# Early-Learning Regularization Prevents Memorization of Noisy Labels

**Carlos Fernandez-Granda**
www.cims.nyu.edu/~cfgranda

# Acknowledgements

Research partially supported by NSF awards NRT-HDR 1922658, DMS 2009752, and HDR 1940097

Joint work with Sheng Liu (NYU), Jon Niles-Weed (NYU), Narges Razavian (NYU School of Medicine), and Chhavi Yadav (UC San Diego)

## Early detection of Alzheimer's disease

## Classification with noisy labels

## Early-learning regularization (ELR)

Early detection of Alzheimer's disease

Classification with noisy labels

Early-learning regularization (ELR)

# Early detection of Alzheimer's disease

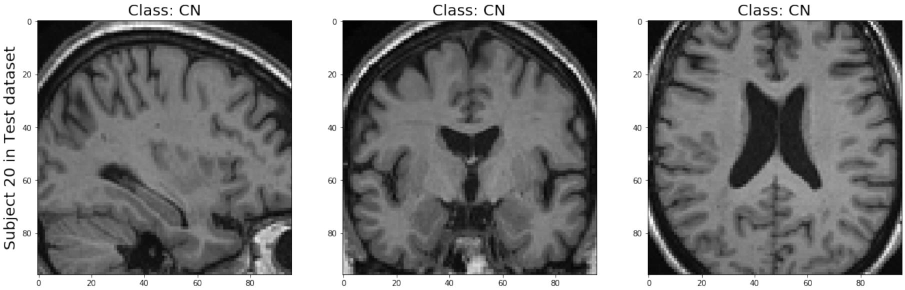
Important to provide treatment, and populate clinical trials

Positron-emission tomography is effective, but invasive and very costly

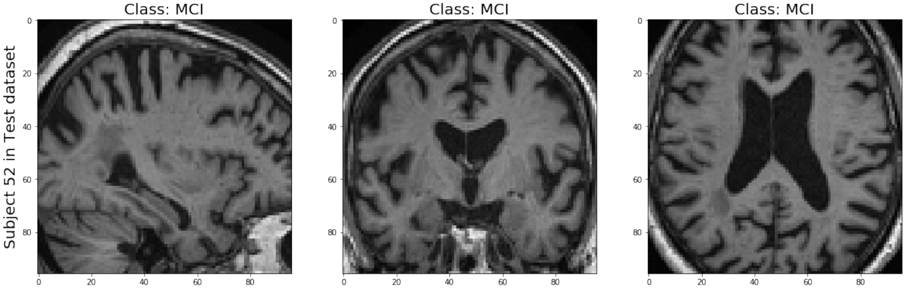
Structural MRI (T1) is less costly, but not so accurate

Goal: Use deep learning to increase accuracy

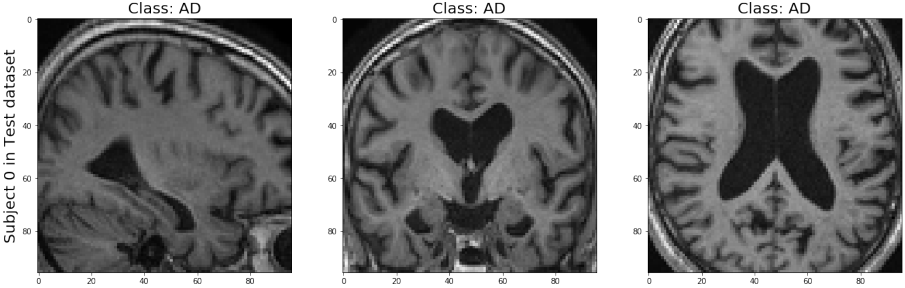
# Structural MRI of cognitively-normal patient



# Structural MRI of mildly cognitively impaired patient (MCI)



# Structural MRI of Alzheimer's patient



# Early diagnostics of Alzheimer's disease

Goal: Distinguish between three classes

1. Cognitively normal (CN)
2. Mild cognitive impairment (MCI)
3. Mild Alzheimer's disease (AD)

Dataset: Alzheimer's Disease Neuroimaging Initiative (ADNI)

# Demographics

| Split | Class | Num. subjects | Num. Scans | Mean Age (std) |
|---|---|---|---|---|
| Train | CN | 140 | 567 | 77.0 (5.4) |
| | MCI | 248 | 840 | 75.9 (7.3) |
| | AD | 193 | 527 | 76.7 (7.4) |
| Val | CN | 33 | 126 | 77.2 (5.6) |
| | MCI | 39 | 138 | 73.3 (7.2) |
| | AD | 41 | 124 | 76.1 (8.3) |
| Test | CN | 24 | 105 | 79.0 (6.1) |
| | MCI | 43 | 140 | 76.7 (6.5) |
| | AD | 45 | 135 | 76.4 (5.1) |

# Entorhinal cortex and hippocampus

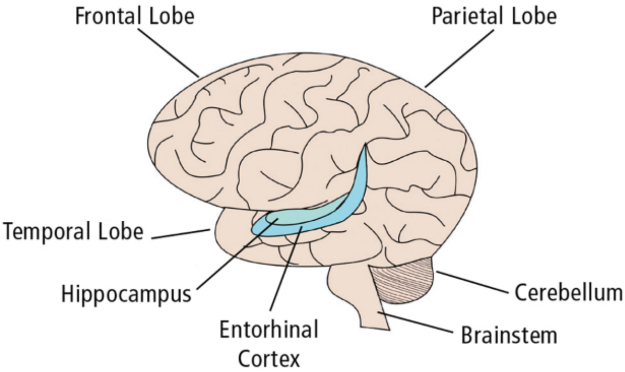
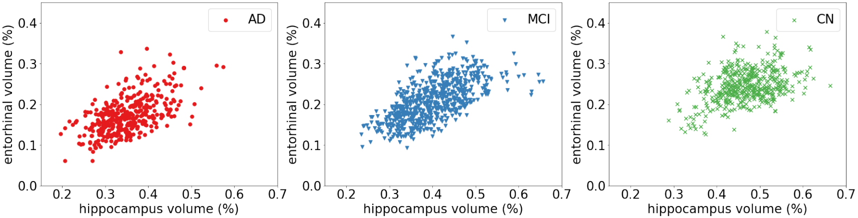



Image source: https://blog.edvotek.com/2020/04/09/insights-into-alzheimers-disease/

# Simple biomarker (normalized volumes)




Accuracy: around 62%

# Proposed methodology

Register images to common template

Train 3D convolutional neural network

Performance is improved by:

- Using small (1x1) filter sizes in first layer
- Widening the network (as opposed to deepening)
- Using instance normalization instead of batch normalization
- Encoding age using a sinusoidal embedding

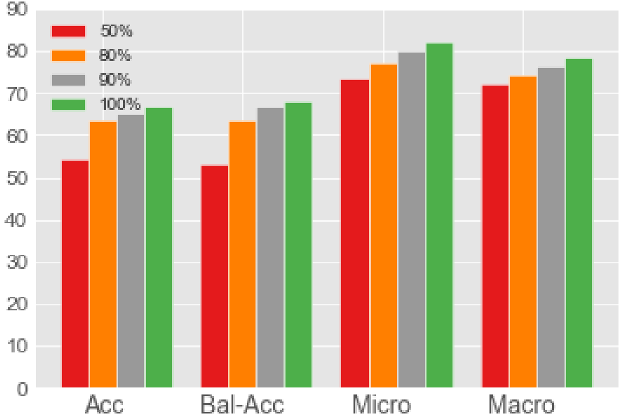
# Results

| Method | Accuracy | Balanced Acc | Micro-AUC | Macro-AUC |
|---|---|---|---|---|
| Volume-based | 61.9% | 62.1% | 78.0 % | 76.1% |
| ResNet-18 3D | 50.1% | 51.3% | 71.2% | 72.4% |
| AlexNet 3D | 57.2% | 56.2% | 75.1% | 74.2% |
| Proposed | 66.9% | 67.9% | 82.0% | 78.5% |
| Proposed + Age | 68.2% | 70.0% | 82.0% | 80.0% |

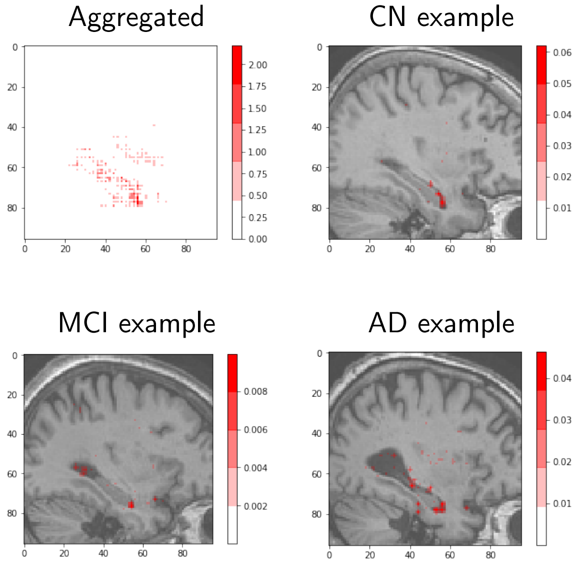
# Independent dataset
# (National Alzheimer's Coordinating Center)

| Method | Accuracy | Balanced Acc | Micro-AUC | Macro-AUC |
|---|---|---|---|---|
| Volume-based | 56.3% | 53.2% | 72.0% | 74.0% |
| Proposed | 74.2% | 60.1% | 87.0% | 80.0% |

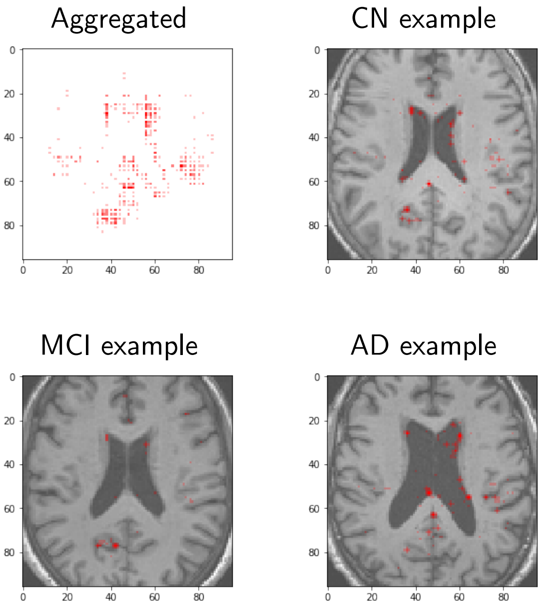
# Main obstacle for improvement: Limited data



# Visualization of gradient with respect to input (axial view)



# Visualization of gradient with respect to input (sagittal view)



# Challenge

Labeling is highly subjective and noisy

For example, MCI diagnosis criteria are:

1. Subjective memory complaints
2. Objective memory loss (scoring below education-adjusted cut-off on Logical Memory Test)
3. Global Clinical Dementia Rating (interview-based rating) of 0.5
4. Diagnosis of dementia could not be made by physician

# For more information

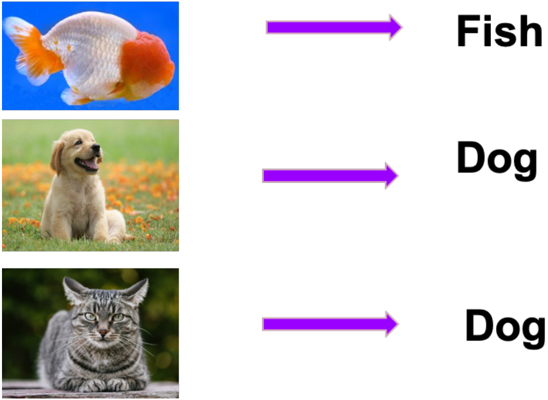
On the Design of Convolutional Neural Networks for Automatic Detection of Alzheimer's Disease S. Liu, C. Yadav, C. Fernandez-Granda, N. Razavian
*NeurIPS Machine Learning for Healthcare (ML4H) workshop 2019*
*Proceedings of Machine Learning Research, PMLR 116 171-183*

Early detection of Alzheimer's disease

Classification with noisy labels

Early-learning regularization (ELR)

# Noisy labels

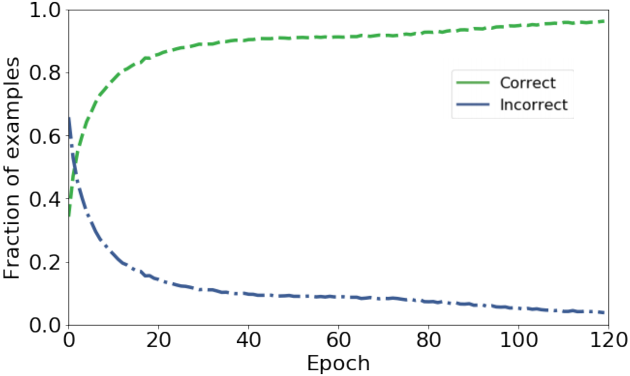



Idealized setting (definitely not what is happening in Alzheimer's data)

# Experiment

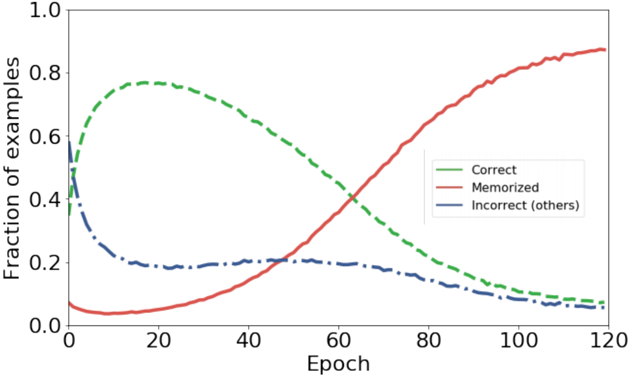
Train ResNet-18 on CIFAR10, with 40% of labels flipped at random

# Predictions on training examples with clean labels



# Predictions on training examples with incorrect labels

Two stages: Early learning and then memorization



# Early learning + Memorization

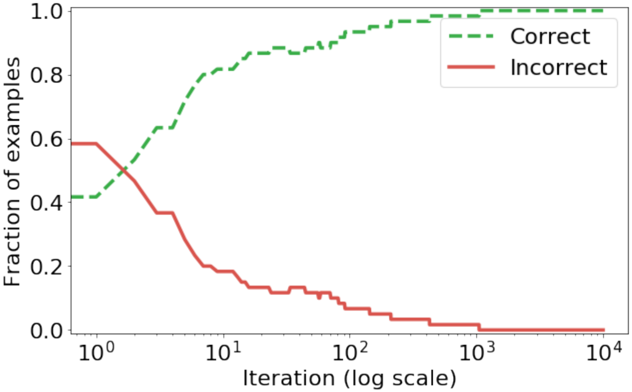
Well known phenomenon in deep learning

But is it unique to deep learning?

Let's see what happens for a separable 2-class problem

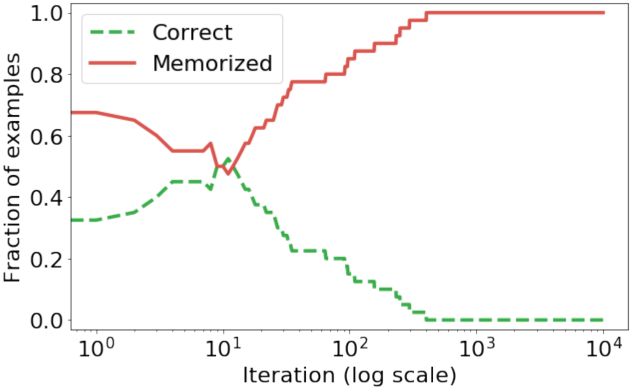
We flip 40% of the labels and fit linear model

# Predictions on training examples with clean labels



# Predictions on training examples with incorrect labels

Two stages: Early learning and then memorization



# Analysis of linear model

Cross-entropy loss function for a single example

$$\mathrm{CE} := -\sum_{c=0}^{1} \mathbf{y}_c \log \mathbf{p}_c$$

$$\mathbf{p}_c := \frac{\exp(\Theta_c^T \mathbf{x})}{\exp(\Theta_0^T \mathbf{x}) + \exp(\Theta_1^T \mathbf{x})}$$

- $\mathbf{y}$: label
- $\mathbf{p}$: model estimate
- $\Theta_0$, $\Theta_1$: model parameters
- $\mathbf{x}$: feature vector

# Gradient of linear model

Gradient of cross-entropy loss function for a single example

$$\nabla_{\Theta_c} \mathrm{CE} = \mathbf{x}\left(\mathbf{p}_c - \mathbf{y}_c\right)$$

Separable model:

- Label 1: $\mathbf{x}^{[i]} := +\mathbf{v} +$ random vector
- Label 0: $\mathbf{x}^{[i]} := -\mathbf{v} +$ random vector

Ideally, we would learn $\Theta_1 = +\mathbf{v}$

But some labels are flipped

# Early learning

Gradient of cross-entropy loss function for $n$ examples

$$\nabla_{\Theta_1} \mathrm{CE} = \sum_{i=1}^{n} \mathbf{x}^{[i]} \left( \mathbf{p}_1^{[i]} - \mathbf{y}_1^{[i]} \right)$$

Sum of all examples with label 1 weighted by $\mathbf{p}_1^{[i]} - 1$ and examples with label 0 weighted by $\mathbf{p}_0^{[i]}$

During gradient descent correct labels push $\Theta_1$ towards $\mathbf{v}$ (*random vectors cancel out*)

Majority (60%) are correct so early learning occurs

# Memorization

Gradient of cross-entropy loss function for $n$ examples

$$\nabla_{\Theta_1}\mathrm{CE} = \sum_{i=1}^{n} \mathbf{x}^{[i]} \left(\mathbf{p}_1^{[i]} - \mathbf{y}_1^{[i]}\right)$$

After some time $\Theta_1$ aligns with $\mathbf{v}$

Good news: Correct labels are well classified!

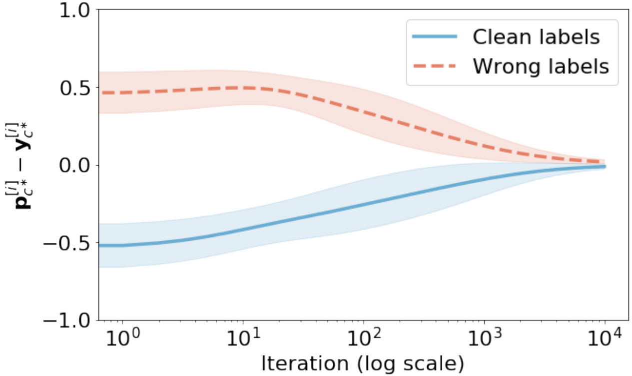
Bad news: Their influence on the gradient **vanishes**

Incorrectly labeled data dominate gradient

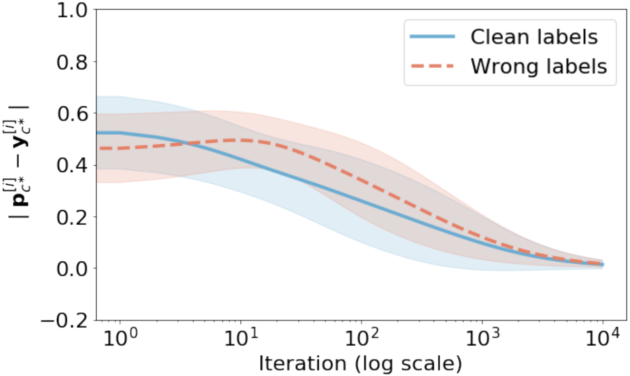
In high dimensions we can find a hyperplane that fits any set of labels

Eventually we find it and memorization occurs

# Gradient of linear model



# Gradient of linear model



# Analysis of deep-learning model

Cross-entropy loss function for a single example

$$\mathrm{CE} := -\sum_{c=1}^{C} \mathbf{y}_c \log \mathbf{p}_c$$

$$\mathbf{p}_c := \frac{\exp(f_\Theta(\mathbf{x})_c)}{\sum_{k=1}^{C} \exp(f_\Theta(\mathbf{x})_k)}$$

- $\mathbf{y}$: label
- $\mathbf{p}$: model estimate
- $\Theta$: model parameters
- $f_\Theta$: deep neural network
- $\mathbf{x}$: feature vector

# Gradient of deep-learning model

Gradient of cross-entropy loss function for a single example

$$\nabla_{\Theta} \mathrm{CE} = J(\mathbf{x}) \, (\mathbf{p} - \mathbf{y})$$

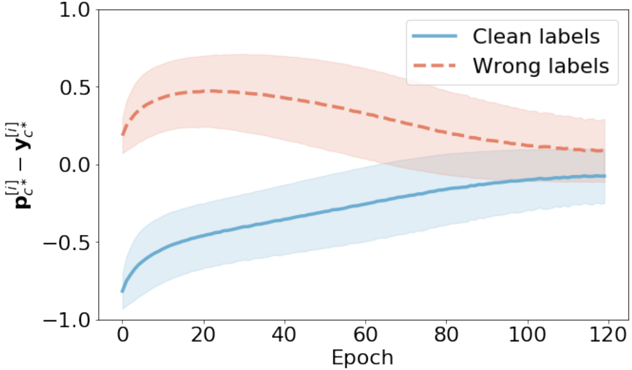
$J(\mathbf{x})$ is the Jacobian of $f_{\Theta}(\mathbf{x})$ with respect to $\Theta$

Same as linear model except that $J(\mathbf{x})$ replaces $\mathbf{x}$

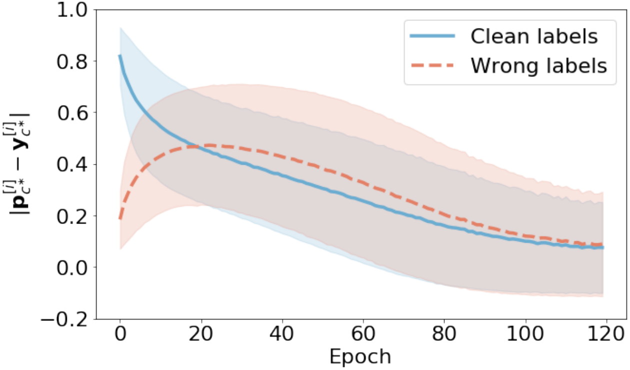
Intuition from linear model:

- At first, correct labels dominate $\rightarrow$ early learning
- Then, their contribution to gradient vanishes and incorrect labels dominate $\rightarrow$ memorization

# Gradient of deep-learning model



# Gradient of deep-learning model



Early detection of Alzheimer's disease

Classification with noisy labels

Early-learning regularization (ELR)

# Idea

Use early learning model to neutralize effect of incorrect labels and avoid memorization

# ELR loss function

For each example $i$, target $\mathbf{q}^{[i]}$ is a *corrected* label estimate based on past model outputs

$$\mathrm{ELR} := \mathrm{CE} + \frac{\lambda}{n} \sum_{i=1}^{n} \log\left(1 - \langle \mathbf{p}^{[i]}, \mathbf{q}^{[i]} \rangle\right)$$

Regularization tries to align model

Crucial insight: Targets don't need to be right, they just need to neutralize gradient from incorrect labels

# Gradient of ELR loss function

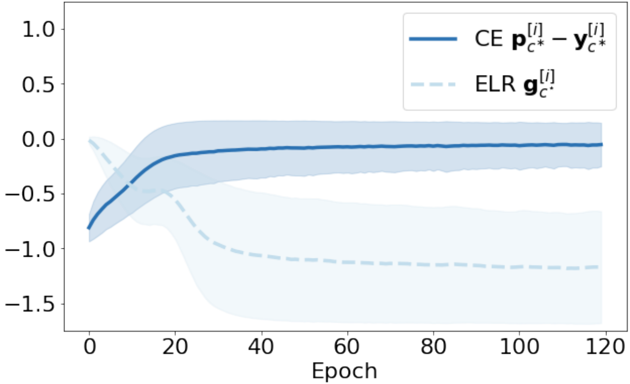
$$\nabla_{\Theta}\mathrm{CE} = J(\mathbf{x})\left(\mathbf{p} - \mathbf{y} + \lambda\mathbf{g}\right)$$

After early learning, regularization term neutralizes incorrect labels and boosts influence of correct labels

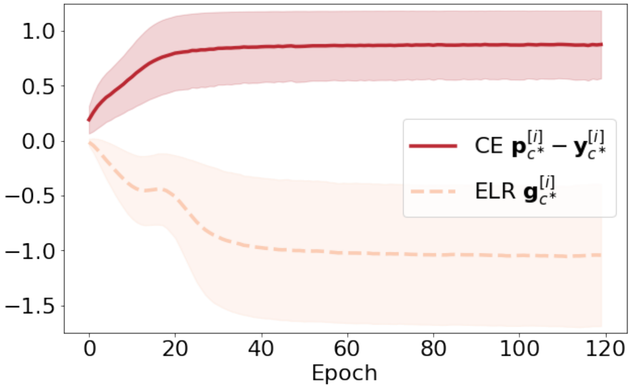
# Experiment

Train ResNet-18 on CIFAR10, with 40% of labels flipped at random

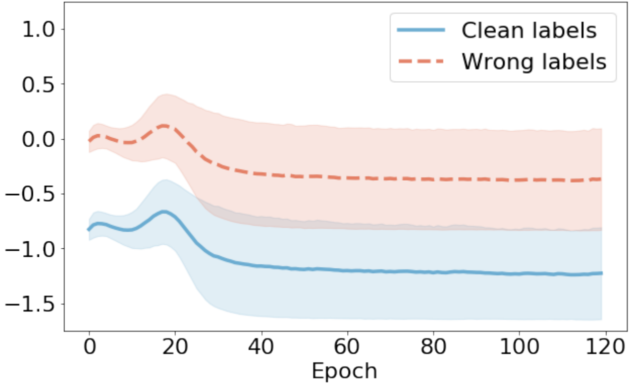
# Gradient for examples with correct labels



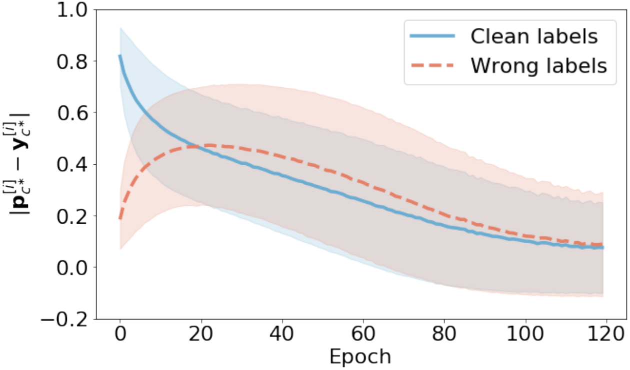
# Gradient for examples with incorrect labels



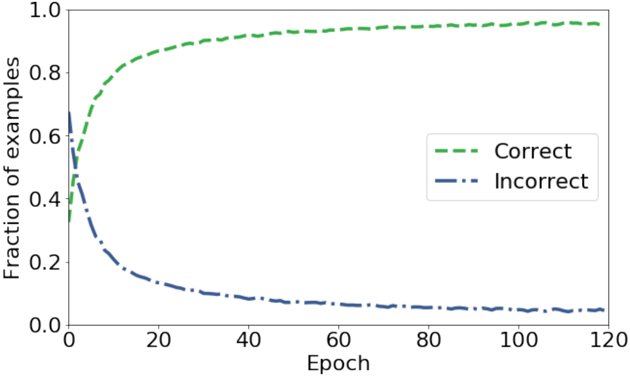
# Gradient of ELR loss function



# Gradient of cross-entropy loss function

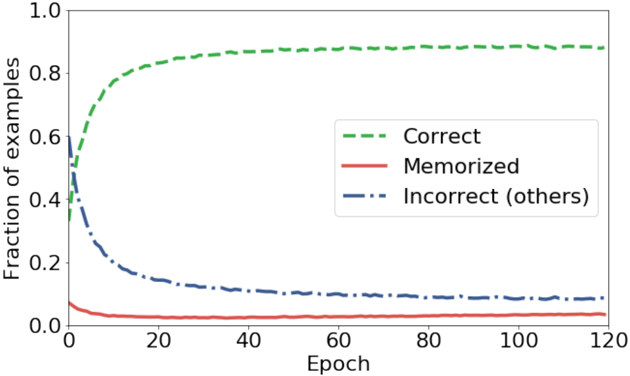


# ELR predictions on training examples with clean labels



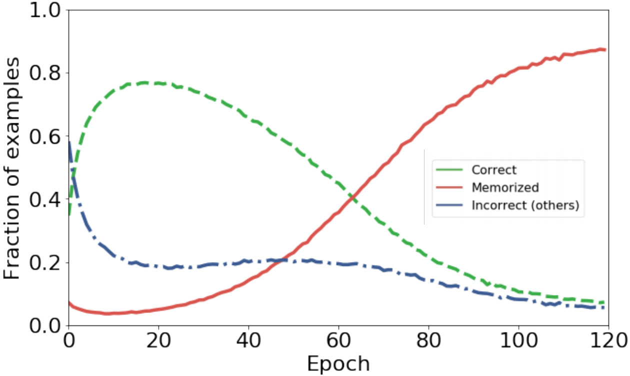
# ELR predictions on training examples with incorrect labels

After early learning no memorization



# CE predictions on training examples with incorrect labels

Two stages: Early learning and then memorization



# How do we estimate the targets

Using ideas from semi-supervised learning

- Temporal averaging of model output
- Temporal averaging of model weights
- Training two models
  (targets of one estimated from output of the other)

# Results

State of the art results on CIFAR-10, CIFAR-100, and two real-world datasets (Clothing-1M and WebVision)

# Future work

- Theoretical analysis for deep learning models
- Other choices of regularization?
- Focus on more realistic noise models

# For more information

Early-Learning Regularization Prevents Memorization of Noisy Labels

S. Liu, J. Niles-Weed, N. Razavian, C. Fernandez-Granda. NeurIPS 2020