Decoupling "when to update" from "how to update"

Eran Malach and Shai Shalev-Shwartz
School of Computer Science, The Hebrew University, Israel

Introduction

- Deep learning requires enormous amounts of labeled data
- Labeling is often done by crowdsourcing or data-mining - a fast labeling process which does not require manual work
- This usually results in noisy labels
- Using data with noisy labels can crucially fail the training process:
  - Updates are performed (mostly) on examples with wrong predictions
  - As the network improves most of its updates will be on noisy examples, thus further improvement is impossible

Method

- Solution idea: decoupling the decision of "when to update" from the update rule itself
- "When to update": train two classifiers and update only when they disagree
- Meta-algorithm for training a classifier from hypothesis class $\mathcal{H}$ on data sampled from noisy distribution $\tilde{\mathcal{D}}$ using update rule $\mathbf{U}$

for $t = 1, 2, \ldots, N$ **do**
  draw mini-batch $(x_1, y_1), \ldots, (x_b, y_b) \sim \tilde{\mathcal{D}}^b$
  let $S = \{(x_i, y_i) : h_1(x_i) \neq h_2(x_i)\}$
  $h_1 \leftarrow U(h_1, S)$
  $h_2 \leftarrow U(h_2, S)$
end for

Theoretical Analysis

- Suppose $\mathcal{D}$ is a linearly separable distribution, and $\tilde{\mathcal{D}}$ is the same distribution with labels flipped with probability $\mu < 0.5$
- Denote w^* the linear separator: $y\langle x, w^*\rangle \geq 1$ for $(x, y) \sim \mathcal{D}$
- Initialize two linear classifiers with norm $\leq K$
- Perceptron as the base classifier, update rule: $w_t^{(i)} = w_{t-1}^{(i)} + \tilde{y}_t x_t$
- The algorithm converges within a number of iterations that is similar up to a constant factor to that of a perceptron on noise-less data

Theorem 1. *Let T be the number of updates performed by the algorithm. Then,* $\mathbb{E}[T] \leq \frac{3(4K+1)}{(1-2\mu)^2} \|w^*\|^2$

- Does not imply convergence to an optimum, as the algorithm can get "stuck" before reaching an optimum
- Empirical evaluation shows that this can be avoided

Proof Idea

- The proof is essentially similar to the classical perceptron convergence proof
- Denote $w_t^{(1)}, w_t^{(2)}$ the weight vectors of the two classifiers after t updates
- By the disagreement rule: $\mathbf{sign}\langle w_{t-1}^{(1)}, x_t\rangle \neq \mathbf{sign}\langle w_{t-1}^{(2)}, x_t\rangle$
- Upper bound $\|w_t^{(i)}\|$:

$$\begin{aligned}\|w_t^{(i)}\|^2 &= \|w_{t-1}^{(i)}\|^2 + 2y_t\langle w_{t-1}^{(i)}, x_t\rangle + \|x_t\|^2 \\ &\leq \|w_{t-1}^{(i)}\|^2 + \underbrace{\|w_{t-1}^{(1)} - w_{t-1}^{(2)}\|}_{\text{constant}}\|x_t\| + \|x_t\|^2\end{aligned}$$

- Lower bound $\mathbb{E}_{\tilde{\mathcal{D}}}[\langle w_t^{(i)}, w^*\rangle]$:

$$\begin{aligned}\mathbb{E}[\langle w_t^{(i)}, w^*\rangle] &= \mathbb{E}[\langle w_{t-1}^{(i)}, w^*\rangle] + \mathbb{E}[\tilde{y}_t\langle x_t, w^*\rangle] \\ &= \mathbb{E}[\langle w_{t-1}^{(i)}, w^*\rangle] + (1-2\mu)\underbrace{\mathbb{E}[y_t\langle x_t, w^*\rangle]}_{\geq 1}\end{aligned}$$

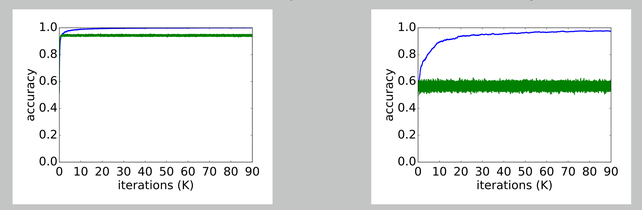
- Combining these bounds:

$$\begin{aligned}(1-2\mu)\,\mathbb{E}[T] &\lesssim \mathbb{E}[\langle w_t^{(i)}, w^*\rangle] \leq \mathbb{E}[\|w_t^{(i)}\|\|w^*\|] \\ &\lesssim \|w^*\|\,\mathbb{E}[\sqrt{T}] \leq \|w^*\|\sqrt{\mathbb{E}[T]}\end{aligned}$$

Notation: we use $\lesssim$ to hide constants

Experiments: Linear Classification

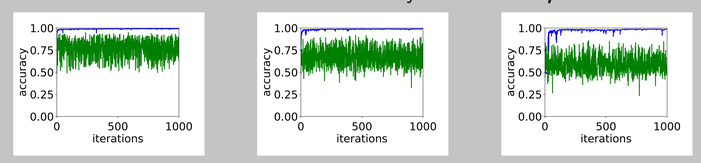
- Linear classification on Gaussian synthetic data with noise μ



$\mu = 0.01$ $\mu = 0.4$

Figure 1: Our algorithm (blue) compared to a vanilla perceptron (green) on Gaussian data

- Linear classification on MNIST data with synthetic noise μ



$\mu = 0.2$ $\mu = 0.3$ $\mu = 0.4$

Figure 2: Mean accuracy of our algorithm vs. perceptron on MNIST classification with noise

Experiments: Deep Learning (Dataset)

- Testing on deep neural-networks trained with momentum SGD
- Construct a real-world noisy dataset for gender classification:
  - Labeled Faces in the Wild benchmark — celebrities images with names [1]
  - To provide labeling, use an online service to match gender to name
  - Filter examples with small confidence to get a dataset with less noise

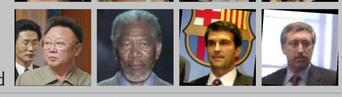
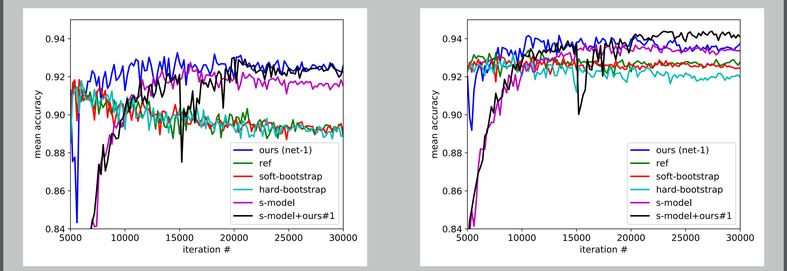
Name	Kim	Morgan	Joan	Leslie
Confidence	88%	64%	82%	88%
Correct				
Mislabeled				

Figure 3: Images from the dataset tagged as female

Experiments: Deep Learning (Results)

- This method proves to be more effective than competing methods [2, 3]
- Combination with other methods can further improve results



Dataset #1 - more noise Dataset #2 - less noise

Figure 4: Balanced accuracy of all methods on clean test data, trained on different datasets.

Conclusion

- A simple solution to supervised learning in the presence of noisy labels, which can be adapted to any update-based algorithm
- Theoretical analysis shows fast convergence for linear classification
- Empirical results show that the method improves the robustness of neural-networks to noisy labels in a real-world scenario

References

[1] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller.
Labeled faces in the wild: A database for studying face recognition in unconstrained environments.
Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[2] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich.
Training deep neural networks on noisy labels with bootstrapping.
arXiv preprint arXiv:1412.6596, 2014.

[3] Jacob Goldberger and Ehud Ben-Reuven.
Training deep neural networks using a noise adaptation layer.
ICLR, 2017.

eran.malach@mail.huji.ac.il