# TTIC 31230, Fundamentals of Deep Learning

David McAllester, April 2017

## Some Generalization Theory

## Generalization Bounds

A generalization bound is a theorem guaranteeing a certain performance on test data.

PAC-Bayesian generalization bounds are sufficiently general to handle arbitrary feed-forward computation (or even arbitrary prediction rules).

However, PAC-Bayesian Bounds for circuits with real-valued parameters typically involve simultaneous weight norm regularization and ensemble regularization.

## A PAC-Bayesian Generalization Bound

For $\Theta \in \mathbb{R}^d$ and $(x, y) \in \mathcal{Z}$ let $\ell(x, y, \Theta)$ be any loss function such that for all $\Theta$, $x$ and $y$ we have $\ell(x, y, \Theta) \in [0, 1]$.

Assume a data distribution from which we can draw problem instances and a training set $(x_1,\ y_1),\ \ldots,\ (x_N,\ y_N)$ drawn IID form that distribution.

$$\hat{\ell}(\Theta) \ = \ \frac{1}{N} \sum_{i=0}^{N-1} \ell(x_i,\ y_i,\ \Theta)$$

$$\ell(\Theta) \ = \ \mathrm{E}_{(x,y)} \left[\ell(x,\ y,\ \Theta)\right]$$

## A Generalization Theorem

Now we consider an ensemble model.

Each model in the ensemble is defined by a random vector $\epsilon$ (analogous to the random mask $\mu$ of dropout).

Here $\epsilon$ is drawn from $\mathcal{N}(0,1)^d$.

$$\hat{\ell}_{\text{ens}}(\Theta) = \mathrm{E}_\epsilon \left[ \hat{\ell}(\Theta + \epsilon) \right]$$

$$\ell_{\text{ens}}(\Theta) = \mathrm{E}_\epsilon \left[ \ell(\Theta + \epsilon) \right]$$

## The Theorem

For any $\lambda > 1/2$, with probability at least $1-\delta$ over the draw of the sample, we have that the following holds **simultaneously** for all $\Theta \in \mathbb{R}^d$.

$$\ell_{\text{ens}}(\Theta) \leq \frac{1}{1-\frac{1}{2\lambda}} \left( \hat{\ell}_{\text{ens}}(\Theta) + \frac{\lambda}{N} \left( \frac{1}{2} ||\Theta||^2 + \ln \frac{1}{\delta} \right) \right)$$

## More General Ensembles

We can consider any probability density $Q$ on $\mathbb{R}^d$ as defining an ensemble of models.

$$\hat{\ell}_Q(\Theta) = \mathrm{E}_{\Theta \sim Q}\left[\hat{\ell}(\Theta)\right]$$

$$\ell_Q(\Theta) = \mathrm{E}_{\Theta \sim Q}\left[\ell(\Theta)\right]$$

## More General Ensembles

For any "prior" distribution $P$ on $\mathbb{R}^d$, for any $\lambda > 1/2$, with probability at least $1 - \delta$ over the draw of the sample, we have that the following holds **simultaneously** for all ensembles $Q$.

$$\ell_Q(\Theta) \leq \frac{1}{1 - \frac{1}{2\lambda}} \left( \hat{\ell}_Q(\Theta) + \frac{\lambda}{N} \left( KL(Q, P) + \ln \frac{1}{\delta} \right) \right)$$

# A Proof

$$\ell_Q(\Theta) \leq \frac{1}{1-\frac{1}{2\lambda}}\left(\hat{\ell}_Q(\Theta) + \frac{\lambda}{N}\left(KL(Q,P) + \ln\frac{1}{\delta}\right)\right)$$

$$\ell_{\text{ens}}(\Theta) \leq \frac{1}{1-\frac{1}{2\lambda}}\left(\hat{\ell}_{\text{ens}}(\Theta) + \frac{\lambda}{N}\left(\frac{1}{2}||\Theta||^2 + \ln\frac{1}{\delta}\right)\right)$$

To prove the second from the first we take $P$ to be $\mathcal{N}(0,1)^d$ (the noise distribution) and $Q$ to be the distribution defined by $\Theta + \epsilon$ for noise $\epsilon$. We then have

$$KL(Q,P) = \frac{1}{2}||\Theta||^2$$

## Compression Regularization

Highly compressed models may have improved generalization.

Consider finite precision representations of $\Theta$. **Let $|\Theta|$ be the number of bits it takes to represent $\Theta$.**

**Theorem:** With probability at least $1 - \delta$ over the draw of the sample the following holds *simultaneously* for all $\Theta$ and $\lambda > 1/2$.

$$\ell(\Theta) \leq \frac{1}{1 - \frac{1}{2\lambda}} \left( \hat{\ell}(\Theta) + \frac{\lambda}{N} \left( (\ln 2)|\Theta| + \ln \frac{1}{\delta} \right) \right)$$

Here there is no ensemble involved.

## Proof

Note that for a fixed $\Theta$ we have that $\hat{\ell}(\Theta)$ is different for different training sets.

Note that $\hat{\ell}(\Theta)$ is an average over a large number of measurements. The law of large numbers says that such an average should be normally distributed.

$$p(\hat{\ell}(\Theta)) \approx e^{\frac{-(\hat{\ell}(\Theta)-\ell(\Theta))^2}{2\sigma^2}}$$

More precisely we have the following Chernoff bound.

$$P_{\text{draw\_of\_train}}\left(\hat{\ell}(\Theta) \leq \ell(\Theta) - \epsilon\right) \leq e^{-N\frac{\epsilon^2}{2\ell(\Theta)}}$$

For a given $h \in \mathcal{H}$ the relative Chernoff bound states that

$$P\left(\hat{\ell}(\Theta) \leq \ell(\Theta) - \epsilon\right) \leq e^{-N\frac{\epsilon^2}{2\ell(\Theta)}}$$

Let $Pr$ be any "prior" distribution on $\Theta$ and take

$$\epsilon = \sqrt{\frac{2\ell(\Theta)\left(\ln\frac{1}{Pr(\Theta)} + \ln\frac{1}{\delta}\right)}{N}}$$

we get

$$P\left(\ell(\Theta) > \hat{\ell}(\Theta) + \sqrt{\frac{2\ell(\Theta)\left(\ln\frac{1}{Pr(\Theta)} + \ln\frac{1}{\delta}\right)}{N}}\right) \leq Pr(\Theta)\delta$$

$$P\left(\ell(\Theta) > \hat{\ell}(\Theta) + \sqrt{\frac{2\ell(\Theta)\left(\ln\frac{1}{Pr(\Theta)} + \ln\frac{1}{\delta}\right)}{N}}\right) \leq Pr(\Theta)\delta$$

Union Bound: $P(\exists x\ \Phi[x]) \leq \sum_x P(\Phi[x])$

$$P\left(\exists\Theta\ \ \ell(\Theta) > \widehat{\ell}(\Theta) + \sqrt{\ell(\Theta)\left(\frac{2\left(\ln\frac{1}{Pr(\Theta)} + \ln\frac{1}{\delta}\right)}{N}\right)}\right) \leq \delta$$

With probability at least $1-\delta$

$$\forall\Theta \;\; \ell(\Theta) \leq \widehat{\ell}(\Theta) + \sqrt{\ell(\Theta)\left(\frac{2\left(\ln\frac{1}{Pr(\Theta)} + \ln\frac{1}{\delta}\right)}{N}\right)}$$

using $\sqrt{ab} = \inf_{\lambda>0} \; \frac{a}{2\lambda} + \frac{\lambda b}{2}$ we have

$$\ell(\Theta) \leq \widehat{\ell}(\Theta) + \frac{\ell(\Theta)}{2\lambda} + \left(\frac{\lambda\left(\ln\frac{1}{Pr(\Theta)} + \ln\frac{1}{\delta}\right)}{N}\right)$$

$$\ell(\Theta) \leq \widehat{\ell}(\Theta) + \frac{\ell(\Theta)}{2\lambda} + \left( \frac{\lambda \left( \ln \frac{1}{Pr(\Theta)} + \ln \frac{1}{\delta} \right)}{N} \right.$$

Solving for $\ell(\Theta)$ we have

$$\ell(\Theta) \leq \frac{1}{1 - \frac{1}{2\lambda}} \left( \hat{\ell}(\Theta) + \frac{\lambda}{N} \left( \ln \frac{1}{Pr(\Theta)} + \ln \frac{1}{\delta} \right) \right)$$

$$\ell(\Theta) \leq \frac{1}{1-\frac{1}{2\lambda}}\left(\hat{\ell}(\Theta) + \frac{\lambda}{N}\left(\ln\frac{1}{Pr(\Theta)} + \ln\frac{1}{\delta}\right)\right)$$

Using $Pr(\Theta) = 2^{-|\Theta|}$ we get

$$\ell(\Theta) \leq \frac{1}{1-\frac{1}{2\lambda}}\left(\hat{\ell}(\Theta) + \frac{\lambda}{N}\left((\ln 2)|\Theta| + \ln\frac{1}{\delta}\right)\right)$$

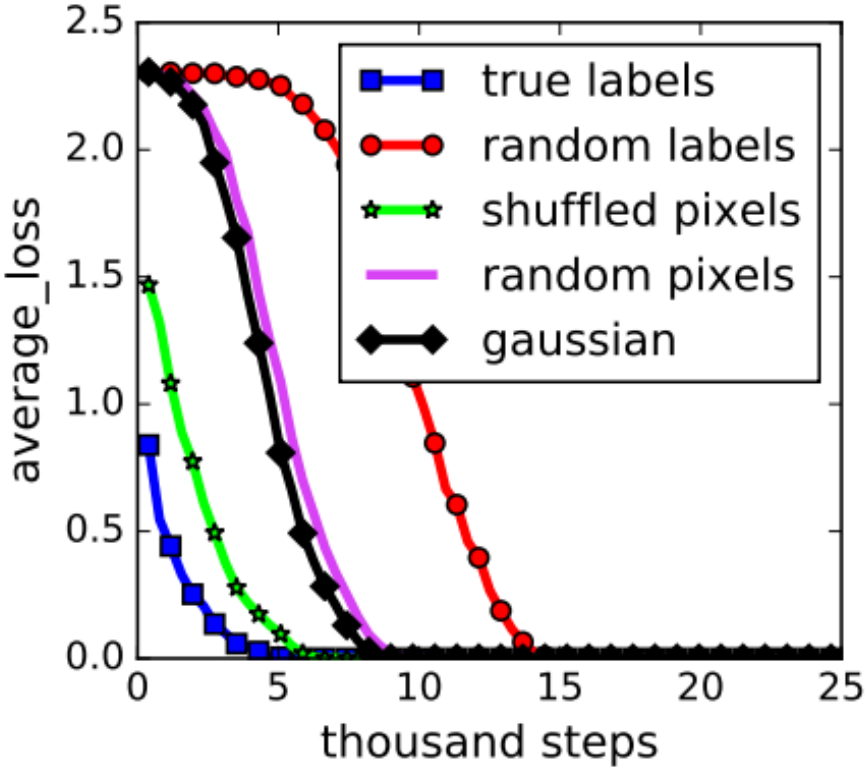
## Implicit Regularization

"Understanding deep learning requires rethinking generalization", Zhang et al. (November 2016).

## Troubling Experiments

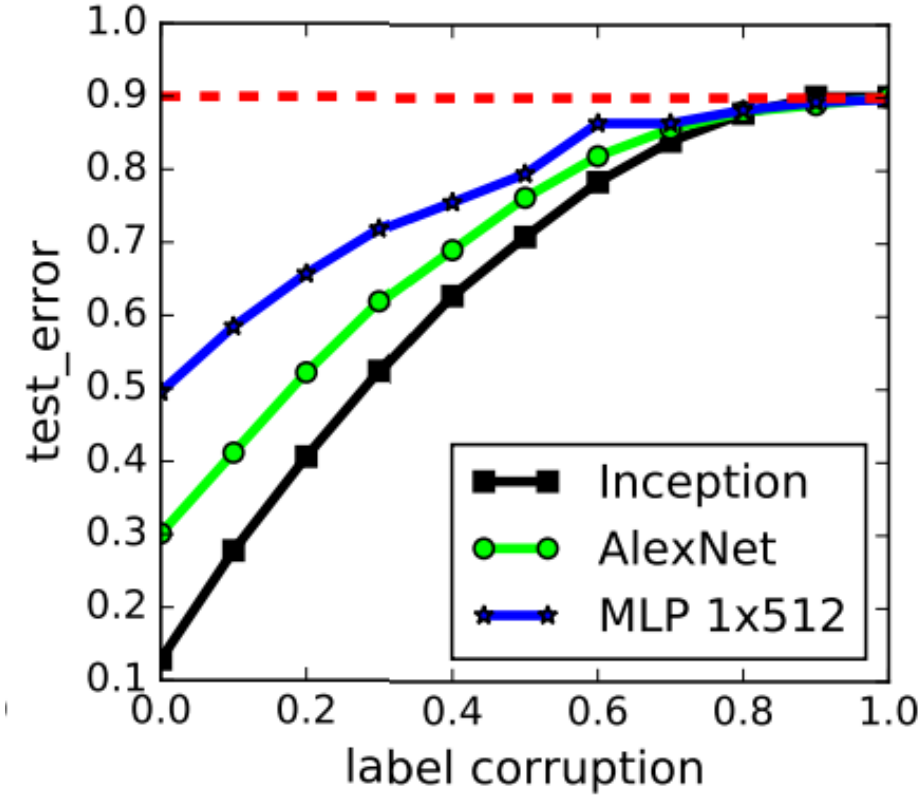
"Our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data."

# Training on Corrupted Data




Inception on CIFAR10

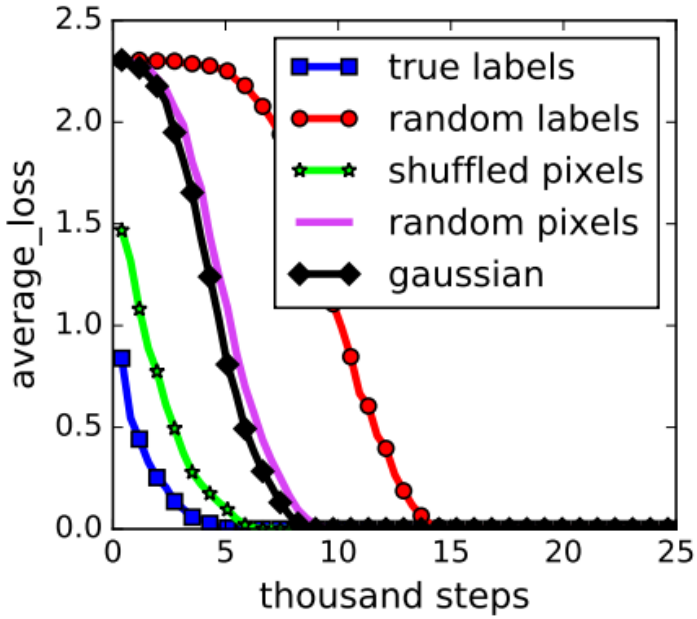
# Test Error as a Function of Training Label Corruption




(c) generalization error growth

# Implicit Regularization

One can modify the PAC-Bayesian bound (and other bounds) to replace $||\Theta||^2$ with $||\Theta - \Theta_{\text{init}}||^2$.



# END