Regularization with a pruning prior

Cyril Goutte[*,†,§] Lars Kai Hansen[‡,§]

Department of Mathematical Modelling, Bygn. 349
Denmark's Technical University
DK-2800 Lyngby, Denmark.

Running title: Pruning prior.

---

[*]Requests for reprint should be sent to Cyril Goutte, phone: +45 4525 5738, fax: +45 4288 0117, e-mail: goutte@ei.dtu.dk

[†]Cyril Goutte is a visiting student, 1995-1996, with a grant from the Danish Rector's Conference. Permanent address: Équipe connexionniste - LAFORIA, Université Paris VI, 4 place Jussieu, 75252 Paris Cedex 05, France. Email: goutte@laforia.ibp.fr

[‡]This research is supported by the Danish Research Councils for the Natural and Technical Sciences through the Danish Computational Neural Network Center.

[§]Both authors wish to acknowledge stimulating discussions with Jan Larsen.

Regularization with a pruning prior

Abstract

We investigate the use of a regularization prior and its pruning properties. We illustrate the behavior of this prior by conducting analyses both using a Bayesian framework and with the generalization method, on a simple toy problem. Results are thoroughly compared with those obtained with a traditional weight decay.

Keywords: Regularization, Generalization method, Bayesian learning, Evidence framework, Laplace prior, Comparison with weight decay.

List of symbols

w parameter.

$\hat{w}$ parameter estimate.

$\tilde{w}$ teacher parameter.

ξ regularization parameter.

$S(w)$ training error.

$C(w)$ regularized training error.

$E(w)$ generalization error.

$\mathcal{D}$ data set.

N number of examples = card($\mathcal{D}$).

y_i individual example, $i \in \{1 \ldots N\}$.

$\bar{y}$ empirical mean of the examples.

σ^2 variance of the examples.

σ_y^2 empirical variance of the examples $= \sum_{i=1}^N (y_i - \bar{y})^2 / N$.

α hyper-parameter.

w_{ML} maximum likelihood estimator.

w_{MP} most probable estimator.

w_{G} best generalization estimator.

$\Phi(x)$ area under the normal curve.

1 Gauss and Laplace priors

Regularization by weight decay, and the associated Bayesian interpretation involving a Gaussian prior is a well established part of modern neural computation, see e.g., (Krogh and Hertz, 1992). (Hansen and Rasmussen, 1994) show that when the weight decay parameter is determined by data—either using MacKay's evidence procedure, or by minimizing generalization error—*pruning* may result. Recently, (Williams, 1995) elucidated the interesting properties of the Laplace prior; in particular he showed that this prior directly results in pruning some weights, for any fixed value of the prior "strength". Here we investigate the statistical properties of the Laplace prior, using the toy model of (Hansen and Rasmussen, 1994) to compare its regularization and pruning properties.

If we consider a system producing input output pairs (x_i, y_i) which we try to approximate using, say, a neural network performing a mapping f_w parameterized by a "weight vector" w, the cost function to minimize is:

$$C(w) = S(w) + \xi . R(w) \tag{1}$$

$R(w)$ is the regularization function: it is the squared L_2 norm in the case of a weight decay, and the L_1 norm for Laplace regularization, i.e. $R(w) = \sum |w_j|$. $S(w)$ is the data misfit and $C(w)$ then represents the cost function subject to the constraint R. As noted in (Williams, 1995), both the Laplace and Gauss distribution are maximum entropy distributions, either for a positive quantity constrained on its mean (Laplace) or for a quantity constrained on its mean and variance (Gauss).

Some insight on the effect of a Laplace prior can be gained by studying the behavior of (1) at its minimum $\hat{w}$. The sensitivity of the data misfit to a non-zero parameter w_j is:

$$\left| \frac{\partial S}{\partial w_j}(\hat{w}) \right| = \xi \qquad |w_j| > 0 \tag{2}$$

This means that the sensitivity does not change with the parameter's value, as is the case with a *weight decay*. This is a crucial difference, as it explains why this prior deserves the name of "pruning prior". A useless parameter induces a sensitivity level related only to the noise, thus very small. Let us consider what happens when ξ grows. Without regularization, weights are optimized so that the "sensitivities" $\frac{\partial S}{\partial w_j}$ are all zero. When ξ increases, the solution is drawn away from the unregularized solution towards 0. The optimal regularized weights solve the tradeoff in (2). If there is a place where the sensitivity w.r.t. w_j can match ξ, equation (2) can be fulfilled; otherwise its value is set to 0.

As a result, weights under a common regularization constraint ξ fall in two categories. If the sensitivity of the data misfit to a given parameter achieves a certain level (linked to ξ), they are adapted with a *uniform* regularization constraint; if not, they are pruned out of the network.

We now proceed to an analysis of Laplace regularization on a simple problem, for which detailed insight can be obtained, that supplements the analysis of (Williams, 1995). The following study is conducted along the lines of (Hansen and Rasmussen, 1994; MacKay, 1992). (Williams, 1995) also uses a Bayesian framework—though in a more general situation—with the so-called "integrate out" approach. The end of section 2 features a comparison with his results. We will use the same toy problem as in (Hansen and Rasmussen, 1994): it consists in estimating the mean of a Gaussian variable of known variance. We consider a *teacher parameter* $\tilde{w}$ providing a set of N examples:

$$y_i = \tilde{w} + \epsilon_i \qquad \epsilon_i \in \mathcal{N}(0, \sigma^2) \qquad i = 1 \ldots N$$

where the variance σ^2 is known. This problem is extremely simple compared to, e.g. a full neural network. Furthermore, it is a typical location parameter problem, and many authors have noted that location parameters such as biases in neural networks should not be regularized. Nevertheless, we choose it as an *illustrative* problem for three main reasons: it allows for a full calculation of the quantities involved in the frameworks studied here with no approximation; even though simple, (Hansen and Rasmussen, 1994) have shown that this problem leads to nontrivial decision processes; and lastly, it fulfills convincingly its illustrative purpose by demonstrating clearly the effect of this prior on learning.

2 Bayesian analysis

The *maximum likelihood* estimator is obtained by maximizing the likelihood of the parameter associated with the training set $\mathcal{D} = \{y_i | i = 1..N\}$:

$$P(\mathcal{D}|w) = \prod_{i=1}^{N} \frac{\exp\left[-\frac{(y_i-w)^2}{2\sigma^2}\right]}{\sqrt{2\pi\sigma^2}} \tag{3}$$

Maximizing $P(\mathcal{D}|w)$ with respect to w shows that the maximum likelihood estimator is, as we know, the empirical mean:

$$w_{\mathrm{ML}} = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y} \tag{4}$$

The probability of a parameter conditioned on the data and the hyperparameter α is obtained using "Bayes' rule"[1]:

$$P(w|\mathcal{D},\alpha) = \frac{P(\mathcal{D}|w,\alpha)P(w|\alpha)}{P(\mathcal{D}|a)} \tag{5}$$

The prior here is the Laplace distribution:

$$P(w|\alpha) = \frac{\alpha}{2}\mathrm{e}^{-\alpha|w|} \tag{6}$$

where α is the inverse of the mean absolute value of the parameter. Using (5), (3) and (6) gives the *posterior* distribution:

$$P(w|\mathcal{D},\alpha) = \frac{\alpha \mathrm{e}^{-\alpha|w|}}{2P(\mathcal{D}|\alpha)} \prod_{i=1}^{N} \frac{\mathrm{e}^{-\frac{(y_i-w)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \tag{7}$$

[1] According to (Jaynes, 1985), Laplace rediscovered this rule and used it extensively for years.

Maximizing (7) with respect to w leads to the *most probable* parameter:

$$w_{\mathrm{MP}} = \begin{cases} \left(1 - \frac{\alpha\sigma^2}{N|w_{\mathrm{ML}}|}\right) w_{\mathrm{ML}} & |w_{\mathrm{ML}}| > \frac{\alpha\sigma^2}{N} \\ 0 & |w_{\mathrm{ML}}| \leq \frac{\alpha\sigma^2}{N} \end{cases} \tag{8}$$

NOTE: It should now be obvious that (6) is a *pruning prior*. In the case of a *weight decay*, pruning sometimes occurs when the most likely value of the hyper-parameter is infinity (see e.g. Hansen & Rasmussen (1994)). Pruning then arises from the hyper-parameter, but not from the prior. This requires one individual hyper-parameter per weight. On the other hand, the Laplace prior *alone* provides effective pruning for all parameters under a certain sensitivity, as shown by (8). One hyper-parameter can encompass several weights, and adapting its value allows to tune the extent of the pruning.

The hyper-parameter α will be set by maximizing the "evidence". This is done by invoking Bayes' rule again:

$$P(\alpha|\mathcal{D}) = \frac{P(\mathcal{D}|\alpha)P(\alpha)}{P(\mathcal{D})} \tag{9}$$

We assume a flat prior on $\log\alpha$, since α is a scale parameter on which we have no *a priori* opinion. The most likely value of α is then reached by maximizing the evidence:

$$P(\mathcal{D}|\alpha) = \int_{-\infty}^{+\infty} P(\mathcal{D}|w,\alpha)P(w|\alpha)\; dw \tag{10}$$

This integral leads to the intricate expression:

$$P(\mathcal{D}|\alpha) = Z_p \frac{\alpha}{2} \left[\mathrm{e}^{\frac{A^2}{2}} \Phi(A) + \mathrm{e}^{\frac{B^2}{2}} \Phi(-B) \right] \tag{11}$$

where Z_p is a constant with respect to α, $A = \frac{\sqrt{N}}{\sigma}\left(w_{\mathrm{ML}} - \frac{\alpha\sigma^2}{N}\right)$ and $B = \frac{\sqrt{N}}{\sigma}\left(w_{\mathrm{ML}} + \frac{\alpha\sigma^2}{N}\right)$. Φ is the "area under the normal curve":

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\, dt$$

(11) can only be maximized numerically, which results in a numerical estimate of α_{ML}. This value is then used together with w_{ML} in (8) to

obtain the optimal w_{MP}. We will now refer to this value as the most probable estimator. (Hansen and Rasmussen, 1994) derive the expression of the same estimator in the Gaussian case.

Let us now perform a brief comparison. (Williams, 1995) uses the so-called "integrate out" approach (Buntine and Weigend, 1991) to propose optimal settings for the scaling and regularization parameter. Using the same method on our toy problem with a non-informative prior over the regularization parameter leads to:

$$-\ln P(w|\mathcal{D}) = \frac{N(\bar{y}-w)^2}{2\sigma^2} + \ln|w| + Z \tag{12}$$

where Z is a constant with respect to w. Maximizing (12) leads to an estimator, the "brave generalization student" of (Hansen and Rasmussen, 1994):

$$w_{\mathrm{BG}} = \begin{cases} \frac{w_{\mathrm{ML}}}{2}\left(1+\sqrt{1-\frac{4\sigma^2}{N{w_{\mathrm{ML}}}^2}}\right) & {w_{\mathrm{ML}}}^2 > 4\frac{\sigma^2}{N} \\ 0 & {w_{\mathrm{ML}}}^2 \leq 4\frac{\sigma^2}{N} \end{cases} \tag{13}$$

It is clear in this case that the result of both frameworks are slightly different, as will be shown by the comparison at the end of the paper[2]. The results however do not display radical differences. In particular, the two procedures lead to the two largest decreases in generalization error for small teacher weights (cf. section 4). This brings us to the conclusion that, in this case, maximizing the evidence provides a good estimator.

It should be noted that the usual "evidence procedure" involves a Gaussian approximation in order to perform the integration over weight space in (10). This approximation has been criticized as potentially unreliable (Wolpert, 1995) and was discarded here in favor of the exact calculation. In our case, the Gaussian approximation results in an estimator that is slightly more conservative than (13).

[2]where the "brave generalization student" is labelled "GeneralizationB"

3 Generalization method

For a given *student parameter* w trying to estimate $\tilde{w}$ with the training set as the only information, the training error is:

$$S(w) = \frac{1}{N}\sum_{i=1}^{N}(y_i - w)^2 = (w - \bar{y})^2 + \sigma_y^2 \tag{14}$$

where $\bar{y}$ and σ_y are the empirical average and standard deviation of the examples, respectively. The measure of success of the *student parameter* is the generalization error:

$$E(w) = \int (\tilde{w} + \epsilon - w)^2 P(\epsilon)\, d\epsilon = (\tilde{w} - w)^2 + \sigma^2 \tag{15}$$

The likelihood detailed in (3) is linked to the data misfit S and the prior $P(w|\alpha)$ over the weight is related to the regularization constraint R in (1). We have $-\ln\left(P(\mathcal{D}|w,\alpha)\right) \propto S(w)$ and $-\ln\left(P(w|\alpha)\right) \propto R(w)$. This links (1) and (5) through (3) and (6), as:

$$\begin{aligned} -\ln P(w|\mathcal{D},\alpha) &\propto -\ln P(\mathcal{D}|w) - \ln P(w|\alpha) \\ &\propto S(w) + \xi . R(w) \end{aligned}$$

Thus, maximizing the posterior probability $P(w|\mathcal{D},\alpha)$ is equivalent to minimizing the regularized cost $C(w)$. Minimizing (14) shows immediately that the unregularized solution is the empirical mean $\bar{y}$, labelled above the *maximum likelihood* estimator. Averaging both (14) and (15) for this estimator with respect to all possible training sets amounts to integrating with respect to $P(\bar{y})$:

$$\langle S \rangle = \left(1 - \frac{1}{N}\right)\frac{\sigma^2}{2} \qquad \langle E \rangle = \left(1 + \frac{1}{N}\right)\frac{\sigma^2}{2} \tag{16}$$

Equation (16) reflects the standard FPE (Akaike, 1969) calculation in our special, one-parameter case.

The regularized cost (1) using the Laplace prior is now[3]:

$$C(w) = (w - \bar{y})^2 + \sigma_y^2 + 2\xi|w| \tag{17}$$

[3]for convenience, we here use 2ξ instead of ξ as in (1)

Minimizing (17) with respect to w produces the regularized optimum:

$$w_{\mathrm{G}} = \begin{cases} \left(1 - \frac{\xi}{|w_{\mathrm{ML}}|}\right) w_{\mathrm{ML}} & |w_{\mathrm{ML}}| > \xi \\ 0 & |w_{\mathrm{ML}}| \leq \xi \end{cases} \tag{18}$$

The link between ξ and α from section 2 is made obvious by comparing (8) and (18): $\xi = \alpha\sigma^2/N$.

The effect of the pruning prior is illustrated in figure 1. As the regularization level increases, the optimal weight (minimum of the transversal curve on the surface) decreases very regularly until it reaches 0, where it is effectively "pruned".

Figure 2 illustrates the effect of our regularizing functional on this simple case and allows for comparison with the *weight decay*. We see that Laplace regularization prunes some weights and gives a constant bias to the remaining ones. As noted above, pruning arises directly from the prior. On the other hand, a finite weight-decay never prunes and gives highly biased estimates for high values of the maximum likelihood (e.g. big teacher parameter). All quantities in figure 2 have been scaled by $\sqrt{\sigma^2/N}$. Let us carry out further comparison by considering not just a common level of regularization, but the "optimal" regularization level in each case.

Section 2 presented the *most probable* solution. We will now proceed to the solution that corresponds to the largest expected decrease in the generalization error, i.e. the *best* solution w.r.t. generalization. For this purpose, we plug (18) into (15) and average E over all possible training sets as above. The average generalization error is now:

$$\begin{aligned} \langle E \rangle &= \left(1 + \frac{1}{N}\right)\sigma^2 + \xi^2 && (19) \\ &+ \frac{\sigma^2}{N}\left[(AB-1)\left(\Phi(B) - \Phi(A)\right)\right. \\ &+ \left.\frac{1}{\sqrt{2\pi}}\left(A\mathrm{e}^{-\frac{B^2}{2}} - B\mathrm{e}^{-\frac{A^2}{2}}\right)\right] && (20) \end{aligned}$$

where $A = \frac{\sigma}{\sqrt{N}}(\tilde{w} - \xi)$ and $B = \frac{\sigma}{\sqrt{N}}(\tilde{w} + \xi)$.

Predictably, this expression can only be minimized numerically. Furthermore, $\tilde{w}$ being unknown, it has to be approximated. The obvious choice is to replace $\tilde{w}$ with the maximum likelihood estimator w_{ML}. Numerical minimization[4] leads to an estimate of the optimal value for ξ, which is in turn plugged into (18) to obtain w_{G}. w_{G} can then be used as a new approximation for w_{ML} in (20), and the optimization iterated to obtain yet another estimate of the teacher parameter which, in turn, can be used in (20). Unfortunately, the numerous optimizations necessary make this scheme extremely time-consuming.

4 Results

We now have two methods (Bayesian inference and generalization method) to produce estimators of the teacher weight. How well do they fare, and how do they compare to similar methods in the case of a weight decay?

Figure 3 presents w_{MP} and w_{G} as a function of w_{ML}, together with the most probable estimator in the case of weight decay. It can be seen that the big advantage of w_{G} over the other two is that the bias is reduced extremely fast, so that it is almost nonexistent for weights of intermediate to big sizes. On the other hand the width of the pruning is smaller. The "Bayesian" estimator w_{MP} has a much broader pruning section, so we can expect a bigger decrease in error around 0, but it is no better at reducing the bias than the equivalent estimator in the case of a weight decay.

These estimators depend on w_{ML} and thus indirectly on the training set. We can easily compute the average excess generalization error, i.e. the error on top of the irreducible noise level σ^2. The average excess error for

[4] the results presented here have been obtained using a simple golden search method (Press et al., 1992).

generic estimator E, $\Delta E(\tilde{w}) = \langle E\left(w_{\mathrm{E}}\right) - \sigma^2\rangle$, is then:

$$\Delta E(\tilde{w}) = \sqrt{\frac{N}{2\pi\sigma^2}} \int_{-\infty}^{+\infty} (\tilde{w} - w_{\mathrm{E}})^2 \, \mathrm{e}^{-\frac{N(w_{\mathrm{ML}}-\tilde{w})^2}{2\sigma^2}} \, dw_{\mathrm{ML}} \tag{21}$$

For the maximum likelihood estimator the excess error, according to (16), is σ^2/N, which is scaled to 1. When the scaled $\Delta E(\tilde{w})$ exceeds 1, our estimator E performs worse, on average, than the maximum likelihood. On the other hand, when it is below 1, we can expect a lower generalization error with E for this value of $\tilde{w}$.

This integral is computed numerically for all estimators. The results are presented in figure 4. We also present for comparison the results obtained in the case of a weight decay. Curves corresponding to Bayesian estimators are labelled 'Most probable', while curves obtained with the generalization method are labelled 'Generalization'. In the case of the weight-decay (Gaussian prior), two estimators have been derived with the generalization method. The first one, GeneralizationA, is obtained with the same approximation $\tilde{w} \approx w_{\mathrm{ML}}$ that we made earlier. The second one is the self consistent estimator dubbed "brave generalization student". It also turns out to be the estimator corresponding to the "integrate out" Bayesian approach with the Laplace prior.

Figure 4 shows that all estimators present the same general shape. The average excess error is minimized around 0 at the price of an overshoot for weights of intermediate size. This is the logical consequence of our priors as we introduce some *a priori* information forcing the parameter to be as close to zero as possible. The most probable estimator w_{MP} gives rise to extremely good performance for small weights, and the increase produced when the teacher is of intermediate size is noticeably reduced compared to the "brave generalization student". On the other hand, w_{G} results in the smallest decrease in error for small weights, but it also performs much better for other weights. There is even virtually no increase in error when $\tilde{w}$ grows larger. Iterating the approximation as mentioned above, we can get estimators that produce sharper decreases in error and still display this nice property of bias cancellation.

(Hansen and Rasmussen, 1994) note that integrating these functions on a logarithmic ordinate axis with a uniform measure favors the function with the biggest decrease in 0. Indeed this is not surprising as the prior is in this case rightfully used and the most conservative estimator[5] will perform better.

5 Conclusion

We have studied the behaviour of a regularization prior using a simple example: estimating the mean of a Gaussian variable. Predictably, no estimators using this prior perform better, on average, than the maximum likelihood. This is no surprise as it is the optimal unbiased estimator, and the only one that does not use a crooked prior. It should be noted, though, that our goal was by no means to suggest an improved estimator of a Gaussian random variable, but rather to illustrate the effect obtained by using a Laplace prior for regularization purposes.

These results tend to show that this prior provides a more straightforward mean of pruning than the use of a weight-decay and adaptive regularization, while the performances in terms of generalization are comparable for both methods. The use of this pruning prior together with neural networks on problems related to signal processing, as presented in (Goutte, 1996), has proved very satisfactory.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243–247.

[5]i.e. one that sets the most weights to 0.

Buntine, W. L. and Weigend, A. S. (1991). Bayesian back-propagation. *Complex Systems*, 5:603–643.

Goutte, C. (1996). On the use of a pruning prior for neural networks. In *Neural Networks for Signal Processing VI – Proceedings of the 1996 IEEE Workshop*, number VI in NNSP, Piscataway, New Jersey. IEEE.

Hansen, L. K. and Rasmussen, C. E. (1994). Pruning from adaptive regularization. *Neural Computation*, 6:1223–1232.

Jaynes, E. T. (1985). Bayesian methods: general background. In Justice, J., editor, *Maximum Entropy and Bayesian Methods in Applied Statistics*, pages 1–25. Cambridge University Press.

Krogh, A. and Hertz, J. A. (1992). A simple weight decay can improve generalization. In Moody, J. E., Hanson, S. J., and Lippman, R. P., editors, *Advances in Neural Information Processing Systems*, volume 4 of *NIPS*.

MacKay, D. (1992). Bayesian interpolation. *Neural Computation*, 4:415–447.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C*. Cambridge University Press, 2nd edition.

Williams, P. M. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143.

Wolpert, D. (1995). What Bayes has to say about the evidence procedure. In Heidbreder, G., editor, *1993 Maximum Entropy and Bayesian Methods Conference*. Kluwer.

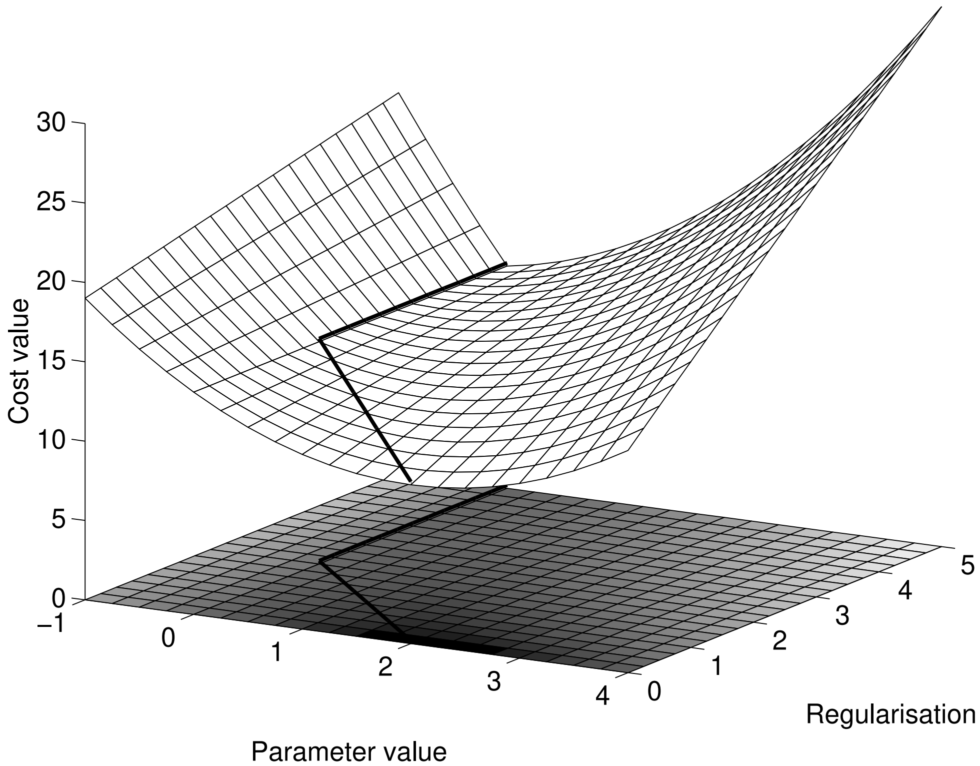

Figure 1: Optimal weight trajectory on the cost surface and its projection. The floor axes are the weight value (from -1 to 4) and the regularization parameter (from 0 to 5). For a given level of regularization, the optimal weight can be found on the bold line. The pruning effect appears when ξ reaches 2, which corresponds to the condition in (18). In this case, $w_{\text{ML}} = 2$.

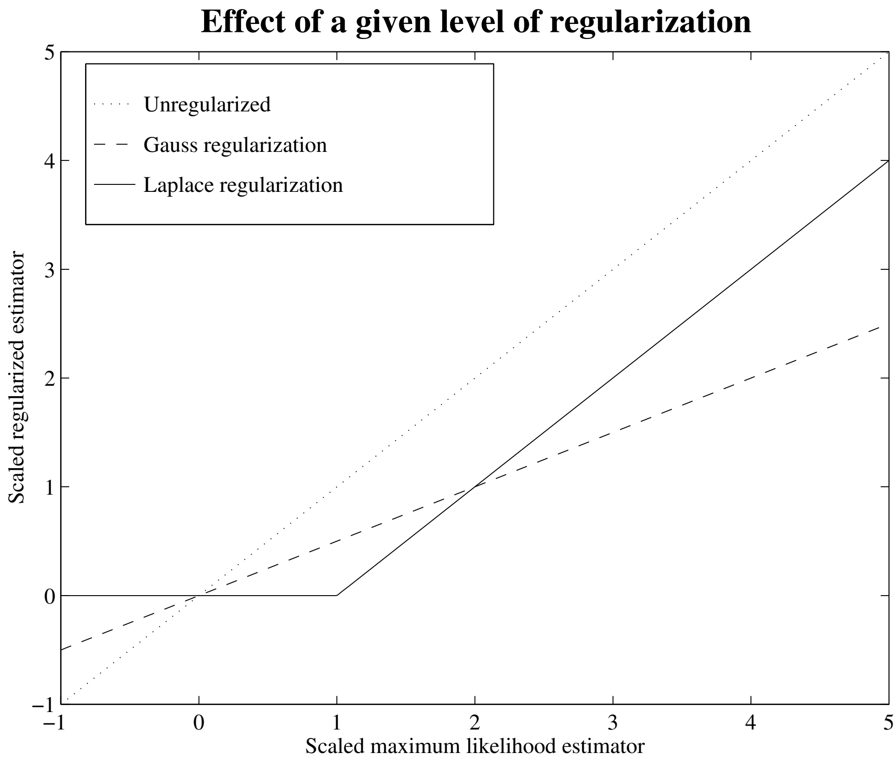

Figure 2: The Laplace prior prunes small weights and introduces a constant bias otherwise. The *weight decay* never prunes and produces a bias linear in the parameter's value. All quantities are scaled.

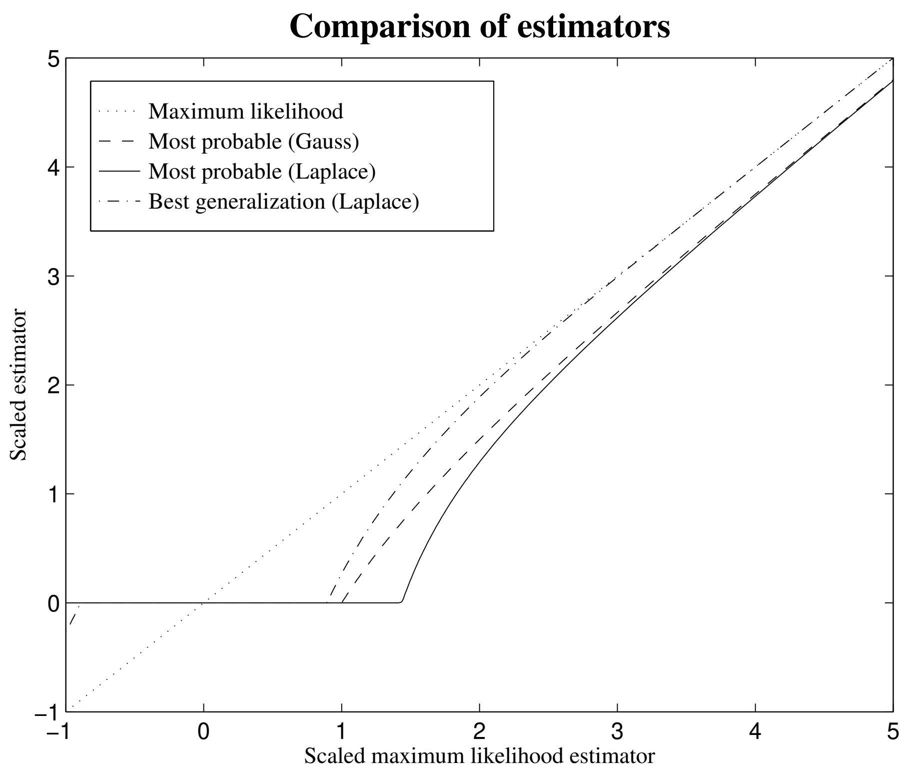

Figure 3: The solid and dash-dotted lines correspond to Laplace regularization, and the dashed line to a simple weight decay (Gaussian prior). All quantities are scaled.

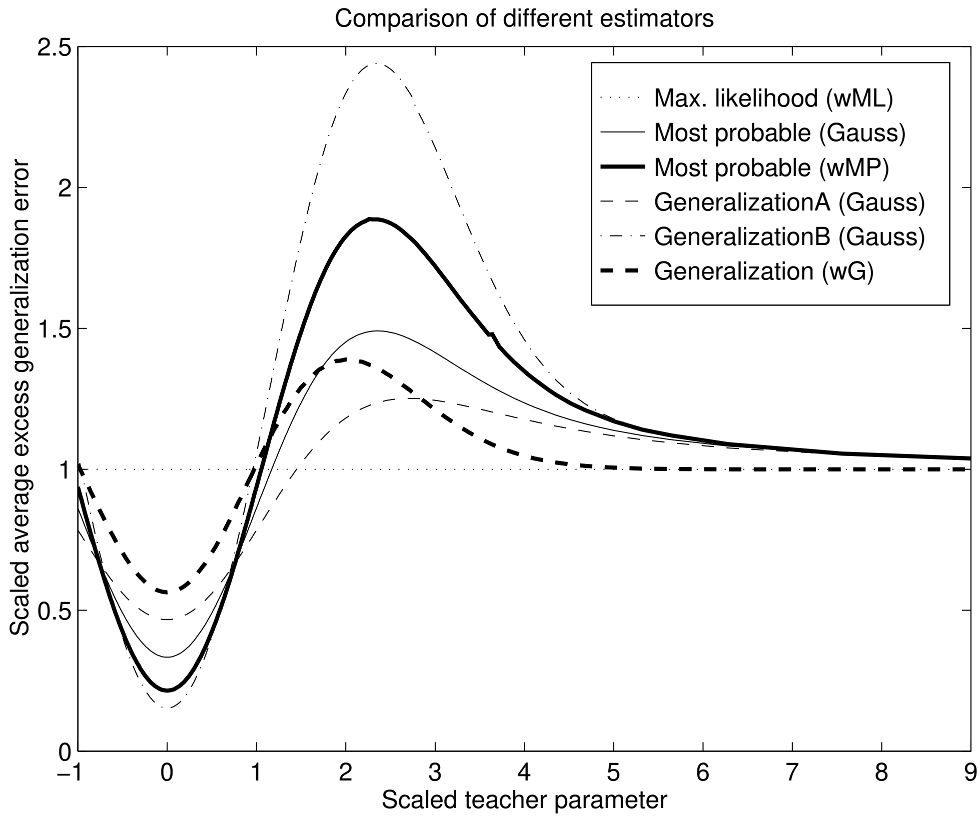

Figure 4: Average excess error for different estimators. Estimators marked Gauss correspond to the weight decay and can be found in (Hansen and Rasmussen, 1994). The bold lines correspond to the estimators calculated in this paper (with estimator name marked). The average excess error is scaled by the excess error of the maximum likelihood estimator σ^2/N.