Regularization for Deep Learning

Yoshua Bengio

May 12, 2016

Date Performed: May, 2012
Partners: Yan Yan

1 Introduction

To reduce the test error, increased training error may occur. Regularization methods were involved. There are a great may forms of regularization available. This chapter we describe regularization in more detail, focusing on regularization strategies for deep models or models that may be used as building blocks to form deep models. Strategies like: put extra constraints, adding restrictions on the parameter values, extra terms. (These constraints and penalties are designed to express a generic preference for a simpler model class in order to promote generalization, or make an underdetermined problem determined, or ensemble methods combine multiple hypotheses that explain the training data).
Regularization strategies are based on regularizing estimators. Regularization of an estimator works by trading increased bias for reduced variance **An effective regularizer is one that makes a profitable trade, reducing variance significantly while not overly increasing the bias**. Three situations where the model family being trained either: (1) excluded the true data generating process - corresponding to under-fitting and inducing bias, or (2) matched the true data generating process, or (3) included the generating process but also many other possible generating processes.
Most applications of deep learning algorithms are to domains where the true data generating process is almost certainly outside the model family. DL methods are typically applied to extremely complicated domains such as images audio sequences and text –the true generation process essentially involves simulating the entire universe.

2 Parameter Norm Penalties

Regularization has been used for decades prior to the advent of deep learning. Linear regression and logistic regression allow simple, straight forward and effective regularization strategies. Many are based on limiting the capacity of

models like neural networks, linear regression, or logistic regression, by adding a parameter norm penalty $\Omega(\theta)$ to the object function J.

$$\tilde{J}(\theta; X, y) = J(\theta; X, y) + \alpha\Omega(\theta) \tag{1}$$

where $\alpha \in [0, \infty)$ is a hyperparameter that weights the relative contribution of the norm penalty term. $\alpha = 0$ results in no regularization, larger values correspond to more regularization.
For NN, we typically choose to use a parameter norm penalty Ω that penalizes only the weights of the affine transformation at each layer and leaves the biased unregularized. The biases requires less data to fit accurately than the weights - each weight specifies how two variables interact.
Need further thinking here at page 232.

2.1 L^2 Parameter Regularization

The L^2 parameter norm penalty commonly known as *weight decay*. L^2 regularization is also known as ridge regression or Tikhonov regularization.

$$\tilde{J}(\omega; X, y) = 0.5\alpha\omega^t\omega + J(\omega; X, y) \tag{2}$$

with gradient

$$\nabla\omega\tilde{J}(\omega; X, y) = \alpha\omega + \nabla_\omega J(\omega; X, y) \tag{3}$$

the update rule:

$$\omega \leftarrow \omega - \epsilon(\alpha\omega + \nabla_\omega J(\omega; X, y)) \tag{4}$$

which equals to

$$\omega \leftarrow (1 - \epsilon\alpha)\omega - \epsilon\nabla_\omega J(\omega; X, y) \tag{5}$$

which means the addition of the weight decay term has modified the learning rule to multiplicatively shrink the weight vector by a constant factor on each step, just before performing the usual gradient update.
For the entire course of training, further make a quadratic approximation to the objective function in the neighbourhood of the value of the weights that obtains minimal unregularized training cost, $\omega^* = argmin_\omega J(\omega)$.

$$\hat{J}(\theta) = J(\omega^*) + 0.5(\omega - \omega^*)^T H(\omega - \omega^*) \tag{6}$$

where H is the Hessian matrix of J with respect to ωevaluated at ω^*. ω^* is defined to be a minimum of J, we can conclude that H is positive semidefinite. The minimum of $\hat{J}$ occurs where its gradient

$$\nabla_\omega\hat{J}(\omega) = H(\omega - \omega^*) \tag{7}$$

is equal to 0.

To study the weight decay, add the weight decay gradient, use $\tilde{\omega}$ to represent the location of the minimum.

$$\alpha\tilde{\omega} + H(\tilde{\omega} - \omega^*) = 0 \tag{8}$$

so we get

$$\tilde{\omega} = (H + \alpha I)^{-1} H\omega^* \tag{9}$$

from which we can see, as $\alpha -> 0$, the regularized $\tilde{\omega} -> \omega^*$. What about α grows?
H is real and symmetric, so $H = Q\Lambda Q^T$ where Λ is a diagonal matrix, and Q is an orthonormal basis of eigenvectors. From which we obtain:

$$\tilde{\omega} = Q(\Lambda + \alpha I)^{-1}\Lambda Q^T\omega^* \tag{10}$$

we see that the effect of weight decay is to rescale ω^* along the axes defined by the eigenvectors of H.

2.2 L^1 Parameter Regularization

Need to be done.

3 Norm Penalties as Constrained Optimization

$$\tilde{J}(\theta; X, y) = J(\theta; X, y) + \alpha\Omega(\theta) \tag{11}$$

We can minimize a function subject to constraints by constructing a generalized Lagrange function, consisting of the original objective function plus a set of penalties. Each penalty is a product between a coefficient called a KKT multiplier, and a function representing whether the constraint is satisfied.
If we want $\Omega(\theta) < k$, we construct a generalized Lagrange function:

$$L(\theta, \alpha; X, y) = J(\theta; X, y) + \alpha(\Omega(\theta) - k) \tag{12}$$

the solution to the constrained problem is given by

$$\theta^* = \arg\max_{\theta} \max_{\alpha, \alpha \geq 0} L(\theta, \alpha) \tag{13}$$

α increase, whenever $\Omega(\theta) > k$;
α decrease, whenever $\Omega(\theta) < k$.
So the optimal value α^* will encourage $\Omega(\theta)$ to shrink, but not so strongly to make $\Omega(\theta)$ less than k.

3.1 Definitions

Stoichiometry The relationship between the relative quantities of substances taking part in a reaction or forming a compound, typically a ratio of whole integers.

Atomic mass The mass of an atom of a chemical element expressed in atomic mass units. It is approximately equivalent to the number of protons and neutrons in the atom (the mass number) or to the average number allowing for the relative abundances of different isotopes.

Mass of empty crucible		7.28 g
Mass of crucible and magnesium before heating		8.59 g
Mass of crucible and magnesium oxide after heating		9.46 g
Balance used		#4
Magnesium from sample bottle		#1
Mass of magnesium metal	= 8.59 g - 7.28 g	
	= 1.31 g	
Mass of magnesium oxide	= 9.46 g - 7.28 g	
	= 2.18 g	
Mass of oxygen	= 2.18 g - 1.31 g	
	= 0.87 g	

Because of this reaction, the required ratio is the atomic weight of magnesium: 16.00 g of oxygen as experimental mass of Mg: experimental mass of oxygen or $\frac{x}{1.31} = \frac{16}{0.87}$ from which, $M_{\mathrm{Mg}} = 16.00 \times \frac{1.31}{0.87} = 24.1 = 24\,\mathrm{g\,mol^{-1}}$ (to two significant figures).

4 Regularization and Under-Constrained Problems

The atomic weight of magnesium is concluded to be $24\,\mathrm{g\,mol^{-1}}$, as determined by the stoichiometry of its chemical combination with oxygen. This result is in agreement with the accepted value.

5 Dataset Augmentation

The accepted value (periodic table) is $24.3\,\mathrm{g\,mol^{-1}}$ **?**. The percentage discrepancy between the accepted value and the result obtained here is 1.3%. Because only a single measurement was made, it is not possible to calculate an estimated standard deviation.

The most obvious source of experimental uncertainty is the limited precision of the balance. Other potential sources of experimental uncertainty are: the reaction might not be complete; if not enough time was allowed for total oxidation, less than complete oxidation of the magnesium might have, in part, reacted with nitrogen in the air (incorrect reaction); the magnesium oxide might have

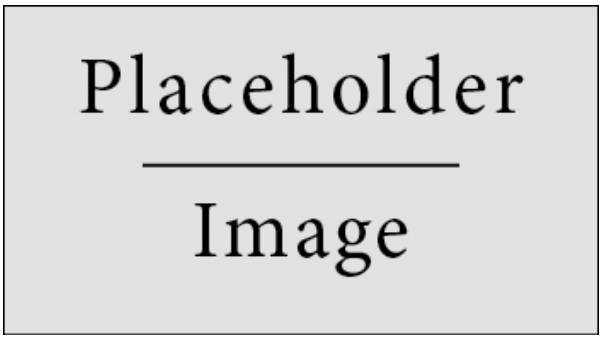

Figure 1: Figure caption.

absorbed water from the air, and thus weigh "too much." Because the result obtained is close to the accepted value it is possible that some of these experimental uncertainties have fortuitously cancelled one another.

6 Noise Robustness

a. The *atomic weight of an element* is the relative weight of one of its atoms compared to C-12 with a weight of 12.0000000..., hydrogen with a weight of 1.008, to oxygen with a weight of 16.00. Atomic weight is also the average weight of all the atoms of that element as they occur in nature.

b. The *units of atomic weight* are two-fold, with an identical numerical value. They are g/mole of atoms (or just g/mol) or amu/atom.

c. *Percentage discrepancy* between an accepted (literature) value and an experimental value is

$$\frac{\text{experimental result} - \text{accepted result}}{\text{accepted result}}$$

7 Semi-Supervised Learning

a. The *atomic weight of an element* is the relative weight of one of its atoms compared to C-12 with a weight of 12.0000000..., hydrogen with a weight of 1.008, to oxygen with a weight of 16.00. Atomic weight is also the average weight of all the atoms of that element as they occur in nature.

b. The *units of atomic weight* are two-fold, with an identical numerical value. They are g/mole of atoms (or just g/mol) or amu/atom.

c. *Percentage discrepancy* between an accepted (literature) value and an experimental value is

$$\frac{\text{experimental result} - \text{accepted result}}{\text{accepted result}}$$

8 Multi-Task Learning

9 Early Stopping

10 Parameter Tying and Parameter Sharing

11 Sparse Representations

12 Bagging and Other Ensemble Methods

13 Dropout

14 Adversarial Training

15 Tangent Distance, Tangent Prop, and Manifold Tangent Classifier