On The Power of Curriculum Learning In Training Deep Networks

Daphna Weinshall

School of Computer Science and Engineering
The Hebrew University of Jerusalem

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

Not my first Jigsaw puzzle

My first Jigsaw puzzle

LEARNING COGNITIVE TASKS (CURRICULUM):

Not my first chair

LEARNING ABOUT OBJECTS' APPEARANCE







Avrahami et al. Teaching by examples: Implications for the process of category acquisition. The Quarterly Journal of Experimental Psychology: Section A, 50(3): 586–606, 1997

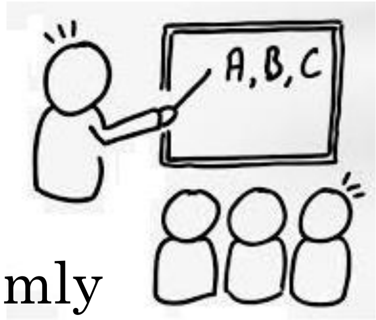
SUPERVISED MACHINE LEARNING

- Data is sampled randomly
- We expect the train and test data to be sampled from the same distribution

- Exceptions:
  - Boosting
  - Active learning
  - Hard data mining

but these methods focus on the more difficult examples…

CURRICULUM LEARNING



- **Curriculum Learning (CL)**: instead of randomly selecting training points, select easier examples first, slowly exposing the more difficult examples from easiest to the most difficult

- **Previous work**: empirical evidence (only), with mostly simple classifiers or sequential tasks

  ⇨ CL speeds up learning and improves final performance

- Q: since curriculum learning is intuitively a good idea, why is it rarely used in practice in machine learning?

A?: maybe because it requires additional labeling…

Our contribution: curriculum by-transfer & by-bootstrapping

Previous empirical work: deep learning

- (Bengio et al, 2009): setup of paradigm, object recognition of geometric shapes using a perceptron; *difficulty is determined by user from geometric shape*

- (Zaremba 2014): LSTMs used to evaluate short computer programs; *difficulty is automatically evaluated from data – nesting level of program*.
- (Amodei et al, 2016): End-to-end speech recognition in english and mandarin; *difficulty is automatically evaluated from utterance length.*
- (Jesson et al, 2017): deep learning segmentation and detection; *human teacher (user/programmer) determins difficulty.*

OUTLINE

1. Empirical study: curriculum learning in deep networks
   - Source of supervision: by-transfer, by-bootstrapping
   - Benefits: speeds up learning, improves generalization

2. Theoretical analysis: 2 simple convex loss functions, linear regression and binary classification by hinge loss minimization
   - Definition of “difficulty”
   - Main result: faster convergence to global minimum

3. Theoretical analysis: general effect on optimization landscape
   - optimization function gets steeper
   - global minimum, which induces the curriculum, remains the/a global minimum

⇨ theoretical results vs. empirical results, some surprises

Definitions

- *Ideal Difficulty Score (IDS)*: the loss of a point with respect to the optimal hypothesis $L(X,h_{opt})$

- *Stochastic Curriculum Learning (SCL)*: variation on SGD. The learner is exposed to the data gradually based on the *IDS* of the training points, from the easiest to the most difficult.

- SCL algorithm should solve two problems:
  - Score the training points by difficulty.
  - Define the scheduling procedure – the subsets of the training data (or the highest difficulty score) from which mini-batches are sampled at each time step.

Curriculum learning: algorithm

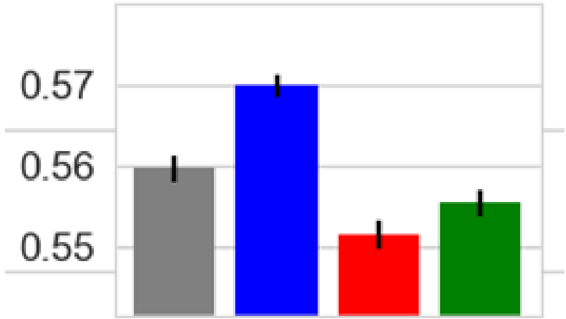
- Data, $\mathbb{X} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$
- Scoring function, $f : \mathbb{X} \rightarrow \mathbb{R}$
- Pacing function, $g_\vartheta : [M] \rightarrow [N]$ ⇨ $\mathbb{X}_1', \ldots, \mathbb{X}_M' \subseteq \mathbb{X}$

Algorithm Curriculum learning method

Input: *pacing function* g_ϑ, *scoring function* f, data $\mathbb{X}$.
Output: sequence of mini-batches $\left[\mathbb{B}_1', \ldots, \mathbb{B}_M'\right]$.
sort $\mathbb{X}$ according to f, in ascending order
$result \leftarrow []$
for all $i = 1, \ldots, M$ **do**
    $size \leftarrow g_\vartheta(i)$
    $\mathbb{X}_i' \leftarrow \mathbb{X}[1, \ldots, size]$
    uniformly sample $\mathbb{B}_i'$ from $\mathbb{X}'$
    append $\mathbb{B}_i'$ to $result$
end for
return $result$

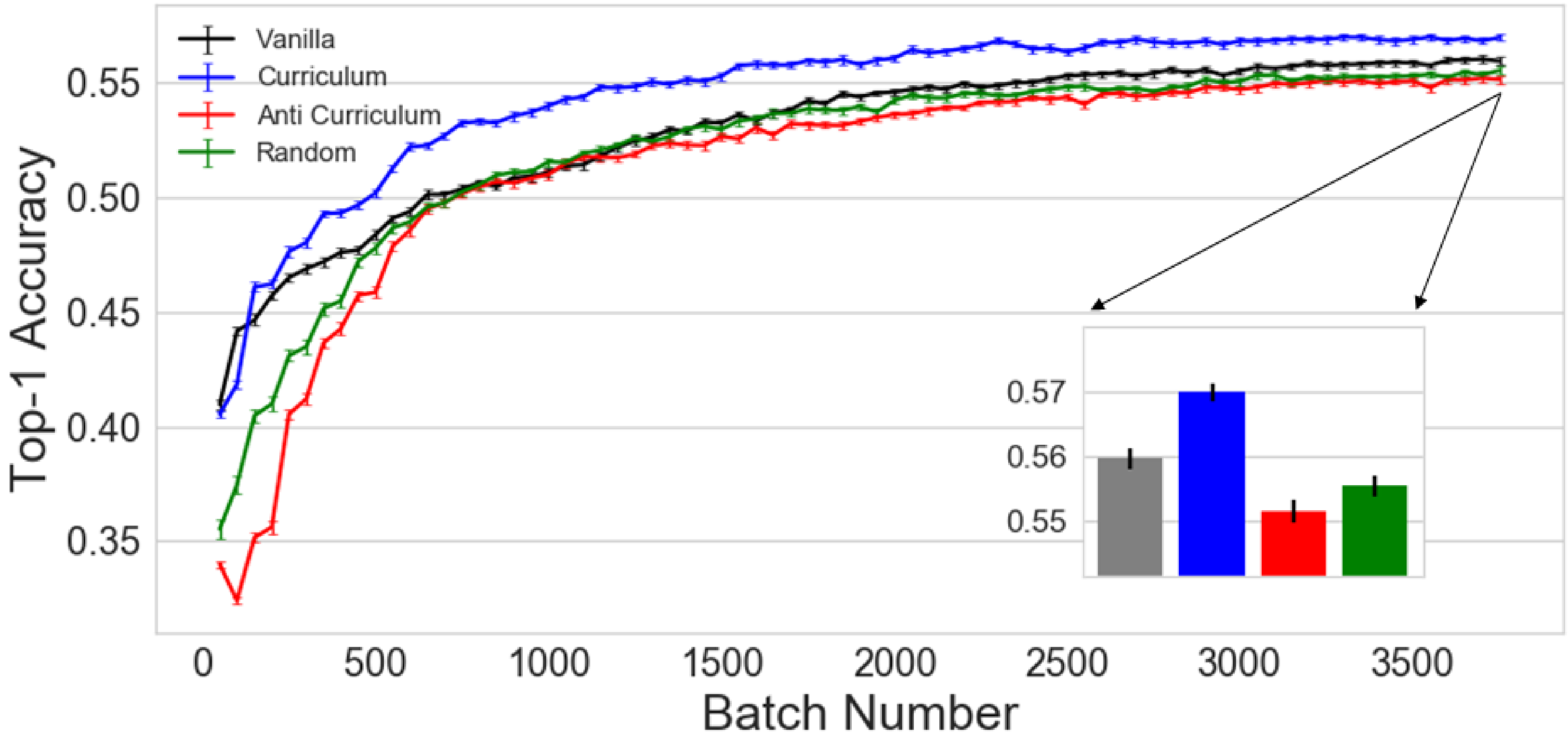
Results

- Vanilla – no curriculum
- Curriculum learning by-transfer
  - Ranking by Inception, a big public domain network pre-trained on ImageNet
  - Similar results with other pre-trained networks
- Basic control conditions
  - Random ranking (benefits from the ordering protocol per se)
  - Anti-curriculum (ranking from most difficult to easiest)



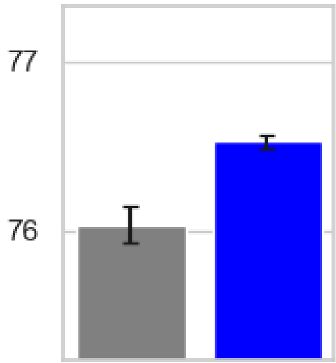
RESULTS: LEARNING CURVE

Subset of CIFAR-100, with 5 sub-classes

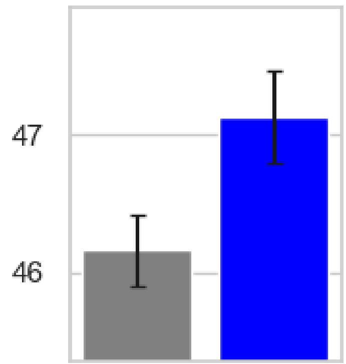


RESULTS: DIFFERENT ARCHITECTURES AND DATASETS, TRANSFER CURRICULUM ALWAYS HELPS

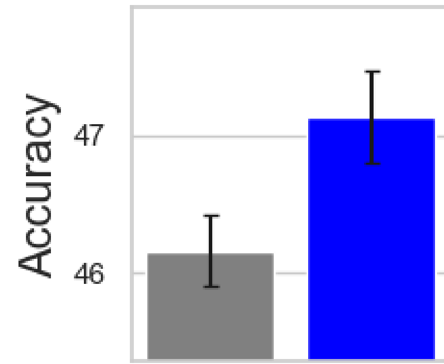
Small CNN trained from scratch



CIFAR-10

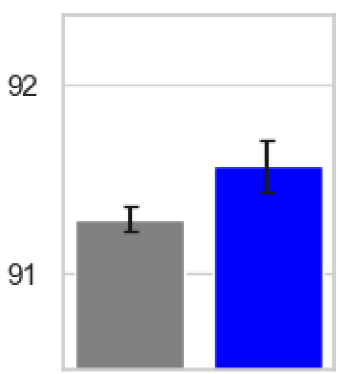


CIFAR-100

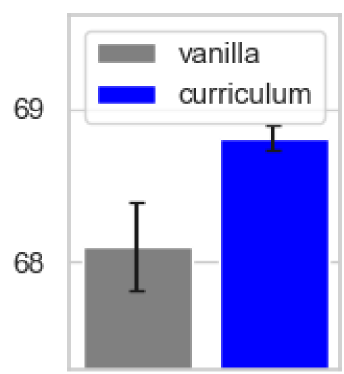


cats (from imagenet)

Pre-trained competitive VGG



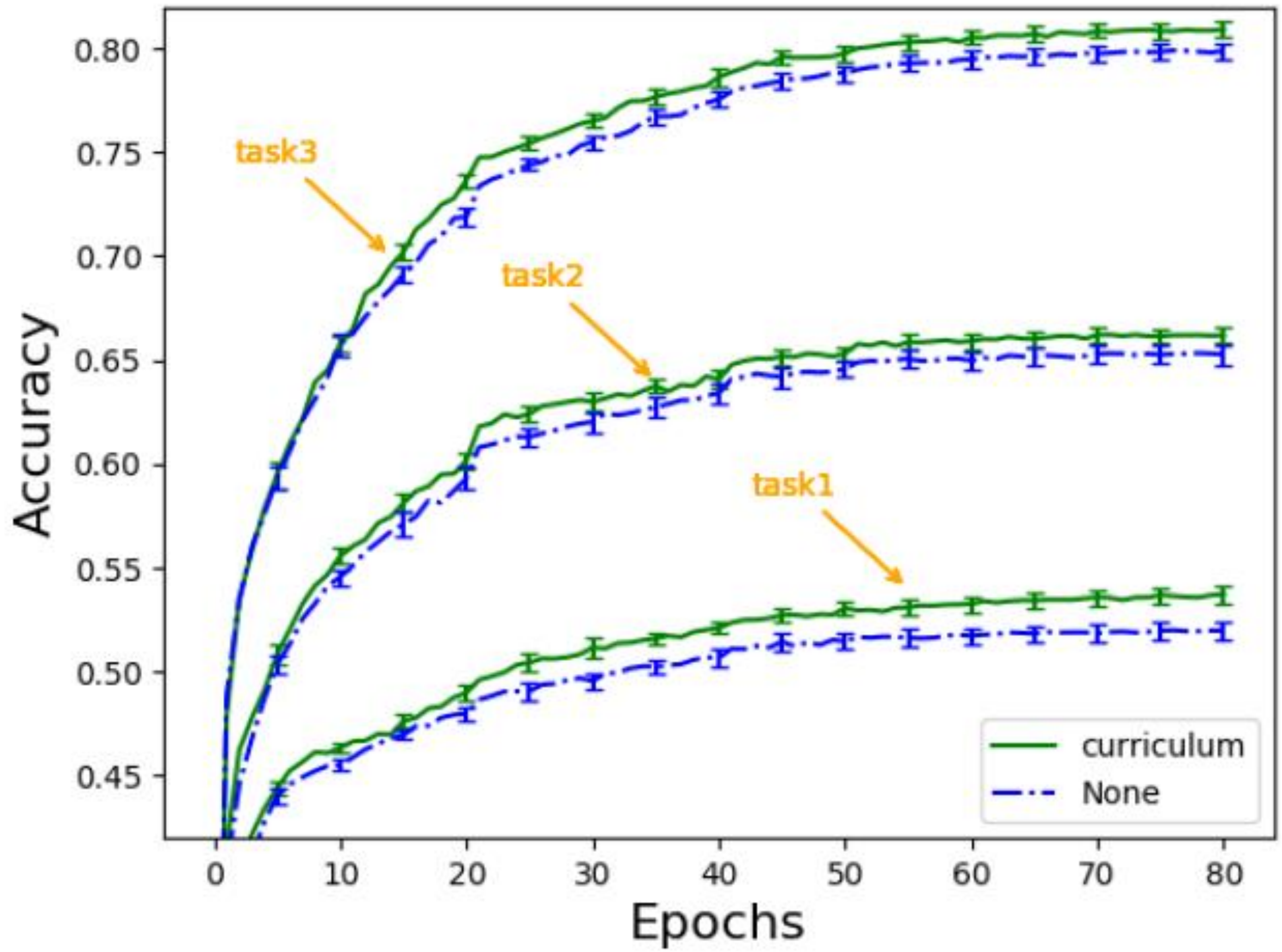
CIFAR-10



CIFAR-100

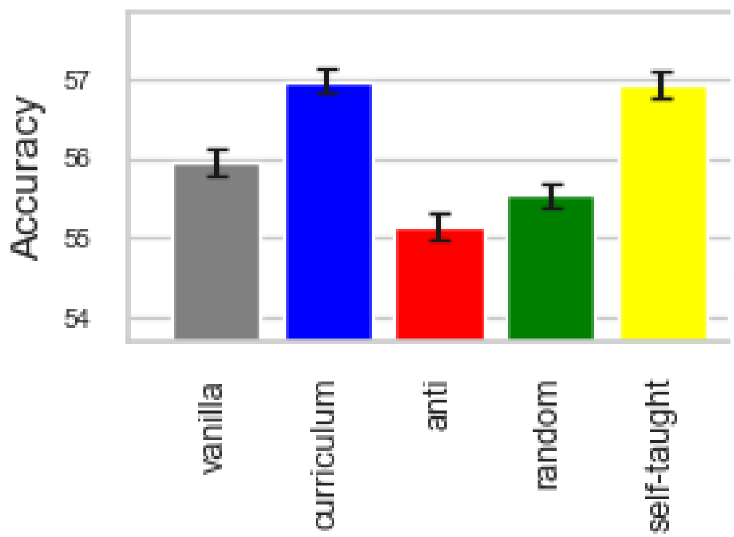
Curriculum helps more for harder problems

3 subsets of CIFAR-100, which differ by difficulty



Additional results

- Curriculum learning by-bootstrapping
  - Train current network (vanilla protocol)
  - Rank training data by final loss using trained network
  - Re-train network from scratch with CL



Outline

1. Empirical study: curriculum learning in deep networks
   - Source of supervision: by-transfer, by-bootstrapping
   - Benefits: speeds up learning, improves generalization

2. Theoretical analysis: 2 simple convex loss functions, linear regression and binary classification by hinge loss minimization
   - Definition of “difficulty”
   - Main result: faster convergence to global minimum

3. Theoretical analysis: general effect on optimization landscape
   - optimization function gets steeper
   - global minimum, which induces the curriculum, remains the/a global minimum

⇨ theoretical results vs. empirical results, some mysteries

Theoretical analysis: linear regression loss, binary classification & hinge loss minimization

- **Theorem**: convergence rate is monotonically decreasing with the *Difficulty Score* of a point.

- **Theorem**: convergence rate is monotonically increasing with the *loss* of a point with respect to the *current hypothesis*.*

- **Corollary:** expect faster convergence at the beginning of training**.**

* when Difficulty Score is fixed

Definitions

- ERM loss $L_D(h) = \mathbb{E}_{\mathbf{X}_t \sim \mathcal{D}}(L(\mathbf{X}_t, h))$
- Definition: point difficulty ⇔ loss with respect to optimal hypothesis $\bar{h}$

$$\Psi(\mathbf{X}) = g(L(\mathbf{X}, \bar{h}))$$

- Definition: transient point difficulty ⇔ loss with respect to current hypothesis h_t

$$\Upsilon(\mathbf{X}) = g(L(\mathbf{X}, h_t))$$

- $\lambda = \|\bar{h} - h_t\|_2$ $\qquad \lambda_t = \|\bar{h} - h_{t+1}\|_2 = f(x)$
- $\Delta(\Psi, \Upsilon) = E[\lambda^2 - \lambda_t^2]$

Theoretical analysis: linear regression loss

- **Theorem**: convergence rate is monotonically decreasing with the *Difficulty Score* of a point Ψ

  Proof: $$\frac{\partial \Delta(\Psi)}{\partial \Psi} \leq 0$$

- **Theorem**: convergence rate is monotonically increasing with the *loss* of a point with respect to the *current hypothesis* Υ

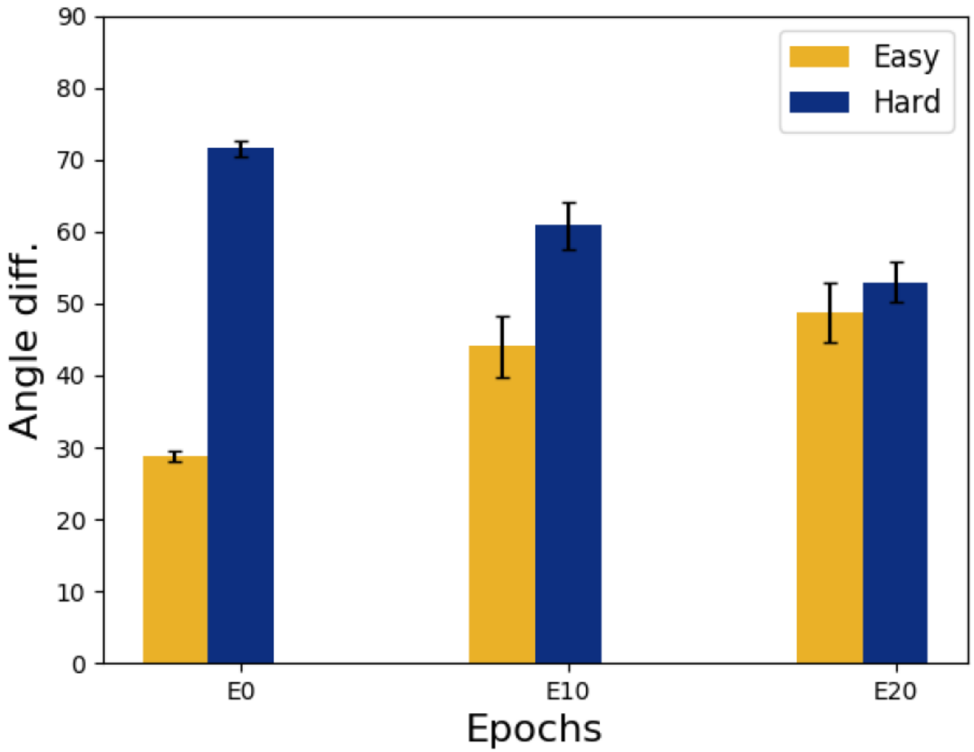
  Proof: $$\frac{\partial \Delta(\Psi_0, \Upsilon)}{\partial \Upsilon} + O(\eta^2) \geq 0 \qquad \forall \Psi_0$$

- **Corollary:** expect faster convergence at the beginning of training (only true for regression loss)

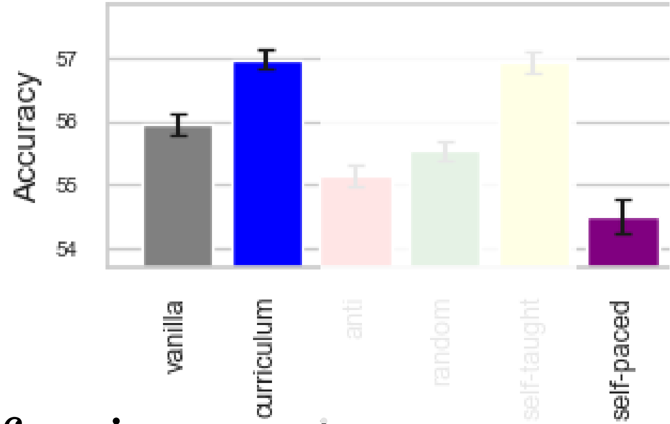
  Proof: $$\frac{\partial \Delta(\Psi)}{\partial \lambda} \geq 0 \qquad \text{when} \quad \eta \leq \frac{\mathbb{E}[r^2 \cos^2 \vartheta]}{\mathbb{E}[r^4 \cos^2 \vartheta]}$$

Matching Empirical results

- Setup: image recognition with deep CNN
- Still, average distance of gradients from optimal direction shows agreement with Theorem 1 and its corollaries



SELF-PACED LEARNING

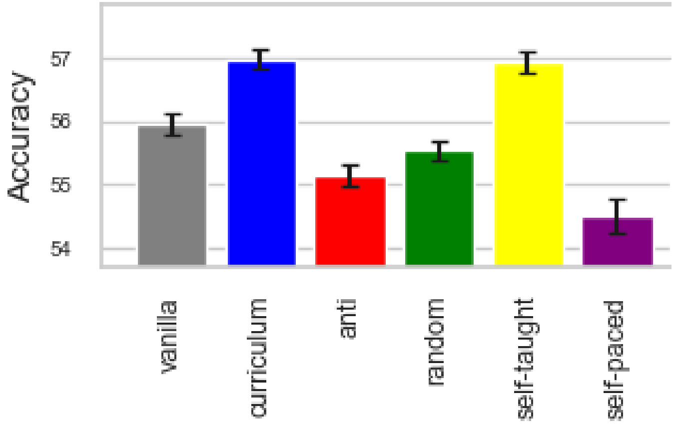


- Self-paced is similar to CL, preferring easier examples, but ranking is based on loss with respect to the current hypothesis (not optimal)

- The 2 theorems imply that one should prefer easier points with respect to the optimal hypothesis, and more difficult points with respect to the current hypothesis

⇨ Prediction: self-paced learning should decrease performance

All conditions



- *Vanilla*: no curriculum
- *Curriculum*: transfer, ranking by inception
- Controls:
  - anti-curriculum
  - random
- *Self taught*: bootstrapping curriculum:
  - training data sorted after vanilla training
  - subsequently, re-training from scratch with curriculum
- Self-Paced Learning: ranking based on local hypothesis

OUTLINE

1. Empirical study: curriculum learning in deep networks
   - Source of supervision: by-transfer, by-bootstrapping
   - Benefits: speeds up learning, improves generalization

2. Theoretical analysis: 2 simple convex loss functions, linear regression and binary classification by hinge loss minimization
   - Definition of "difficulty"
   - Main result: faster convergence to global minimum

3. Theoretical analysis: general effect on optimization landscape
   - optimization function gets steeper
   - global minimum, which induces the curriculum, remains the/a global minimum

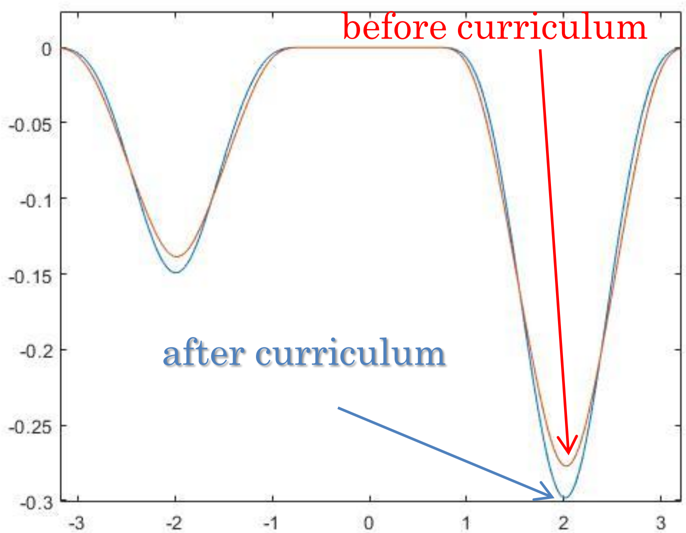
⇨ theoretical results vs. empirical results, some mysteries

Effect of CL on optimization landscape

- Corollary 1: with an ideal curriculum, under very mild conditions, the modified optimization landscape has the same global minimum as the original one

- Corollary 2: when using any curriculum which is positively correlated with the ideal curriculum, gradients in the modified landscape are steeper than the original one

optimization function



Theoretical analysis: optimization landscape

Definitions:

- ERM optimization: $\mathcal{L}(\vartheta) = \hat{\mathbb{E}}[L_\vartheta] = \frac{1}{N}\sum^{N} L_\vartheta(X_i)$

$$\tilde{\vartheta} = \arg\min_{\vartheta} \mathcal{L}(\vartheta) = \arg\max_{\vartheta} \prod_{i=1}^{N} e^{-L_\vartheta(X_i)}$$

- Empirical Utility/Gain Maximization:

$$\mathcal{U}(\vartheta) = \hat{\mathbb{E}}[U_\vartheta] = \frac{1}{N}\sum_{i=1}^{N} U_\vartheta(X_i) \triangleq \frac{1}{N}\sum_{i=1}^{N} e^{-L_\vartheta(X_i)}$$

- Curriculum learning:

$$\mathcal{V}(\vartheta) = \hat{\mathbb{E}}_{\boldsymbol{p}}[U_\vartheta] = \sum_{i=1}^{N} U_\vartheta(X_i)p(X_i) = \sum_{i=1}^{N} e^{-L_\vartheta(X_i)}p(X_i)$$

- Ideal curriculum: $p(X_i) = P(X_i|_{\tilde{\vartheta}}) \propto P(\tilde{\vartheta}|_{X_i})$

Some results

For any prior:

$$\mathcal{V}(\vartheta) = \mathcal{U}(\vartheta) + \mathrm{C\hat{o}v}[U_\vartheta, p]$$

For the ideal curriculum:

$$\mathcal{V}(\vartheta) = \mathcal{U}(\vartheta) + \frac{1}{C}\mathrm{Cov}[U_\vartheta, U_{\tilde{\vartheta}}]$$

which implies

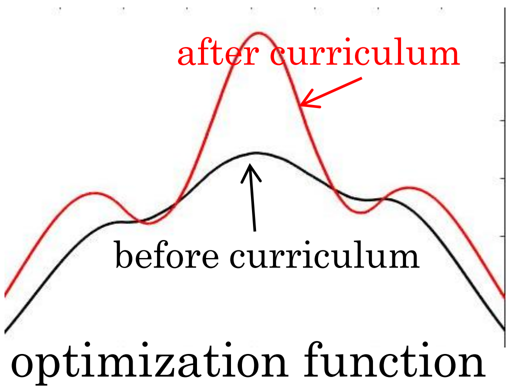
$$\mathcal{V}(\tilde{\vartheta}) - \mathcal{V}(\vartheta) \geq \mathcal{U}(\tilde{\vartheta}) - \mathcal{U}(\vartheta) \quad \forall \vartheta : \mathrm{Cov}[U_\vartheta, U_{\tilde{\vartheta}}] \leq 0$$

and generally

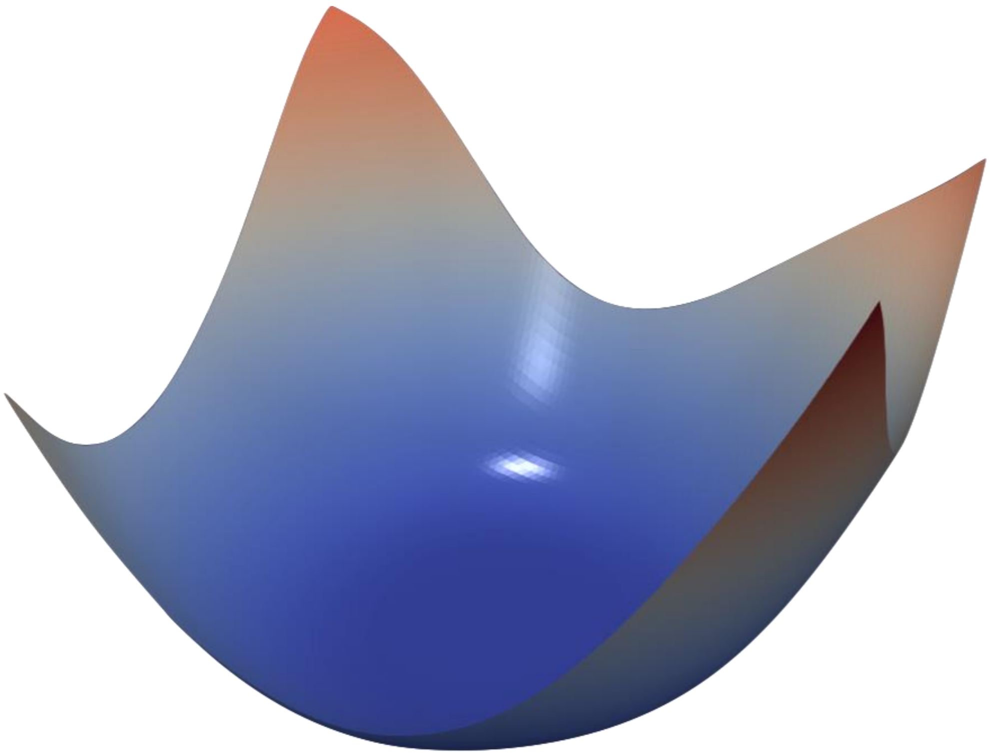
$$\mathcal{V}(\tilde{\vartheta}) - \mathcal{V}(\vartheta) \geq \mathcal{U}(\tilde{\vartheta}) - \mathcal{U}(\vartheta) \quad \forall \vartheta : \mathrm{Cov}[U_\vartheta, U_{\tilde{\vartheta}}] \leq \mathrm{Var}[U_{\tilde{\vartheta}}]$$

REMAINING UNCLEAR ISSUES, WHEN MATCHING THE THEORETICAL AND EMPIRICAL RESULTS…

Empirical findings	Theoretical results
CL steers optimization to better local minimum	steeper landscape
curriculum helps mostly at the beginning (one step pacing function)	Predicts faster convergence at the end, anywhere in final basin of attraction



NO PROBLEM… IF LOSS LANDSCAPE IS CONVEX



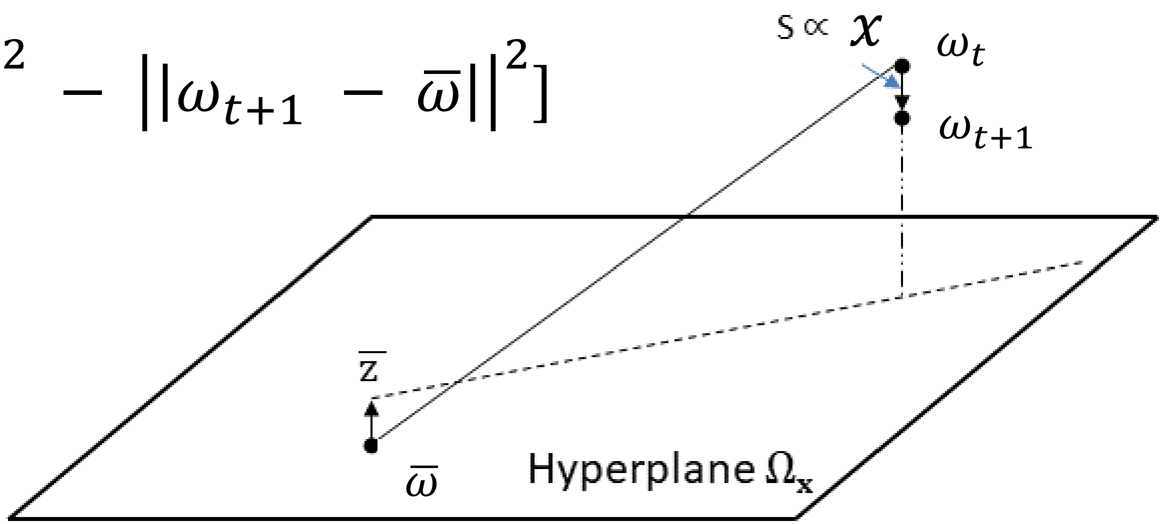
Densenet121 (Tom Goldstein)

Back to the regression loss…

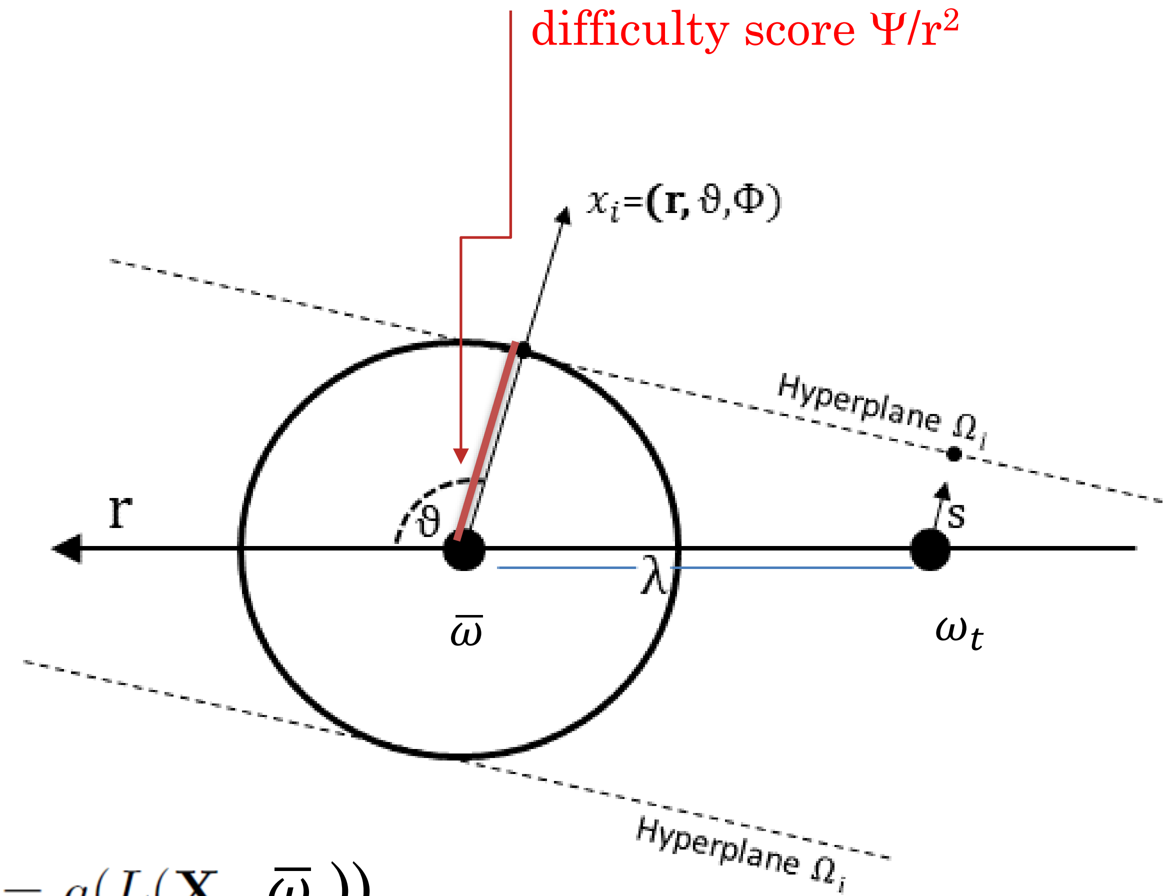
$$L(\omega,(x,y)) = (\omega \cdot x - y)^2$$

$$s = \frac{\partial L(\omega)}{\partial \omega}\Big|_{\omega=\omega_t} = 2\ (\omega_t \cdot x - y)\ x$$

$$\Delta = E\left[\left|\left|\omega_t\ -\ \overline{\omega}\right|\right|^2 - \left|\left|\omega_{t+1}\ -\ \overline{\omega}\right|\right|^2\right]$$



Computing the gradient step



$$\Psi(\mathbf{X}) = g(L(\mathbf{X}\ ,\overline{\omega}\ ))$$

$$\frac{1}{4}\Delta(\Psi) = \eta\mathbb{E}[r^2\lambda^2\cos^2\vartheta] - \eta^2\mathbb{E}[r^4\lambda^2\cos^2\vartheta] - \eta^2\Psi^2\mathbb{E}[r^2]$$

Theoretical analysis: linear regression loss

- **Theorem**: convergence rate is monotonically decreasing with the *Difficulty Score* of a point.

  Proof: $\frac{\partial \Delta(\Psi)}{\partial \Psi} \leq 0$

- **Theorem**: convergence rate is monotonically increasing with the *loss* of a point with respect to the *current hypothesis.*

- **Corollary:** expect faster convergence at the beginning of training (only true for regression loss)

  Proof: $\frac{\partial \Delta(\Psi)}{\partial \lambda} \geq 0$ when $\eta \leq \frac{\mathbb{E}[r^2 \cos^2 \vartheta]}{\mathbb{E}[r^4 \cos^2 \vartheta]}$

THEORETICAL ANALYSIS: LINEAR REGRESSION LOSS

- **Theorem**: convergence rate is monotonically decreasing with the *Difficulty Score* of a point.

- **Theorem**: convergence rate is monotonically increasing with the *loss* of a point with respect to the *current hypothesis.*

- **Corollary:** expect faster convergence at the beginning of training (only true for regression loss)

Loss with respect to current hypothesis

$$\Upsilon(\mathbf{X}) = g(L(\mathbf{X}, \omega_t))$$

$$\frac{1}{4\eta}\Delta(\Psi_0, \Upsilon) = \Psi_0^2 + \Upsilon^2 + 2\Psi_0\Upsilon\nabla$$

$$\nabla = \frac{f(\frac{\Psi+\Upsilon}{\lambda}) - f(\frac{\Psi-\Upsilon}{\lambda}) - f(\frac{-\Psi+\Upsilon}{\lambda}) + f(\frac{-\Psi-\Upsilon}{\lambda})}{f(\frac{\Psi+\Upsilon}{\lambda}) + f(\frac{\Psi-\Upsilon}{\lambda}) + f(\frac{-\Psi+\Upsilon}{\lambda}) + f(\frac{-\Psi-\Upsilon}{\lambda})}$$

Theorem *Assume that the gradient step size is small enough so that we can neglect second order terms $O(\eta^2)$, and that $\frac{\partial\nabla}{\partial\Upsilon} \geq \frac{\Psi}{\Upsilon} - \frac{\Upsilon}{\Psi}\ \forall\Upsilon$. Fix the difficulty score at Ψ_0. At time t the expected convergence rate is monotonically increasing with the local difficulty $\Upsilon(\mathbf{x})$.*

Corollary *For any $c \in \mathbb{R}^+$, if ∇ is $(c - \frac{1}{c})$-Lipschitz then $\frac{\partial\Delta(\Psi,\Upsilon)}{\partial\Upsilon} \geq 0$ for any $\Upsilon \geq c\,\Psi$.*

Hinge loss

$$L(\mathbf{X}, \mathbf{w}) = \max(1 - (\mathbf{x} \cdot \mathbf{w})y, 0)$$

$$\Delta(\Psi) = \mathbb{E}\left[\frac{\mathbf{w}_{t+1} \cdot \bar{\mathbf{w}}}{\|\mathbf{w}_{t+1}\| \|\bar{\mathbf{w}}\|} - \frac{\mathbf{w}_t \cdot \bar{\mathbf{w}}}{\|\mathbf{w}_t\| \|\bar{\mathbf{w}}\|} \,\bigg|_{\Psi} \right]$$

$$= \int_{-\infty}^{\mathcal{B}(\Psi)} \eta[(1-\Psi)\sin^2\vartheta - x_2 \sin\vartheta\cos\vartheta] \cdot f(x_2) dx_2 + O(\eta^2)$$

Theorem *Assume that the gradient step size is small enough so that we can neglect second order terms $O(\eta^2)$. The expected convergence rate decreases monotonically as a function of Ψ for every $\Psi > (1 - \cos\vartheta)$ when $\cos\vartheta > 0$ ($\bar{\mathbf{w}}, \mathbf{w}_t$ are positively correlated), and for every $\Psi < (1 - \cos\vartheta)$ when $\cos\vartheta < 0$. Monotonicity holds $\forall\Psi$ when $\cos\vartheta = 0$.*

Theorem *Assume that the gradient step size is small enough so that we can neglect second order terms $O(\eta^2)$. Assume further that $\cos\vartheta \geq 0$. Fixing Ψ and $\forall\Psi$, the expected convergence rate is monotonically increasing with Υ for every $\Upsilon > 0$.*

Summary and discussion

1. First theoretical demonstration that curriculum learning indeed helps, speeding up convergence during training. Previous related results have relied mostly on empirical evidence.

2. The literature is **confusing**, with 2 apparently conflicting methods:
   - ⇨ Curriculum learning, giving preference to easier examples
   - ⇦ Methods like hard example mining and boosting, which focus on the more difficult examples

   Resolution: results are consistent, it's all in how one measures difficulty:
   - ⇨ **Curriculum**: *Easy,* with respect to *final* hypothesis.
   - ⇦ **Hard example mining**: *Difficult*, with respect to *current* hypothesis.

3. Curriculum learning made practical:
   - **CL by transfer**: source network, which is bigger and more powerful, is used to sort the examples for the weaker network.
   - **CL by bootstrapping**: same pre-trained network is used to sort the examples

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

Guy Hacohen

Gad Cohen

Dan Amir

This Research is supported by the Israeli Science Foundation, the Gatsby Charitable Foundation, and Mafat Center for Deep Leaning.