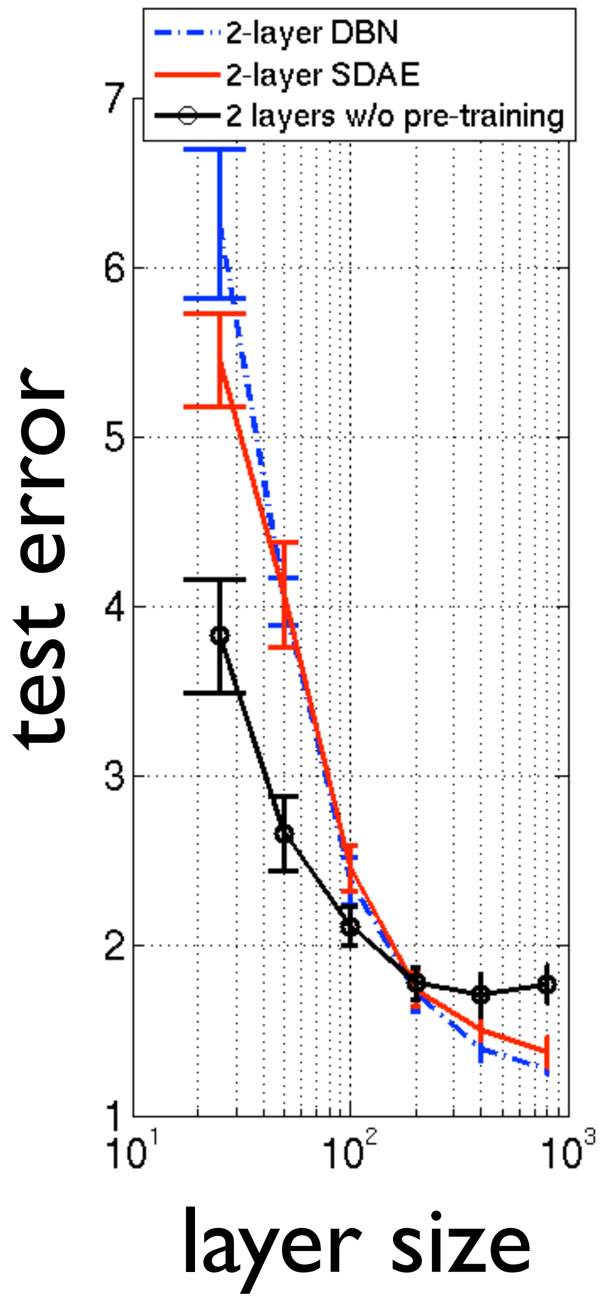


Varying the layer size

- Pre-training + small layer size = *worse* than randomly initialized nets

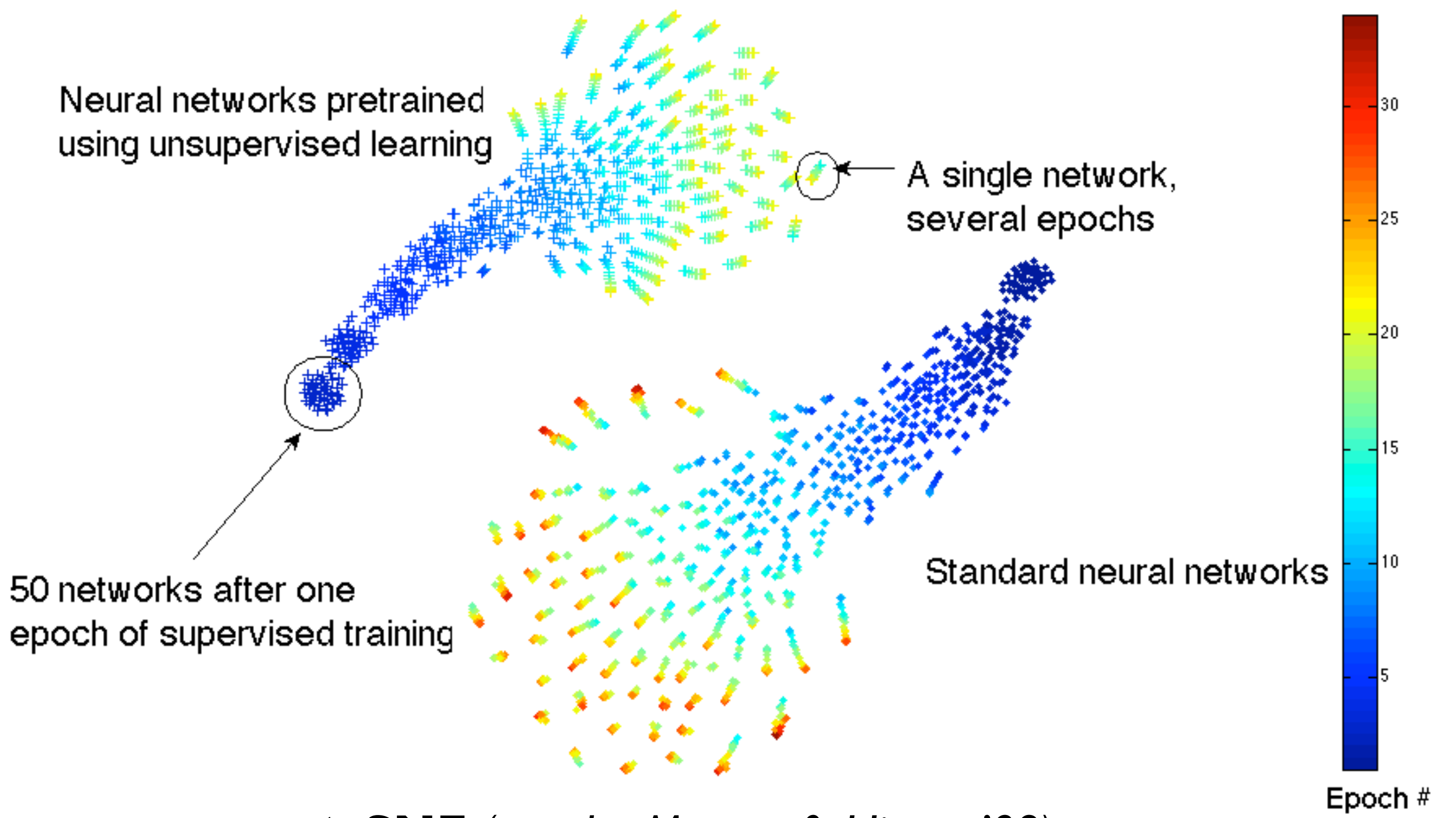
- Additional capacity argument

- Supports a **regularization** explanation.

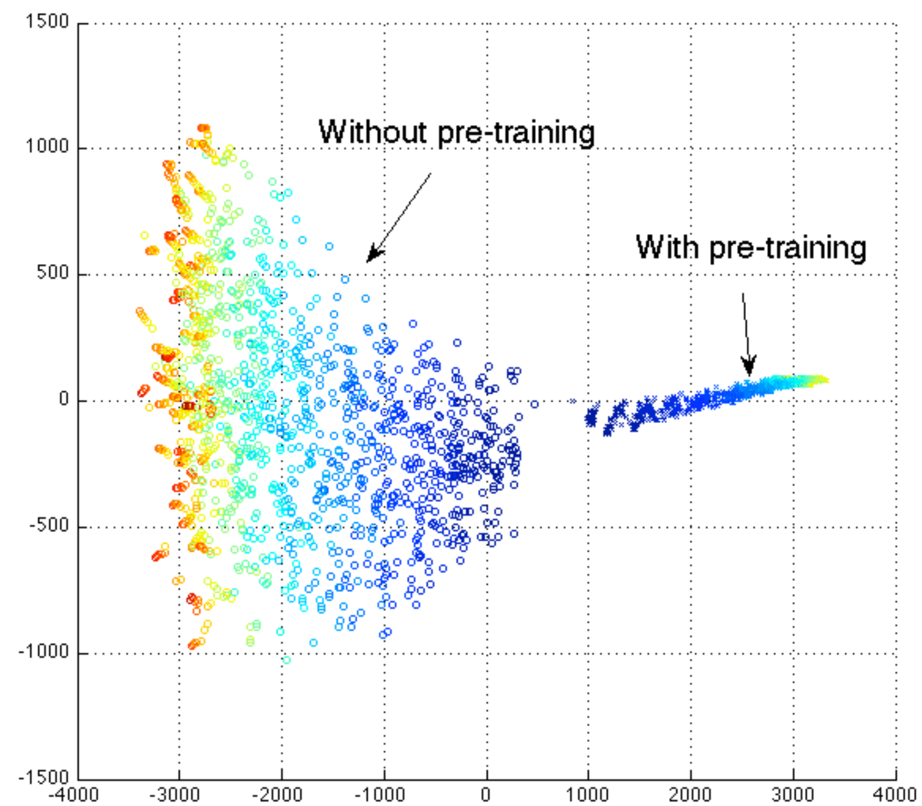


Trajectories in function space

Projecting network outputs (number of test examples x number of top layer units) into 2D:



t-SNE *(van der Maaten & Hinton '08)*



Isomap *(Tenenbaum, de Silva & Langford '00)*

Many apparent local minima

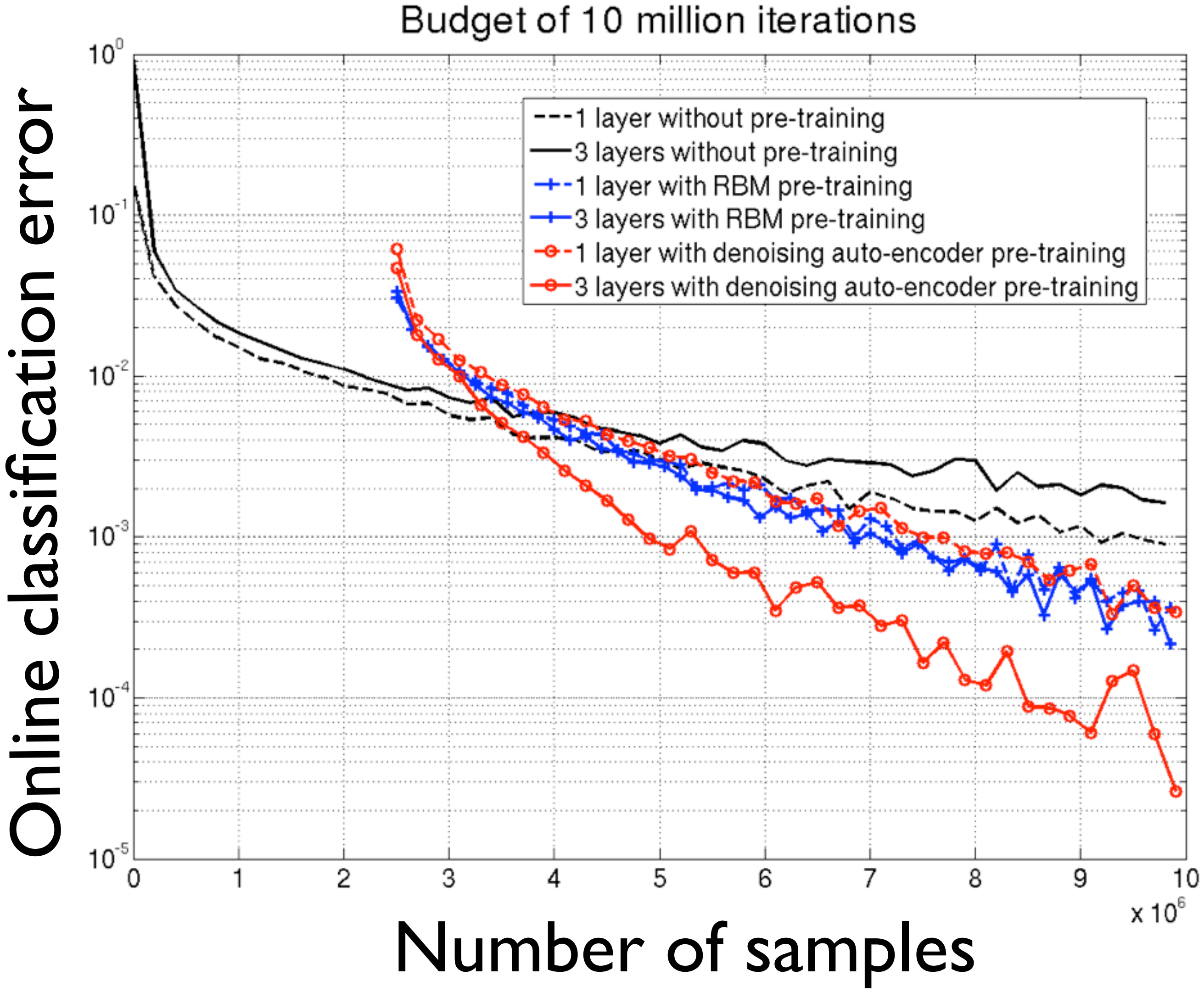
Disjoint regions of space

The role of pre-training

- Pre-training places the networks in a region of the parameter space that is very different from the one given by random initialization.
- Effect of a unique kind of regularizer: one that restricts and influences positively the starting point of supervised optimization.
- Will the pre-training effect disappear in a large-scale (online) learning scenario?

The online learning scenario

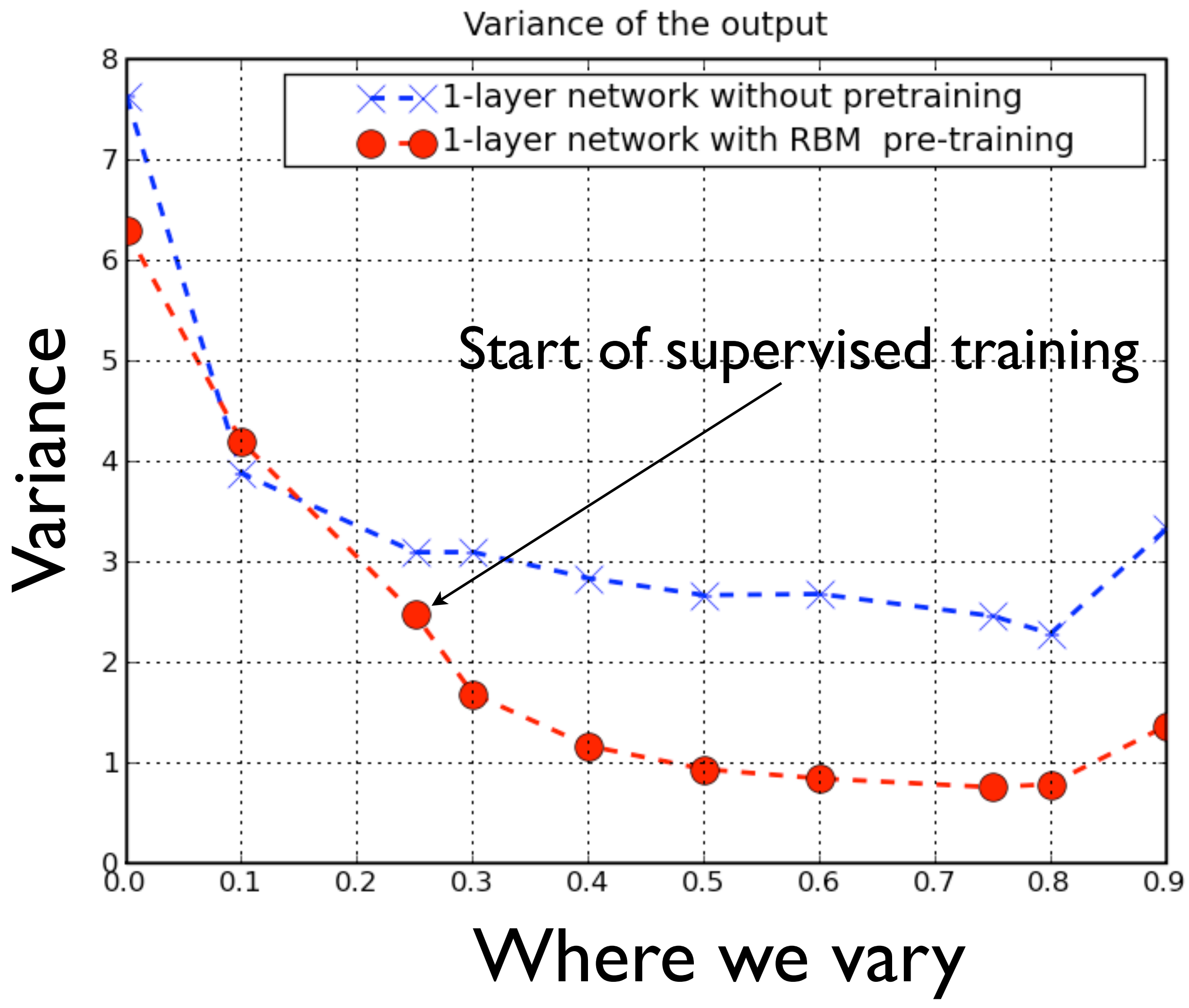
- 10 million examples; (smoothed) online error.
- Pre-training advantage **does not vanish** as dataset size increases.
- Starting point of non-convex optimization clearly **matters,** even in a scenario with essentially unbounded training data.



Surprising as it shows that pre-training does not follow the standard interpretation of a regularizer.

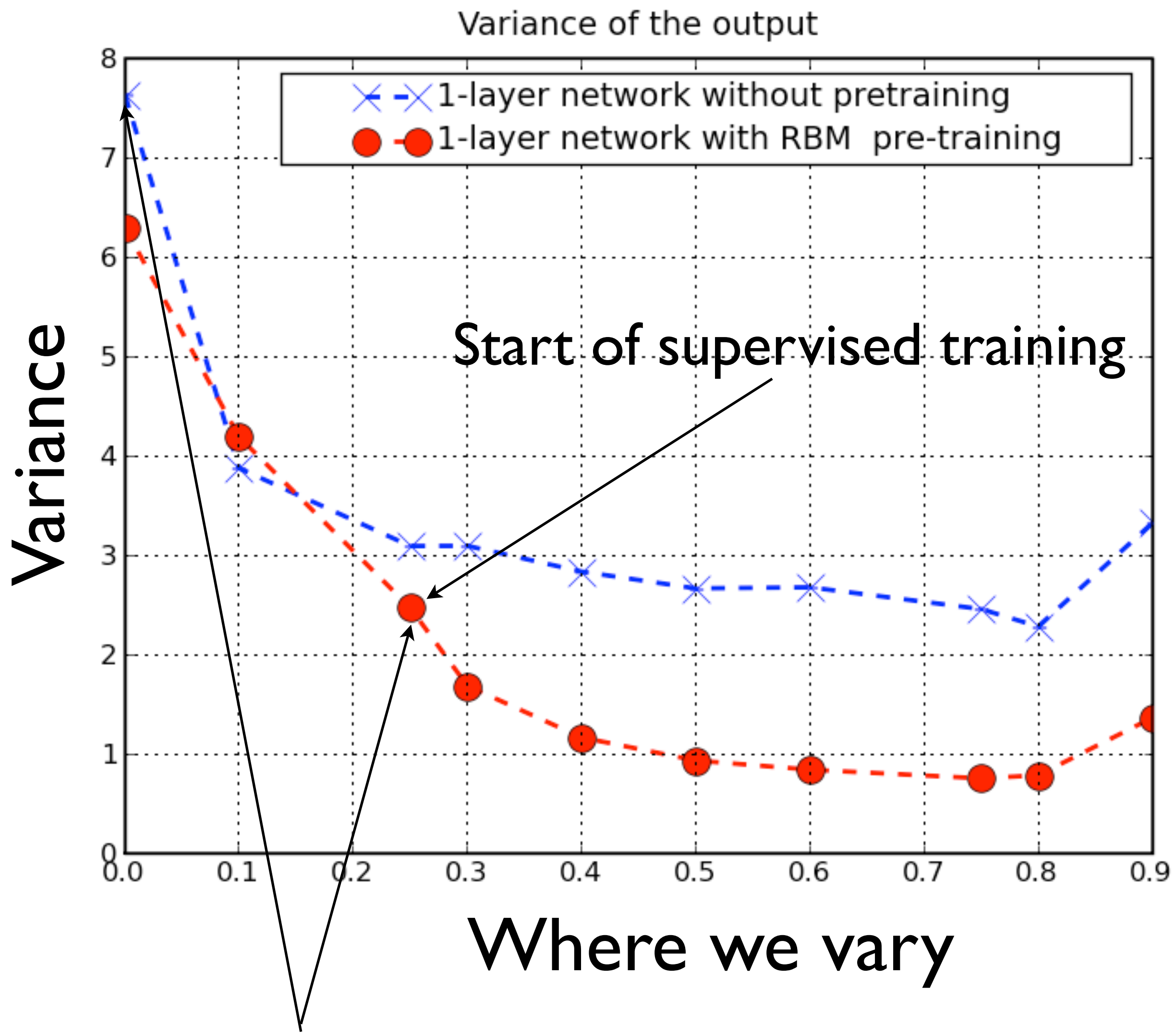
Effect of example ordering

- Online, stochastic, non-convex.
- What is the effect of examples seen at different points during training on the outcome?
- Vary only the 1st one million examples, only the 2nd million, etc.
- Measure the **variance of the output at the end of training** on a fixed test set:
  - Early examples influence more
  - Pre-training = variance reduction



Effect of example ordering

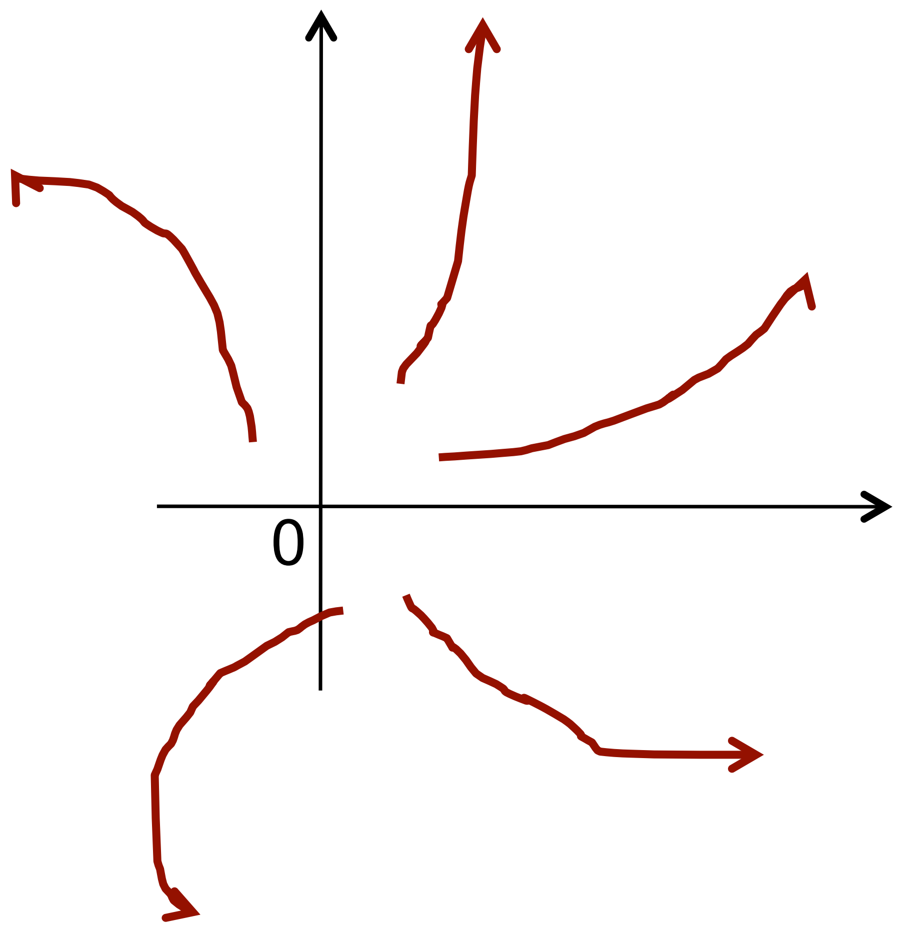
- Online, stochastic, non-convex.
- What is the effect of examples seen at different points during training on the outcome?
- Vary only the 1st one million examples, only the 2nd million, etc.
- Measure the **variance of the output at the end of training** on a fixed test set:
  - Early examples influence more
  - Pre-training = variance reduction



Variance at the onset of supervised training is *lower* for pre-trained networks

Dynamics of unsupervised pre-training initialization

- As weights become larger, they get trapped in a basin of attraction ("quadrant" does not change)
- Initial updates have a crucial influence ("critical period"), explain more of the variance
- Unsupervised pre-training initializes in basin of attraction with good generalization properties



Discussion & take-home

- Early results had pointed towards a regularization hypothesis; we suggest a more nuanced interpretation.
- Explored the online setting and found surprising results: pre-training effect *does not vanish*.
- Pre-training: *variance reduction technique*.
- Positive effect as long as modelling P(x) is useful for P(y|x).
- Influence of early examples could be troublesome.
- Future: understand other semi-supervised deep approaches.
- More results & discussion in our *upcoming JMLR paper*!

Thank you!

Questions? Comments?