EXPLOVA: Choice of Regularization Parameters by Explanation of Variance

Malik Magdon-Ismail*
magdon@cco.caltech.edu
Caltech Learning Systems Group
Department of Electrical Engineering
California Institute of Technology
136-93 Pasadena, CA, 91125

Xubo Song
xubosong@fire.caltech.edu
Caltech Learning Systems Group
Department of Electrical Engineering
California Institute of Technology
136-93 Pasadena, CA, 91125

Abstract

We develop a new method for choosing the regularization parameter (λ) for general learning models and regularization terms, and provide a condition for regularization to be helpful. Our method (EXPLOVA) chooses the regularization term to reduce the variance of the learned function. When noise is present we find that $\lambda \sim O(\sigma^2/N)$. We present experimental verification of our results using neural networks. We emphasize that effective regularization requires a prior estimate of the noise level and some prior belief about the target function.

1 Introduction

Regularization techniques have been used to improve out of sample performance, typically by smoothing the fit to noisy data. For a survey of well established regularization techniques, see [3] Generally, they correspond to adding a penalty term to the empirical risk, that somehow penalizes the complexity of the learned function.

Unfortunately, there is no obvious way in which to choose the regularization parameter. Possible techniques are cross validation eg. [4] or some other principle such as structural risk minimization [9]. Mackay ([5]) describes a method for choosing regularization parameters from the Bayesian point of view, given a prior distribution for the regularization parameter. However, exactly what a prior on the regularization parameter would be and how one could obtain it are not immediately obvious.

For general learning models and regularization functions, we develop a technique

*To whom correspondence should be addressed.

which we call EXPLOVA (Explanation of Variance)[1] in order to derive an expression for the regularization parameter. Using this technique, we also derive a data dependent condition on the learning model that we expect to be satisfied if a given regularization term is to be helpful, and give examples where we expect regularization to help. Using EXPLOVA, we find that the regularization parameter is $O(\sigma^2/N)$. Our results prescribe weight decay when the data are noisy but not necessarily when the data are noiseless. We demonstrate this experimentally using simulations with neural networks.

2 The Optimal Regularization Parameter

Here, we set up the learning problem and then consider the optimal choice of the regularization parameter. We assume the standard learning paradigm as follows. A continuous target function $f : \mathbf{R}^d \rightarrow \mathbf{R}$ generates the training data set $\mathcal{D}_N$, which consists of N input–output pairs $\{\mathbf{x}_i, y_i\}_{i=1}^N$. Each $\mathbf{x}_i \in \mathbf{R}^d$ is drawn from some input probability measure $dF(\mathbf{x})$. Additive noise is present in the training data, $y_i = f(\mathbf{x}_i) + \epsilon_i$. We further assume that the noise realizations are independent and zero mean, so

$$\langle \epsilon \mid \mathbf{x} \rangle_\epsilon = \mathbf{0}, \qquad \langle \epsilon\epsilon^T \mid \mathbf{x} \rangle_\epsilon = diag[\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2] \tag{1}$$

(we use $\langle \cdot \rangle$ to denote expectations, $\sigma = [\sigma_1 \sigma_2 \ldots \sigma_N]$, and $diag[\cdot]$ denotes a diagonal matrix). The functions $g_\omega \in \mathcal{H}$ are parameterized by a weight vector ω. We use mean squared deviation as our error measure, thus the **training error**, $\mathcal{E}_{tr}$, is defined by

$$\mathcal{E}_{tr} = \frac{1}{N} \sum_{i=1}^{N} (g_\omega(\mathbf{x}_i) - y_i)^2 \tag{2}$$

and the **test error**, $\mathcal{E}_N(\sigma)$, is

$$\mathcal{E}_N(\sigma) = \left\langle (f(\mathbf{x}) - g_\omega(\mathbf{x}))^2 \right\rangle_{\epsilon, \mathcal{D}_N, \mathbf{x}} \tag{3}$$

where the expectation is over possible training data sets, noise and the test point. Our goal is the minimization of the test error, $\mathcal{E}_N(\sigma)$. We consider a regularization term of the form $\lambda\Omega(\omega)$, where λ is the regularization parameter and Ω is the regularizing function. The learning algorithm will be to minimize the function

$$\hat{\mathcal{E}} = \mathcal{E}_{tr}(\omega) + \lambda\Omega(\omega)$$

When $\lambda = 0$ the algorithm becomes minimization of the training error. Intuitively, for non-zero λ, we minimize the training error but penalize complexity (due to attempting to fit the noise) with the regularization term.

We restrict ourselves to *stable learning systems*. Stable learning systems possess the two properties "continuity" and "unbiasedness". Continuity ensures that "close" data sets are mapped to "close" functions. For two data sets differing only by the addition of zero-mean noise, unbiasedness requires that at every point, the average value (with respect to the noise) of the functions resulting from the noisy data set is equal to the value of the function resulting from the noiseless data set. (The formal definitions can be found in [6]). Linear learning models are stable, and it has been experimentally verified that neural networks trained using gradient descent are stable to a high degree of precision.

[1] This name is partly due to Alexander Nicholson.

2.1 Explanation of Variance (EXPLOVA)

The regularization term decreases the complexity of the learned hypothesis. The test error usually improves because the variance of the regularized hypothesis decreases – the regularization should compensate for the extra variance "learned" due to the noise in the data. We try to exploit the fact that with noise, the learned function has higher variance in order to deduce some properties that the regularizer should have. We suppose that $\mathcal{E}_{tr}$ is minimized at ω^*. Due to the addition of the regularization term, the minimum of $\hat{\mathcal{E}}$ will be shifted from ω^* by an amount $\Delta\omega$. This produces a change in the variance of the output for the learned function. The presence of noise adds variance to the targets. This implies that the variance of the learned function will be higher than it should be, in proportion to the level of noise. The presence of the regularization term should offset the variance gained due to the noise, and this change should be related to the noise level. We now quantify this intuition.

Initially the learned function was trained to "explain" as much variance in the data as it could. It is well understood that the presence of noise increases the test error (see for example [2, 8, 6]). Due to the presence of noise, we suppose that the test error increases by an amount $\Delta E(\sigma, N)$. Expanding the expression for the test error we have

$$\begin{aligned}\mathcal{E}_N(\sigma) &= \left\langle (f-g)^2 \right\rangle_{\epsilon,\mathcal{D}_N,\mathbf{x}} \\ &= \left\langle f^2 \right\rangle_x - 2\langle f\langle g\rangle_\epsilon\rangle_{\mathcal{D}_N,\mathbf{x}} + \left\langle g^2 \right\rangle_{\epsilon,\mathcal{D}_N,\mathbf{x}}\end{aligned}$$

The first term is independent of noise, and the second term is equal to the term that would replace it if there were no noise (by the unbiasedness assumption). Therefore $\left\langle g^2 \right\rangle$ must increase by $\Delta E(\sigma, N)$. The presence of noise increases the complexity of the learned function by increasing $\left\langle g^2 \right\rangle_{\epsilon,\mathcal{D}_N,\mathbf{x}}$.

For stable learning systems it is shown in [6] that

$$\Delta E(\sigma, N) \leq \frac{\mathcal{C}\overline{\sigma^2}}{N} + O(\frac{1}{N^2}) \tag{4}$$

where $\mathcal{C}$ is a constant that depends on the target function, input distribution and learning system and $\overline{\sigma^2}$ is the average noise variance. For linear models, $\mathcal{C} = (d+1)$. We would like regularization to enable us to perform as well as we would were there no noise present in the data. Unfortunately all we know is that the noise induces an increase in $\left\langle g^2 \right\rangle$. Thus we know that to be performing as well as in the noiseless case, $\left\langle g^2 \right\rangle$ should decrease by $\Delta E(\sigma, N)$. This is a necessary requirement, but it is not sufficient. In a sense, we are using the fact that $\left\langle g^2 \right\rangle$ should decrease as a hint [1]. We enforce this "hint" not because it guarantees getting closer to the noiseless performance, but because if we are getting closer to the noiseless performance, then the hint will be obeyed. Furthermore, $\langle g \rangle$ should remain unchanged due to the regularizer, because the addition of noise did not change $\langle g \rangle$ in the first place.

The general approach is as follows. The training error will be minimized at some weight vector ω^*. If the regularization parameter is small, then the regularization term will be small. The introduction of the regularization term will cause the minimum to shift to $\omega^* + \Delta\omega$, where $\Delta\omega$ will be small. This change in ω will lead to a change in $\left\langle g_\omega^2 \right\rangle$ which we will equate to $-\Delta\mathcal{E}(\sigma, N) \approx -\mathcal{C}\overline{\sigma^2}/N$.

By taking the Taylor expansion of $\hat{\mathcal{E}}(\omega)$ about ω^* up to second order in $\Delta\omega = \omega - \omega^*$, and using the fact that $\nabla_{\omega^*}\mathcal{E}_{tr} = 0$, we find that

$$\hat{\mathcal{E}}(\omega) = \mathcal{E}_{tr}(\omega^*) + \lambda\Omega(\omega^*) + \frac{1}{2}\Delta\omega^T[H_{\omega^*}(\mathcal{E}_{tr}) + \lambda H_{\omega^*}(\Omega)]\Delta\omega + \lambda\Delta\omega^T\nabla_{\omega^*}\Omega \quad (5)$$

where the Hessian $[H_{\omega^*}(\cdot)]_{ij}{=}\partial_{\omega_i}\partial_{\omega_j}(\cdot)\ |_{\omega^*}$ is the matrix of second derivatives. $H_{\omega^*}(\mathcal{E}_{tr})$ is positive definite and can be calculated from the data given the learning model and ω^* (which can be obtained by optimizing the training error). Setting $\nabla_{\Delta\omega}\hat{\mathcal{E}}(\omega) = 0$ we find for the change in the weight vector induced by the regularization term

$$\begin{aligned} \Delta\omega &\approx -\lambda\left[H_{\omega^*}(\mathcal{E}_{tr}) + \lambda H_{\omega^*}(\Omega)\right]^{-1}\nabla_{\omega^*}\Omega && (6)\\ &= -\lambda H_{\omega^*}^{-1}\nabla_{\omega^*}\Omega + O(\lambda^2) && (7) \end{aligned}$$

where $\nabla_{\omega^*}\Omega \neq 0$. We will ignore the $O(\lambda^2)$ term as λ will be small, and write H_{ω^*} for $H_{\omega^*}(\mathcal{E}_{tr})$. Equating the change in $\langle g^2\rangle$ (due to the change $\Delta\omega$) to $-\mathcal{C}\sigma^2/N$ we can deduce what λ should be. The change in $\langle g^2\rangle$ is approximately given by $\Delta\omega\cdot\nabla_{\omega^*}\langle g^2\rangle$. Using (7) and equating this change to $-\mathcal{C}\overline{\sigma^2}/N$, we have

$$\lambda \quad = \quad \frac{\mathcal{C}\overline{\sigma^2}/N}{2\left(H_{\omega^*}^{-1}\nabla_{\omega^*}\Omega\right)\cdot\langle g\nabla_{\omega^*}g\rangle_{\epsilon,\mathcal{D}_N,\mathbf{x}}} + O(\frac{1}{N^2}) \quad (8)$$

Equating the change in $\langle g\rangle$ to zero (due to unbiasedness), the following condition must be satisfied

$$\left(H_{\omega^*}^{-1}\nabla_{\omega^*}\Omega\right)\cdot\langle\nabla_{\omega^*}g\rangle + O(\lambda^2) = 0 \quad (9)$$

This condition can usually be satisfied by a linear transformation of the inputs. This will become clearer when we consider linear models as an example. We conclude from (8) that the regularization term is $O(\sigma^2/N)$. For a positive regularization parameter to be consistent with a decrease in variance, the necessary condition is that

$$\left(H_{\omega^*}^{-1}\nabla_{\omega^*}\Omega\right)\cdot\langle g\nabla_{\omega^*}g\rangle_{\epsilon,\mathcal{D}_N,\mathbf{x}} \geq 0 \quad (10)$$

hold at $\omega = \omega^*$. Since the regularization term should be applicable to any target function (so, for all ω^*), it is necessary that this inequality be true for all ω. This is a condition that has to be satisfied by a given regularization term and is not independent of the learning model – a regularization term will be useful for a certain learning model, but perhaps not for another – for example, consider two identical learning models, one parameterized by ω, the other parameterized by $1/\omega$; if weight decay helps in one case, it will not help in the other.

For practical purposes, one could use sample estimates of $\langle g\nabla_{\omega^*}g\rangle$ and $\langle\nabla_{\omega^*g}\rangle$. The expression for the regularization parameter λ in (8) is calculated as follows. Optimize the training error to obtain ω^*. Given ω^*, the term (10) can be calculated using the data points. This can then be used to evaluate (8) provided that estimates for $\overline{\sigma^2}$ and $\mathcal{C}$ are available. Descending on $\hat{\mathcal{E}}$, starting from ω^*, we then arrive at the final regularized solution.

EXPLOVA is summarized as follows. Noise is present in the data set, and the noise level is known. Further, one needs an estimate for $\mathcal{C}$, an effective model complexity or an "effective number of parameters" [8]. $\mathcal{C}$ will in general depend on the target function, input distribution and the learning model. We then calculate the value of the regularization parameter by equating the the variance gain in the function due to the presence of noise and the variance loss due to the presence of the regularization term. EXPLOVA is more a philosophy than a specific method, applicable to other techniques. We stress that the noise level needs to be known, because, for given a regularization term, one can find a noise level small enough such that regularization will hurt.

We conclude this section by specializing the preceding analysis to the case of weight decay. For simplicity, we present the analysis for the weight decay term

$$\Omega(\omega) = \frac{1}{2}\omega^T H_{\omega^*}\omega$$

where H_{ω^*} can be computed by first optimizing the training error. In this case, $\nabla_{\omega^*}\Omega(\omega^*) = H_{\omega^*}\omega^*$, therefore using (7), we have

$$\Delta\omega \approx -\lambda\omega^*$$

Using (8), we find for the regularization parameter

$$\lambda \approx \frac{\mathcal{C}\sigma^2/N}{\frac{2}{N}\omega^* \cdot \langle g\nabla_{\omega^*} g\rangle} \tag{11}$$

so, for a positive decay parameter to generate a decrease in variance, the necessary condition is that

$$\omega^* \cdot \langle g\nabla_{\omega^*} g\rangle \geq 0 \tag{12}$$

and (9) becomes $\omega^* \cdot \langle \nabla_{\omega^*} g\rangle = 0$. As already pointed out, this can usually be achieved by transforming the input space. (12) should hold for all ω^*. This is a strong condition on the learning model.

2.2 Examples

2.2.1 General Linear Learning Models

For general linear learning models, $g_\omega = \omega^T\mathbf{z}$ where $\mathbf{z}$ is some fixed transformation of $\mathbf{x}$. The training error is given by

$$\mathcal{E}_{tr} = \frac{1}{N}\sum_{i=1}^{N}(\omega^T\mathbf{z}_i - y_i)^2$$

from which we see that $H_{\omega^*} = 2\hat{\Sigma}$ where $\hat{\Sigma}_{\mathbf{z}} = \mathbf{Z}\mathbf{Z}^T/N$ and we use the definition $\mathbf{Z}\mathbf{Z}^T = \sum_{m=1}^{N}\mathbf{z}_m\mathbf{z}_m^T/N$. Let $\Sigma_{\mathbf{z}} = \left\langle \hat{\Sigma}_{\mathbf{z}}\right\rangle$. The consistency condition (12) becomes

$$(\omega^*)^T\Sigma_{\mathbf{z}}\omega^* \geq 0$$

which is satisfied for all ω^* because $\Sigma_{\mathbf{z}}$ is a positive semi-definite matrix. The sample version of (9) becomes $\omega^* \cdot \sum_{i=1}^{N}\mathbf{z}_i = 0$. This can be achieved by an initial translation of variables such that $\sum_{i=1}^{N}\mathbf{z}_i = 0$. (Note that this will change $\hat{\Sigma}_{\mathbf{z}}$). The regularization term becomes

$$\lambda\Omega(\omega) = \frac{1}{2}\frac{\sigma^2 S}{N\omega^{*T}\hat{\Sigma}_{\mathbf{z}}\omega^*}w^T\hat{\Sigma}_{\mathbf{z}}\omega, \tag{13}$$

similar (up to a factor of 2) to the to the expression obtained by directly minimizing the test error [7].

2.2.2 $g_\omega(\mathbf{x}) = \phi(\omega \cdot \mathbf{x})$

If $\phi' \geq 0$, and $x\phi(x) \geq 0$, then

$$\omega \cdot g\nabla_\omega g = \omega \cdot \mathbf{x}\phi(\omega \cdot \mathbf{x})\phi'(\omega \cdot \mathbf{x}) \geq 0$$

and so the consistency condition is satisfied. eg. $\phi(t) = \tanh(t)$

3 Experiments

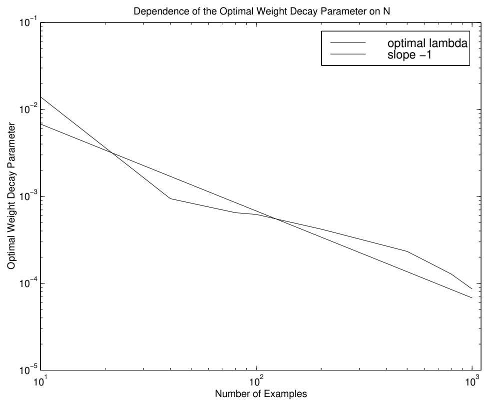

Figure 1: Illustration of the behavior of $\lambda_{optimal}$ on N.

The experiments were run for a teacher 2-5-1 neural network and student 2-10-1 with σ^2=0.3^2. The behavior of the optimal λ minimizing the expected test error is shown in figure 1. A slope -1 line is shown for comparison with the expected $1/N$ decrease which should have slope -1 on this LogLog plot.

4 Comments

We deduced a technique (EXPLOVA) for testing the effectiveness of regularization, which relied upon using the 'hint' that the target function is simpler than indicated by the variance of the targets . This technique offers hope for being able to choose a regularization parameter in some systematic way, given an estimate for the noise level. In the presence of noise, EXPLOVA suggests a regularization parameter of the form

$$\lambda \;=\; \frac{\mathcal{C}\overline{\sigma^2}/N}{\frac{2}{N}\left(H_{\omega^*}^{-1}\nabla_{\omega^*}\Omega\right)\cdot\langle g\nabla_{\omega^*}g\rangle} \tag{14}$$

which is applicable to general learning models and regularization terms. Further, EXPLOVA furnishes a method for evaluating whether or not a given regularization term is useful for a particular learning model. We believe that the philosophy embedded in EXPLOVA bears promise – the existence of noise indicates that the targets have a higher variance than the true function, therefore, we should incorporate this knowledge by requiring that our learned function have smaller variance. Further theoretical and experimental investigation along these lines seems promising.

We emphasize the need for a prior belief on the noise level by noting that for any given regularizer, there is a level of noise small enough such that the regularization will hurt, therefore in order to effectively regularize, estimates of σ^2 and $\mathcal{C}$ are needed.

5 Acknowledgments

We would like to acknowledge Yaser Abu-Mostafa and the Caltech Learning Systems Group for their useful input.

References

[1] Y. Abu-Mostafa. Hints. *Neural Computation*, 4(7):639–671, 1995.

[2] S. Amari and N. Murata. Statistical theory of learning curves under entropic loss criterion. *Neural Computation*, 5:140–153, 1992.

[3] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[4] G. Golub, M. Heath, and G. Wahba. Generalized cross validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–224, 1979.

[5] D. J. C. Mackay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4:448–472, 1992.

[6] M. Magdon-Ismail, A. Nicholson, and Y. S. Abu-Mostafa. Financial markets: very noisy information processing. *To appear In Proceedings of the IEEE Special Issue on Information Processing*, 1998.

[7] M. Magdon-Ismail and X. Song. Regularization in general linear models. *manuscript submitted to NIPS 98*, 1998.

[8] J. E. Moody. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In *NIPS*, volume 4, pages 847–854, 1991.

[9] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer–Verlag, 1995.