Regularization in Generalized Linear Models

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 derive an expression for the optimal weight decay parameter (λ) for generalized linear models. We show that when noise is present, $\lambda \sim O(\sigma^2/N)$ for well specified models. For misspecified models, the degree of misspecification has an effect analogous to noise. 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. Generally, they correspond to adding a penalty term to the empirical risk, that somehow penalizes the complexity of the learned function. Well established techniques include Tikhonov type smoothing regularizers [7], heuristic weight decay regularizers [5], weight elimination [6], soft weight sharing [4]. It is well established that such regularization techniques improve out of sample performance when one attempts to learn from a noisy training set. In the case of noiseless data, it is unclear what kind of performance to expect when using regularization techniques.

Unfortunately, there is no obvious way in which to choose the regularization parameter. Possible techniques are cross validation eg. [1] or some other principle such as structural risk minimization [8]. Mackay ([3]) describes a method for choosing regularization parameters from the Bayesian point of view, given a prior distribution for the regularization parameter. Unfortunately, exactly what a prior on the regularization parameter would be and how one could obtain it are not well defined. It is

*To whom correspondence should be addressed.

shown in [2] that if $\lambda = o(n^{-1/2})$, then the gain in the test error is $o(1/n)$ therefore it is crucial to have a reliable method for choosing λ to take advantage of this small potential gain. Conversely, a poor choice for λ could easily have a negative effect.

When noise is present, attempting to fit the noise will produce a learned function that is more complicated than necessary, so restricting the complexity of the learned function should help the out of sample behavior. By minimizing the out of sample error (test error), we will derive an analytic expression for the optimal weight decay parameter when the learning model is linear in the parameter space (generalized linear models). These results generalize the results of [2] from the one dimensional linear case to the multidimensional general linear case. We will show that for generalized linear models that are well specified, the regularization parameter is $O(\sigma^2/N)$ and the gain in test error is $O(1/N^2)$, analogous to the result in [2] for the one dimensional case. For more general target functions, the regularization parameter turns out to be $O(\sigma^2/N) + A_d$ where A_d is related to the approximation degree (how close the target function is to the learning model). A target function outside the learning model has an effect analogous to noise. Here too we will see that the test error gain is $O(1/N^2)$, and that regularization is considerably less useful when there is no noise.

2 The Optimal Regularization Parameter

In this section, 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 $f : \mathbf{R}^d \to \mathbf{R}$ 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

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

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.

2.1 Generalized Linear Learning Models

Here, we will extend the results of [2] to weight decay for generalized linear learning models. We pick the regularization term so as to minimize the test error directly. Our goal is to analyze the (asymptotic) form of the optimal regularization parameter. We will see that for noisy data, the regularization parameter is $O(\sigma^2/N)$. For noiseless data, there is considerably less gain (and potential loss) in using a regularization term.

Define a set of expansion functions $\{\psi_i(\mathbf{x})\}_{i=1}^{S}$. Thus we map the d dimensional input space into an S dimensional space by $\mathbf{z}_k = \psi_k(\mathbf{x})$. We will assume the existence of all expectations necessary for the following analysis to proceed. Our learning model is linearly parameterized by the weight vector ω:

$$g_\omega(\mathbf{x}) = \sum_{i=1}^{S} w_i \psi_i(\mathbf{x}) = \omega \cdot \mathbf{z} \tag{4}$$

When $\psi_k(\mathbf{x}) = \mathbf{x}_k, k = 1 \ldots d$ and $\psi_{d+1} = 1$, we have standard linear regression. This framework is applicable to general polynomial regression, Fourier series methods, etc. The target vector is given by

$$\mathbf{y}^T = [y_1 \ldots y_N] = \mathbf{f}^T + \epsilon$$

where $\mathbf{f}^T = [f(\mathbf{x}_1) \ldots f(\mathbf{x}_N)]$. Let

$$\mathbf{Z} = [\mathbf{z}_1 \ldots \mathbf{z}_N]$$

be the transformed input matrix. Let $\hat{\Sigma}_\mathbf{z} = \mathbf{Z}\mathbf{Z}^T/N$ and $\hat{\mathbf{g}} = \sum_{i=1}^{N} \mathbf{z}_i f(\mathbf{z}_i)/N$. Let $\Sigma_\mathbf{z} = \langle \hat{\Sigma}_\mathbf{z} \rangle_\mathbf{x}$ and $\mathbf{g} = \langle \hat{\mathbf{g}} \rangle_\mathbf{x}$. To make the analysis more presentable, we assume a regularization term of the form

$$\Omega(\omega) = \lambda \omega^T \hat{\Sigma}_\mathbf{z} \omega$$

Then, $\hat{\mathcal{E}}(w) = \mathcal{E}_{tr}(\omega) + \lambda\Omega(\omega)$ is quadratic in ω. By straight forward differentiation with respect to ω, we find that the weight vector minimizing $\hat{\mathcal{E}}$ is given by

$$\omega(\lambda) = \frac{(\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{y}}{1+\lambda} = \frac{\omega^*}{(1+\lambda)} \tag{5}$$

Where $\omega^* = (\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{y} = \hat{\Sigma}_\mathbf{z}^{-1}(\hat{\mathbf{g}} + \mathbf{Z}\epsilon/N)$. We assume that $\mathbf{g} \neq 0$ else the optimal least squares solution is simply 0 and weight decay will always help. By the law of large numbers, we note that $\mathbf{Z}\mathbf{Z} \underset{N\to\infty}{\longrightarrow} N\Sigma_\mathbf{z}$ and $\mathbf{Z}\mathbf{f} \underset{N\to\infty}{\longrightarrow} N\mathbf{g}$, so we write

$$\mathbf{Z}\mathbf{Z}^T = N\Sigma_\mathbf{z} + \sqrt{N}\mathbf{V}(\mathbf{Z}), \qquad \mathbf{Z}\mathbf{f} = N\mathbf{g} + \sqrt{N}\mathbf{a}(\mathbf{Z}) \tag{6}$$

where $\langle \mathbf{V} \rangle_\mathbf{Z} = \langle \mathbf{a} \rangle_\mathbf{Z} = 0$ and $\mathrm{Var}(\mathbf{V})$ and $\mathrm{Var}(\mathbf{a})$ are $O(1)$, by the central limit theorem[1]. Using (6) and the identity $[1+\lambda\mathbf{A}]^{-1} = 1 - \lambda\mathbf{A} + \lambda^2\mathbf{A}^2 + O(\lambda^3)$, one finds that

$$\hat{\Sigma}_\mathbf{z}^{-1} = \Sigma_\mathbf{z}^{-1} - \frac{\Sigma_\mathbf{z}^{-1}\mathbf{V}\Sigma_\mathbf{z}^{-1}}{\sqrt{N}} + \frac{\Sigma_\mathbf{z}^{-1}\mathbf{V}\Sigma_\mathbf{z}^{-1}\mathbf{V}\Sigma_\mathbf{z}^{-1}}{N} + O\left(\frac{1}{N^{3/2}}\right) \tag{7}$$

Using (4) in (3), and differentiating with respect to ω, we find that the best fit in our model is $w_0 = \Sigma\mathbf{g}$. Let the target function $f(\mathbf{z}) = \omega_0^T\mathbf{z} + \eta(\mathbf{z})$. The maximum

[1] We assume the conditions for this to occur are satisfied.

value of η is a useful measure of how close the target function is to our learning model. Let us define this formally by

Definition 2.1 *The approximation degree (ν_d) of the generalized linear model with respect to the target function and input distribution is given by*

$$\nu_d = \sup_{\mathbf{z}}\{|\eta(\mathbf{z})|\} \tag{8}$$

For good learning models, ν_d will be close to zero. For noiseless data, $\langle\omega^*\rangle \approx \omega_0$. More precisely, $\langle\omega^*\rangle$ converges to ω_0 as $N \to \infty$. The rate of this convergence could also be a useful measure. Formally, we define it as follows.

Definition 2.2 *The convergence degree (c_d) of the generalized linear model with respect to the target function and input distribution is given by*

$$c_d = \sup_k \left\{ k : \langle\omega^*\rangle = \omega_0 + O\left(\frac{1}{N^k}\right)\right\} \tag{9}$$

Examples:

- For a well specified generalized linear model ($\eta(\mathbf{z}) = 0$), after we get S data points, $\omega^* = \omega_0$ therefore $c_d = \infty$.
- $c_d = 2$ implies that $\langle w^*\rangle = w_0 + O(1/N^2)$.
- It can be shown that in general, $c_d \geq 1$ and that if $\left\langle \mathbf{V}\Sigma_{\mathbf{z}}^{-1}\mathbf{V}\right\rangle\Sigma_{\mathbf{z}}^{-1}\mathbf{g} - \Sigma_{\mathbf{z}}^{-1}\left\langle \mathbf{V}\Sigma_{\mathbf{z}}^{-1}\mathbf{a}\right\rangle = 0$, then $c_d \geq 3/2$.

2.1.1 Well Specified Learning Model

The target function is given by $f = \omega_0 \cdot \mathbf{z}$. Thus, $\mathbf{y} = \mathbf{Z}^T\omega_0 + \epsilon$, therefore, $\omega^* = \omega_0 + \hat{\Sigma}_{\mathbf{z}}^{-1}\mathbf{Z}\epsilon/N$. Using (3) and (5), we get for the test error

$$E_{test}(\lambda) = \frac{\lambda^2}{(1+\lambda)^2}\omega_0\Sigma_{\mathbf{z}}\omega_0 + \frac{\sigma^2 \mathrm{Tr}[\left\langle \hat{\Sigma}_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z}}\right\rangle]}{N(1+\lambda)^2} \tag{10}$$

where we used the facts $\left\langle \epsilon^T\mathbf{A}\epsilon\right\rangle_\epsilon = \sigma^2\mathrm{Tr}[\mathbf{A}]$ and $\left\langle \epsilon^T A\right\rangle = 0$ where A is independent of ϵ. When $\sigma^2 = 0$ it is no surprise that any $\lambda \neq 0$ will hurt. In fact it is easy to convince oneself that for any $\lambda > 0$, $\exists\sigma^2 > 0$ such that regularization with that value of λ will hurt. By setting $\partial_\lambda E_{test} = 0$, we find

$$\begin{aligned}
\lambda_{optimal} &= \frac{\sigma^2\mathrm{Tr}[\left\langle \hat{\Sigma}_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z}}\right\rangle]}{N\omega_0\Sigma_{\mathbf{z}}\omega_0} \\
&\stackrel{(a)}{=} \frac{\sigma^2 S}{N\omega_0\Sigma_{\mathbf{z}}\omega_0} + O\left(\frac{1}{N^2}\right) \\
&\stackrel{(b)}{\approx} \frac{\sigma^2 S}{N\omega^*\hat{\Sigma}_{\mathbf{z}}\omega^*} + o\left(\frac{1}{N}\right)
\end{aligned} \tag{11}$$

where (a) follows because $\left\langle \hat{\Sigma}_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z}}\right\rangle = \mathbf{I}_{S\times S} + \left\langle \Sigma_{\mathbf{z}}^{-1}\mathbf{V}\Sigma_{\mathbf{z}}^{-1}\mathbf{V}\right\rangle/N + o(1/N)$ and (b) because $\left\langle (\omega^*)^T\hat{\Sigma}_{\mathbf{z}}\omega^*\right\rangle = \omega_0^T\Sigma_{\mathbf{z}}\omega_0 + O(\sigma^2/N)$. Using (11) in (10) we find for the improvement in the test error, $\Delta E_{test} = E_{test}(\lambda) - E_{test}(0)$,

$$\Delta E_{test} = -\frac{3\sigma^4 \mathrm{Tr}[\left\langle \hat{\Sigma}_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z}}\right\rangle]^2}{N^2 \omega_0 \Sigma_{\mathbf{z}} \omega_0} + o\left(\frac{1}{N^2}\right) = O\left(\frac{\sigma^4}{N^2}\right) \tag{12}$$

Technical Note

For input dependent noise that is independent and has zero conditional mean, the test error can be calculated as

$$E_{test} = \frac{\lambda^2}{(1+\lambda)^2}\omega_0 \Sigma_{\mathbf{z}} \omega_0 + \frac{\left\langle \sigma^2(\mathbf{z}_i)\mathbf{Z}_{ji}\mathbf{Z}_{il}[\hat{\Sigma}_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z}}\hat{\Sigma}_{\mathbf{z}}^{-1}]_{lj}\right\rangle}{N(1+\lambda)^2} \tag{13}$$

where summation over repeated indices is understood; to $O(1/N)$ this expression becomes

$$E_{test} = \frac{\lambda^2}{(1+\lambda)^2}\omega_0 \Sigma_{\mathbf{z}} \omega_0 + \frac{\mathrm{Tr}[\left\langle \sigma^2(\mathbf{z})\mathbf{z}\mathbf{z}^T\right\rangle \Sigma_{\mathbf{z}}^{-1}]}{N(1+\lambda)^2} \tag{14}$$

We can then proceed to obtain $\lambda_{optimal}$ in exactly the same way as has already been used above when the noise is input independent.

2.1.2 Misspecified Learning Model

In this case, the target function is not in the learning model. We can still make some progress because the learned hypothesis can be explicitly constructed. Using (3) and (5) the test error can be calculated as

$$E_{test} = \frac{\left\langle \hat{\mathbf{g}}^T\hat{\Sigma}_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z}}\hat{\Sigma}_{\mathbf{z}}^{-1}\hat{\mathbf{g}} + \frac{\sigma^2}{N}\mathrm{Tr}[\Sigma_{\mathbf{z}}\hat{\Sigma}_{\mathbf{z}}^{-1}]\right\rangle}{(1+\lambda)^2} - \frac{2\left\langle \hat{\mathbf{g}}^T\hat{\Sigma}_{\mathbf{z}}^{-1}\mathbf{g}\right\rangle}{1+\lambda} + \left\langle f^2\right\rangle$$

We can write this in a form similar to (10) as follows

$$E_{test} = \frac{\lambda^2\left\langle \hat{\mathbf{g}}^T\hat{\Sigma}_{\mathbf{z}}^{-1}\mathbf{g}\right\rangle}{(1+\lambda)^2} + \frac{\frac{\sigma^2}{N}tr[\left\langle \Sigma_{\mathbf{z}}\hat{\Sigma}_{\mathbf{z}}^{-1}\right\rangle] + A_d}{(1+\lambda)^2} + \left\langle f^2\right\rangle - \left\langle \hat{\mathbf{g}}^T\hat{\Sigma}_{\mathbf{z}}^{-1}\mathbf{g}\right\rangle \tag{15}$$

Where

$$A_d = \left\langle \hat{\mathbf{g}}^T\hat{\Sigma}_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z}}\hat{\Sigma}_{\mathbf{z}}^{-1}\hat{\mathbf{g}} - \hat{\mathbf{g}}^T\hat{\Sigma}_{\mathbf{z}}^{-1}\mathbf{g}\right\rangle$$

A_d deviates from zero to the extent that $\Sigma_{\mathbf{z}}\hat{\Sigma}_{\mathbf{z}}^{-1}\hat{\mathbf{g}}$ deviates from $\mathbf{g}$, thus A_d is small and in fact decreases at least as fast as $1/N$. A more detailed analysis of A_d is given in the appendix, where it is shown that $A_d \approx \left\langle \mathrm{Tr}[\eta(\mathbf{z})\mathbf{z}\mathbf{z}^T\Sigma_{\mathbf{z}}^{-1}]\right\rangle$. Setting $\partial_\lambda E_{test}$=0 we solve for $\lambda_{optimal}$.

$$\lambda_{optimal} = \frac{\frac{\sigma^2}{N}\left\langle \mathrm{Tr}[\Sigma_{\mathbf{z}}\hat{\Sigma}_{\mathbf{z}}^{-1}]\right\rangle + A_d}{\hat{\mathbf{g}}^T\Sigma_{\mathbf{z}}^{-1}\mathbf{g}} \tag{16}$$

$$\stackrel{(a)}{\approx} \frac{\sigma^2\left\langle \mathrm{Tr}[\Sigma_{\mathbf{z}}\hat{\Sigma}_{\mathbf{z}}^{-1}]\right\rangle + \left\langle \mathrm{Tr}[\eta(\mathbf{z})^2\mathbf{z}\mathbf{z}^T\Sigma_{\mathbf{z}}^{-1}]\right\rangle}{N\hat{\mathbf{g}}^T\Sigma_{\mathbf{z}}^{-1}\mathbf{g}} \tag{17}$$

where (a) follows from (24) in appendix 4. Note that $\hat{\Sigma}_{\mathbf{z}}^{-1}\hat{\mathbf{g}} = \omega^*$ so as $\hat{\mathbf{g}}^T\Sigma_{\mathbf{z}}^{-1}\mathbf{g} = {w^*}^T\hat{\Sigma}_{\mathbf{z}}\omega^* + O(1/\sqrt{N})$. Notice that the extra term induced by the misspecification is exactly what would arise if instead we added noise to the data with a variance equal to $\eta^2(\mathbf{z})$, dependent on $\mathbf{z}$ (14). Using our expression for $\lambda_{optimal}$ in (15), we see that the gain in test error performance is

$$\Delta E_{test} = -\frac{3(\frac{\sigma^2}{N}\mathrm{Tr}[\left\langle\hat{\Sigma}_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z}}\right\rangle] + A_d)^2}{\hat{\mathbf{g}}^T\Sigma_{\mathbf{z}}^{-1}\mathbf{g}} + o\left(\frac{1}{N^2}\right) \tag{18}$$

which is $O(1/N^2)$. If $c_d > 1$ then it can be shown that $A_d = o(1/N)$ and so can be ignored (in the asymptotic limit) in the expression for $\lambda_{optimal}$. Otherwise A_d can be bounded in terms of ν_d, therefore, for good learning models, A_d will be small.

3 Comments

We analyzed regularization (specifically weight decay) for generalized linear models (G.L.M's). We found an explicit formula for $\lambda_{optimal} = O(\frac{\sigma^2+n^2}{N})$, where the n^2 arises from the component of the target function outside the learning model when the G.L.M. is misspecified – the component of the target function outside the learning model has an effect equivalent to input dependent noise. When the data is noiseless, regularization is of limited use (especially for $c_d > 1$ or $\nu_d \ll 1$), contrary to the idea that regularization will help by curtailing the model complexity. If there is nothing to over fit, then the model will not over fit.

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 (where now we treat a target function outside the learning model as a noisy target function in the learning model) such that the regularization will hurt. Our results depended on parameters related to the target function. The gains due to regularization are $O(1/N^2)$ for the *optimal* regularization parameter, so it is important to have an accurate estimate for $\lambda_{optimal}$ in order to avoid potential disaster. To this end, we need σ^2 and A_d (for G.L.M.'s) and w_0. These quantities can only be calculated given some prior belief about the target function.

4 Appendix

Here, we will present a detailed analysis of A_d, given by

$$A_d = \left\langle \hat{\mathbf{g}}^T\hat{\Sigma}_{\mathbf{z}}^{-1}\Sigma_{\mathbf{z}}\hat{\Sigma}_{\mathbf{z}}^{-1}\hat{\mathbf{g}} - \hat{\mathbf{g}}^T\hat{\Sigma}_{\mathbf{z}}^{-1}\mathbf{g}\right\rangle \tag{19}$$

Using (7) and (6), we find that

$$A_d = \frac{\langle 2\mathbf{g}^T\Sigma_{\mathbf{z}}^{-1}\mathbf{V}\Sigma_{\mathbf{z}}^{-1}\mathbf{V}\Sigma_{\mathbf{z}}^{-1}\mathbf{g} + \mathbf{a}^T\Sigma_{\mathbf{z}}^{-1}\mathbf{a} - 3\mathbf{g}^T\Sigma_{\mathbf{z}}^{-1}\mathbf{V}\Sigma_{\mathbf{z}}^{-1}\mathbf{a}\rangle}{N} + O\left(\frac{1}{N^{3/2}}\right) \tag{20}$$

Thus, A_d is $O(1/N)$. We can develop a more intuitive understanding of A as follows. Let $f(\mathbf{z}) = \mathbf{z}^T\omega_0 + \eta(\mathbf{z})$, where the best fit ω_0 is given by $\omega_0 = \Sigma_{\mathbf{z}}^{-1}\mathbf{g} = \Sigma_{\mathbf{z}}^{-1}\langle\mathbf{z}f(\mathbf{z})\rangle$. Consistency requires that $\Sigma_{\mathbf{z}}^{-1}\langle\mathbf{z}\eta(\mathbf{z})\rangle = 0$, so $\langle\mathbf{z}\eta(\mathbf{z})\rangle = 0$. Let

$$\hat{\mathbf{g}} = \frac{\mathbf{Zy}}{N} = \hat{\Sigma}_{\mathbf{z}}\omega_0 + \xi \tag{21}$$

where we have defined $\xi = \mathbf{Z}\eta/N = \sum_i \mathbf{z}_i \eta(\mathbf{z}_i)$ is an $S \times 1$ vector. $\langle \xi \rangle = 0$ and the covariance matrix for ξ can be calculated as

$$\langle \xi \xi^T \rangle = \frac{\langle \mathbf{z}\mathbf{z}^T \eta^2(\mathbf{z}) \rangle_{\mathbf{z}}}{N} \tag{22}$$

The expression for A_d can now be reduced to

$$A_d = \left\langle \xi^T \hat{\Sigma}_{\mathbf{z}}^{-1} \Sigma_{\mathbf{z}} \hat{\Sigma}_{\mathbf{z}}^{-1} \xi \right\rangle + \left\langle \omega_0 \Sigma_{\mathbf{z}} \hat{\Sigma}_{\mathbf{z}}^{-1} \xi \right\rangle \tag{23}$$

We ignore the second term because it only involves ξ. Also by construction, ω_0 is approximately orthogonal to ξ so if $\Sigma_{\mathbf{z}} \hat{\Sigma}_{\mathbf{z}}^{-1} \approx \mathbf{I}$ then this is justified[2]. The first term can be simplified by taking the trace. To order $1/N$, we then find

$$A_d = \frac{\mathrm{Tr}[\langle \eta(\mathbf{z})^2 \mathbf{z}\mathbf{z}^T \rangle \Sigma_{\mathbf{z}}^{-1}]}{N} \tag{24}$$

Notice that this is equivalent to the term that would arise were there input dependent noise with noise variance $\eta(\mathbf{z})^2$.

5 Acknowledgments

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

References

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

[2] P. Koistinen. Asymptotic theory for regularization: One-dimensional linear case. In *Advances in Neural Information Processing Systems (NIPS)*, volume 10, 1997.

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

[4] S. Nowlan and G. Hinton. Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4(4):473–493, 1992.

[5] D. Plaut, S. Nowlan, and G. Hinton. Experiments on learning by backpropagation. Technical Report CMU-CS-86-126, Carnegie Mellon University, 1986.

[6] R. Scalettar and A. Zee. Emergence of grandmother memory in feedforward networks: Learning with noise and forgetfulness. In D. Waltz and J. Feldman, editors, *Connectionist Models and Their Implications: Readings from Cognitive Science*. Ablex Pub. Corp, 1988.

[7] A. N. Tikhonov and V. I. Arsenin. *Solutions of Ill-Posed Problems*. Scripta Series in Mathematics. Distributed solely by Halsted Press, Winston; New York, 1977. Translation Editor: Fritz, John.

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

[2]Technically, one can show that the $O(1/N)$ coefficient is $\omega_0^T \langle \mathbf{V} \Sigma_{\mathbf{z}}^{-1} \alpha \rangle$ where alpha is a random variable with the same statistics as $\sqrt{N}\xi$. However, we expect that $\mathbf{V}$ should be uncorrelated with α, so we can ignore this term.