Neural Networks 67 (2015) 121–130

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A biological mechanism for Bayesian feature selection: Weight decay and raising the LASSO

Patrick Connor [a,*], Paul Hollensen [a], Olav Krigolson [b], Thomas Trappenberg [a]

[a] *Faculty of Computer Science, Dalhousie University, Halifax, Nova Scotia, Canada*

[b] *Neuroeconomics Laboratory, University of Victoria, Victoria, British Columbia, Canada*

ARTICLE INFO

Article history:
Received 16 May 2014
Received in revised form 18 February 2015
Accepted 12 March 2015
Available online 25 March 2015

Keywords:
Feature selection
Weight decay
Bayesian prior
Online learning
LASSO
Regularization

ABSTRACT

Biological systems are capable of learning that certain stimuli are valuable while ignoring the many that are not, and thus perform feature selection. In machine learning, one effective feature selection approach is the least absolute shrinkage and selection operator (LASSO) form of regularization, which is equivalent to assuming a Laplacian prior distribution on the parameters. We review how such Bayesian priors can be implemented in gradient descent as a form of weight decay, which is a biologically plausible mechanism for Bayesian feature selection. In particular, we describe a new prior that offsets or "raises" the Laplacian prior distribution. We evaluate this alongside the Gaussian and Cauchy priors in gradient descent using a generic regression task where there are few relevant and many irrelevant features. We find that raising the Laplacian leads to less prediction error because it is a better model of the underlying distribution. We also consider two biologically relevant online learning tasks, one synthetic and one modeled after the perceptual expertise task of Krigolson et al. (2009). Here, raising the Laplacian prior avoids the fast erosion of relevant parameters over the period following training because it only allows small weights to decay. This better matches the limited loss of association seen between days in the human data of the perceptual expertise task. Raising the Laplacian prior thus results in a biologically plausible form of Bayesian feature selection that is effective in biologically relevant contexts.

1. Introduction

A supervised learning problem where there are few training examples relative to the number of input features is difficult. Often, however, there are only a few relevant features such that the feature space can be condensed somehow. The fewer model parameters needed, the less chance that the model will overfit or memorize the training data and generalize poorly to test data. So, the researcher will often employ feature reduction (Back & Trappenberg, 2001; Fodor, 2002; Guyon & Elisseeff, 2003; Saeys, Inza, & Larrañaga, 2007) or, more specifically, feature selection to find a useful subset of the features to employ. An embedded form of feature selection (Saeys et al., 2007) is regularization or weight decay, which is combined with regression to reduce the weight of uninformative features. The least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) is an effective such form of regularization.

It is useful to frame regularization and weight decay into the Bayesian framework of prior assumptions. The general feature selection assumption is that each relevant feature comes from a uniform probability distribution and that only a few features are relevant leaving the rest with zero influence, as shown in Fig. 1. It has been shown that applying certain prior distributions to regression is functionally equivalent to specific forms of weight decay or regularization (MacKay, 1992; Williams, 1995), which provides a mechanism for biologically plausible Bayesian feature selection. In our simulations of a regression task with few relevant features, we found that "raising" the Laplacian prior led to substantially lower prediction errors than the LASSO regularization (i.e., the Laplacian prior, Tibshirani, 1996) and ridge regression (i.e., the Gaussian prior). This is because a raised Laplacian prior better approximates the distribution from which the underlying generative parameters were drawn in our task, which followed the general feature selection assumption.

Investigating weight decay as a biologically plausible mechanism for Bayesian feature selection, we found that the raised Laplacian prior also has desirable properties in an online setting. Although it is possible that batch learning is employed in biological

[*] Corresponding author.
E-mail address: patrick.connor@dal.ca (P. Connor).

http://dx.doi.org/10.1016/j.neunet.2015.03.005
0893-6080/© 2015 Elsevier Ltd. All rights reserved.

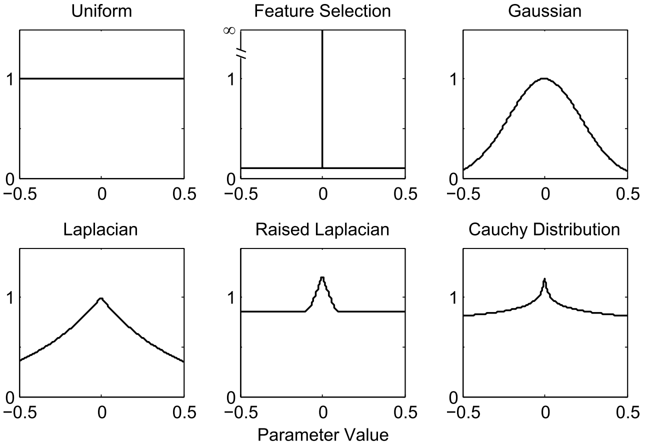

Fig. 1. Prior distributions. The uniform distribution is naturally assumed for regression. When feature selection is used, however, this implies that the expected prior is a low probability uniform prior (representing few relevant features) plus a high probability (delta function) prior at zero (representing irrelevant features). The Gaussian and Laplacian priors are often used to perform an embedded form of feature selection. However, these priors with settings that achieve representative results in later simulations do not match the feature selection 'prior' very well. The raised Laplacian and Cauchy priors model the assumed feature selection prior and thus provide lower prediction errors, as we will show in a specific regression task.

systems to some degree as a consequence of hippocampal replay (Ji & Wilson, 2006; Lansink, Goltstein, Lankelma, McNaughton, & Pennartz, 2009), it is presumed that much biological learning (e.g., during active awake periods) occurs online. How do priors compare in this setting? In a simulation of online learning, we show that raising the Laplacian prior avoids the natural erosion of the parameters for relevant features that occurs with the other priors or forms of regularization evaluated here. Also, in a specific association task (Krigolson, Pierce, Holroyd, & Tanaka, 2009), we show that raising the Laplacian prior reduces prediction errors during training and decays little during between-training rest periods, which matches the experimental data. These findings lend support to the raised Laplacian prior as a plausible form of biological feature selection.

There are a number of approaches related to using a raised Laplacian prior. The raising of the Laplacian prior is somewhat similar to the so-called "spike-and-slab" mixture priors (Ishwaran & Rao, 2005; Lempers, 1971; Mitchell & Beauchamp, 1988). These hierarchical priors define the probability of how often the true prior is derived from the spike-shaped prior (e.g., a narrow Gaussian) versus the slab (e.g., uniform prior). These can then be used to sample the posterior. Our proposed raised Laplacian flattens the hierarchy of the approach, summing the spike and slab elements and leading to the regularization or weight decay style terms commonly used in embedded feature selection. There are also a variety of LASSO variants which aim at reducing its prediction error bias (Fan & Li, 2001; Zhang, 2013; Zou, 2006). We show how several of these relate to some of the Bayesian feature selection priors we investigate (and vice versa).

2. Bayesian priors, regularization, and synaptic weight decay

We first discuss our proposed approach for data that are generated as

$$y = \psi^T x + \mathcal{N}(0, \sigma^2) \tag{1}$$

where x is an input vector, y is the outcome, ψ is a vector of function parameters, and 0 and σ^2 are the mean and variance of a Gaussian random variable. This linear world model with Gaussian noise is appropriate here for several reasons: (1) it is simple enough to highlight the features of our approach (2) many real world problems are actually at least locally linear, and (3) system noise is often Gaussian due to the central limit theorem. The major issue we consider here is that we have many features in the world, but only a few may be relevant to the quantity to be predicted. Furthermore, this task is even more challenging when given only a limited number of examples, or real-world experiences, from which the parameters must be estimated. This is the world in which biological systems navigate and must therefore possess the ability to overcome.

It is appropriate to model the relationship between predictive features (x) and an outcome (y) as a probability density function (PDF). Common ways of proceeding are to make an explicit parameterized assumption, as used below, or an implicit assumption expressed in methods such as neural networks. Here we know that the data conforms to a specific PDF model, whose parameters we can configure using the available data. The PDF that represents the Gaussian random variable with linear mean (i.e., Eq. (1)) is given as

$$p(y, x|\phi) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\phi^T x)^2}{2\sigma^2}} \tag{2}$$

where ϕ are the model parameters to be inferred from data generated by Eq. (1) (seeking to discover ψ). A standard way of doing this is by finding the most likely parameters given the data (the posterior probability) according to Bayes rule,

$$P(\phi|D) = \frac{P(D|\phi)P(\phi)}{P(D)} \tag{3}$$

where $D = \{x, y\}$ is the data (i.e., a collection of inputs and associated outcomes). The likelihood term $P(D|\phi)$ is the PDF in Eq. (2). Without prior knowledge, the prior distribution on the data, $P(D)$, becomes irrelevant since it is the same for all ϕ hypotheses and we are only interested in finding the most likely hypothesis. The factor $P(\phi)$ defines a priori how likely an individual hypothesis or parameter combination is. It is the choice of this prior that is the focus of the present work. From Eq. (3), the relative likelihood that a specific data set is generated according to a certain ϕ hypothesis is

$$\begin{aligned} L(\phi) &= p(y^{(1)}, \ldots, y^{(m)}, x^{(1)}, \ldots, x^{(m)}|\phi)p(\phi) \\ &= p(\phi)\prod_{i=1}^{m} p(y^{(i)}, x^{(i)}|\phi) \end{aligned} \tag{4}$$

where m is the number of training data points. To find the set of parameters that maximizes the likelihood of observing the data provided, that is, the maximum a posteriori (MAP) estimate,

$$\phi^{MAP} = \underset{\phi}{\operatorname{argmax}}(L(\phi)) \tag{5}$$

we take the argument's log,

$$\begin{aligned} \log\ L(\phi) &= \log\ p(\phi) + \log \prod_{i=1}^{m} p(y^{(i)}, x^{(i)}|\phi) \\ &= \log\ p(\phi) + \sum_{i=1}^{m} \log\ p(y^{(i)}, x^{(i)}|\phi) \\ &= \log\ p(\phi) + -m\ \log\left(\sqrt{2\pi}\sigma\right) \\ &\quad - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y^{(i)} - \phi^T x^{(i)})^2 \end{aligned} \tag{6}$$

and ascend its gradient,

$$\frac{\partial\ \log\ L(\phi)}{\partial\phi_j} = \frac{\partial\ \log\ p(\phi)}{\partial\phi_j} + \frac{1}{\sigma^2}\sum_{i=1}^{m}(y^{(i)} - \phi^T x^{(i)})x_j^{(i)}. \tag{7}$$

The gradient can be ascended by iteratively updating the model parameters,

$$\phi_j =: \phi_j + \alpha \frac{\partial \log L(\phi)}{\partial \phi_j} \tag{8}$$

where the learning rate is

$$\alpha = \frac{\sigma^2}{m(n+2)} \tag{9}$$

and n is the number of input features. Notice that α's numerator contains the variance of the Gaussian random variable, which will conveniently cancel the $\frac{1}{\sigma^2}$ in part of the expression. The prior itself, through an internal parameter λ, will offset this variance term. The form of α's denominator was chosen because it is optimal for the case when the input, x, has a mean of zero and variance of one, which we use in the first of our simulations. The entire formulation we present here is equivalent to least mean squares (LMS) regression when the prior is uniform (i.e., $p(\phi) = 1$, $\frac{\partial \log p(\phi)}{\partial \phi_j} = 0$).

2.1. The Gaussian prior, ridge regression, and weight-proportional decay

A simple non-uniform prior often used is the (unnormalized) Gaussian distribution, $p(\phi) = (e^{-\frac{\phi^T \phi}{2}})^{\lambda}$, where λ dictates the strength of the prior or, in this case, the Gaussian's variance. The gradient ascent update term provided by the prior (whether normalized or not) becomes

$$\frac{\partial \log p(\phi)}{\partial \phi_j} = -\lambda \phi_j. \tag{10}$$

Thus, the Gaussian prior reduces a parameter in proportion to its size. This prior is equivalent to "ridge regression" (Hoerl & Kennard, 1970), a special case of Tikhonov regularization (Tikhonov, 1963), which employs an L_2 norm penalty on the parameters. Krogh and Hertz (1992) showed that the optimal value of λ is the variance of the Gaussian random variable divided by the average squared value of the underlying linear model parameters which generate the data, or $\lambda = \frac{\sigma^2}{\frac{1}{m}\sum_i \psi_i^2}$ relative to Eq. (1). When the parameters of the linear function are seen as weights on a linear perceptron or model neuron, this prior can also be viewed as a synaptic weight decay, where large weights decay faster than small weights. This relationship between a Bayesian prior probability, a form of regularization, and synaptic weight decay can be seen in other cases as well, as described below.

2.2. The Laplacian prior, the LASSO regularization, and constant decay

An often more effective prior commonly used is based on the exponential function, $p(\phi) = (e^{-\sum_j |\phi_j|})^{\lambda}$. This prior is referred to as the double exponential or the Laplacian prior (Williams, 1995), which gives an update term of

$$\frac{\partial \log p(\phi)}{\partial \phi_j} = -\lambda \, \mathrm{sign}(\phi_j). \tag{11}$$

This is equivalent to the LASSO regularization (Tibshirani, 1996), which employs an L_1 norm penalty on the parameters (see Vidaurre, Bielza, & Larrañaga, 2013 for a survey of its use). The gradient is technically non-differentiable at parameter values of $\phi_j = 0$, but in our implementations the sign() function is used to set the gradient in such cases to zero. A number of solutions involving quadratic programming have been devised (Schmidt, 2005; Tibshirani, 2011) that find the exact solution. Yet, because of the convex nature of the problem, the approximate gradient will lead to a good approximation of the global maximum. In the following simulations, this direction is taken in an effort to provide a weight decay interpretation of the approach. The primary difference is that, very small parameter values will constantly jump back-and-forth across zero instead of shrinking to zero as in exact LASSO solutions. In the results section, we calculate the maximum error due to the approximate gradient for our task and show that its effect is negligible.

The LASSO has become a very popular feature selection technique, having many variants (e.g., Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005; Yamada, Jitkrittum, Sigal, Xing, & Sugiyama, 2014; Zou, 2006) and new applications being added regularly (e.g., Colombani et al., 2013; Loo et al., 2014; Toiviainen, Alluri, Brattico, Wallentin, & Vuust, 2014). The LASSO's implicit prior naturally tends to encourage large values or zero values more than ridge regression's implicit Gaussian prior, placing more of the probability mass at zero and in its tails (Tibshirani, 1996). This leads to smaller regression parameters for irrelevant features.

2.3. Raising the Laplacian prior leads to a decay zone

When considering the shapes of the Gaussian and Laplacian priors from Fig. 1, we see that neither of these match well with the expected distribution of parameter values when only a few of a great many features are relevant. Raising the Laplacian prior offers a novel and useful prior distribution

$$p(\phi) = \begin{cases} \prod_j \left(\frac{z}{2s} + \frac{\beta(1-z)}{2(1-e^{-\beta s})} e^{-\beta|\phi_j|} \right)^{\lambda}, \\ \quad \text{for } -s < \phi_j < s \\ 0, \quad \text{otherwise} \end{cases} \tag{12}$$

with a gradient of

$$\frac{\partial \log p(\phi)}{\partial \phi_j} = \begin{cases} -\lambda \, \mathrm{sign}(\phi_j) \frac{s\beta^2(1-z)}{s\beta(1-z) + (1-e^{-\beta s}) z e^{\beta|\phi_j|}}, \\ \quad \text{for } -s < \phi_j < s \\ 0, \quad \text{otherwise} \end{cases} \tag{13}$$

where z is the probability that a parameter is chosen randomly with a uniform distribution (bounded between $-s$ and s) and β is inversely proportional to the width of the Laplacian centered about zero. As in the above priors, λ establishes the effect or weight of the prior on the posterior probability. In short, this distribution represents a uniform distribution plus a relatively narrow Laplacian distribution, suggesting that some proportion of the parameter values are drawn around 0 and some are drawn with a uniform distribution. During updates with this prior, weights decay rapidly in a zone surrounding zero and almost not at all for larger values of ϕ. The decay zone is rather sharply bounded with a width proportional to β and $1 - z$.

When $z = 1$ this approach reverts back to a uniform distribution and when $z = 0$, it is equivalent to the LASSO (if $\beta = 1$). This bears resemblance to Elastic Nets (Zou & Hastie, 2005), which combine the LASSO and ridge regression in a similar weighted way. The benefit of combining the LASSO's Laplacian prior with a uniform distribution is that it guarantees a significant probability that parameters will be drawn from the extremes of the range. Adding a uniform distribution to the Laplacian prior frees it to have a narrow width about zero instead of forcing it to be wide enough to capture extreme parameters. Adding a Gaussian prior, as in Elastic Nets, does not offer the same freedom.

2.4. The Cauchy prior

Of all of the common prior distributions one might use, the most similar to the raised Laplacian might be the Cauchy prior (albeit with an uncommon parameterization). It is a special case of the

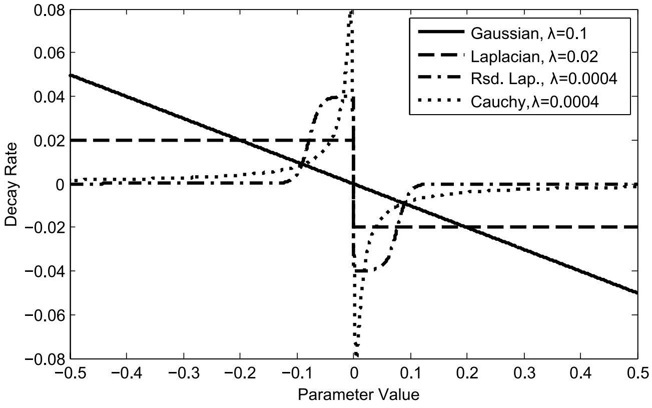

Fig. 2. The decay rate of parameters for each prior. For the Gaussian, weights decay in proportion to their magnitude. In contrast, the Cauchy prior's weights decay inversely proportional to their magnitude. For the Laplacian prior, weights are reduced by a constant amount. Raising the Laplacian has the effect of nearly shutting off decay for large weights, which fall outside of the decay zone. The width of the decay zone is affected by the choice of values for z and β.

student-t prior (as is the Gaussian prior), where the number of degrees of freedom is 1. The Cauchy prior is defined as

$$p(\phi_j) = \left(\frac{1}{\pi\gamma\left(1 + \left(\frac{\phi_j}{\gamma}\right)^2\right)} \right)^{\lambda} \tag{14}$$

where γ is proportional to the width of the high-probability region. This prior has a compact mathematical expression relative to the general purpose student-t prior, but its most important feature is that it has heavy tails (see Fig. 1). Given an appropriate choice of γ, the Cauchy prior's heavy tails also begin to approximate the prior assumption implicit in feature selection. However, such extremely heavy tails appear to be an uncommon parameterization of this prior. The batch update term for the Cauchy prior can be defined exactly and is

$$\frac{\partial \log p(\phi)}{\partial \phi_j} = -\lambda \frac{2\phi_j}{\gamma^2 + \phi_j^2}. \tag{15}$$

Although not as widely used as either ridge regression or the LASSO, the Cauchy prior has been suggested as a good default prior (Gelman, Jakulin, Pittau, & Su, 2008).

Fig. 2 shows a plot of the decay rate for values of ϕ between −0.5 and 0.5 for all of the priors examined here, using the same parameter settings as the representative results in the following section. The forms of weight decay, which are calculated from the priors' gradients $\frac{\partial \log p(\phi)}{\partial \phi_j}$ are vastly different. The Gaussian prior leads to a weight proportional weight decay, where large weights are decreased more with each epoch than small weights. The Cauchy prior leads to the decay of small weights faster than large weights. The Laplacian prior leads to constantly decaying weights, regardless of their size. Raising the Laplacian prior has the effect of stopping the decay of larger weights, which are outside of its decay zone.

3. Experiments and results

We expect that raising the Laplacian prior will reduce prediction errors in a feature selection setting where there are few relevant and many irrelevant features, because it better approximates the prior probability on the feature relevance distribution. To evaluate this, we employ a batch learning task with only a few relevant features and stress test the priors' ability to reduce prediction errors by separately increasing the number of irrelevant features (to as many as ten times the number of data instances) and the amount of Gaussian noise added to the training data.

As discussed earlier, it seems that awake, biological systems operate on a primarily online learning basis. Thus, we seek to determine if raising the Laplacian has a beneficial effect on prediction errors in an online setting. One issue of concern for conventional priors is that associations or weights are expected to degrade during quiescent periods due to the weight decay forms they impose. We first evaluate all of the priors in an online learning feature selection task that is similar to the batch learning experiment, and based on synthetic data. Then we simulate a specific association task (Krigolson et al., 2009) in an online learning fashion and compare it to recorded human data, which exhibits an association degradation following a quiescent period.

3.1. *Batch learning with few data points and many irrelevant inputs*

In this section, we evaluate what happens in a simple linear regression task as we vary the number of irrelevant inputs and amount of noise when there is little data. For each simulation, the linear model used to generate the data has two ψ parameters with real values and the rest are set to zero. Each data point in a simulation (x) is drawn from a Gaussian distribution centered about zero with a variance of one, which suits the learning rate in Eq. (9). The data point is passed through the generative linear model according to Eq. (1) to get the output which our models will attempt to match.

Setting only the values of two of the generative linear model parameters represents the case when there are few relevant features and many irrelevant features, for which feature selection is commonly employed. To get a representative sampling of the combinations of the two relevant features in our evaluation, we repeat each simulation for all possible positive–negative parameter combinations (e.g., 0.25, −0.1) and positive–positive parameter combinations (e.g., 0.25, 0.1) between 1 and −1 from a grid with a resolution of 1/11, which gives giving 72 combinations. The results of training and testing on these combinations provide the mean prediction errors displayed in the figures. The training set contains 103 data points, which is equal to the number of model parameters when there are 100 irrelevant parameters (see below), 2 relevant parameters, and a bias parameter. The test set contains 1000 data points to provide sufficient accuracy for our comparisons.

Training must be terminated before the method "memorizes" the specifics of the data rather than the true linear model to reduce overfitting. For this, a "validation" data set is presented to the model after every cycle (epoch) through the data. In this strategy, training will continue as long as the prediction error on the validation set decreases with each epoch. In our implementation, we actually run for 25 epochs between evaluations of the validation set to allow for somewhat noisy ascents. Once the prediction error begins to grow again, training stops and the test set is evaluated.

Fig. 3 provides prediction errors for the various priors. When the results for an approach enter the upper shaded area, they are performing more poorly than the uniform prior and thus provide no benefit there. No result will enter the lower shaded area because this boundary represents the Bayesian optimal result. The Bayesian optimal prediction error is found by selecting the two relevant features manually and training with LMS. Here, the uniform prior is accurate so that LMS can be used to find optimal parameter values. Representative results for each prior are shown with appropriate λ values (discussed below) for the range of interest (Gaussian: $\lambda = 0.1$, Laplacian: $\lambda = 0.02$, raised Laplacian: $\lambda = 4 \times 10^{-4}$, and Cauchy: $\lambda = 4 \times 10^{-4}$), where the λ values reported are per data

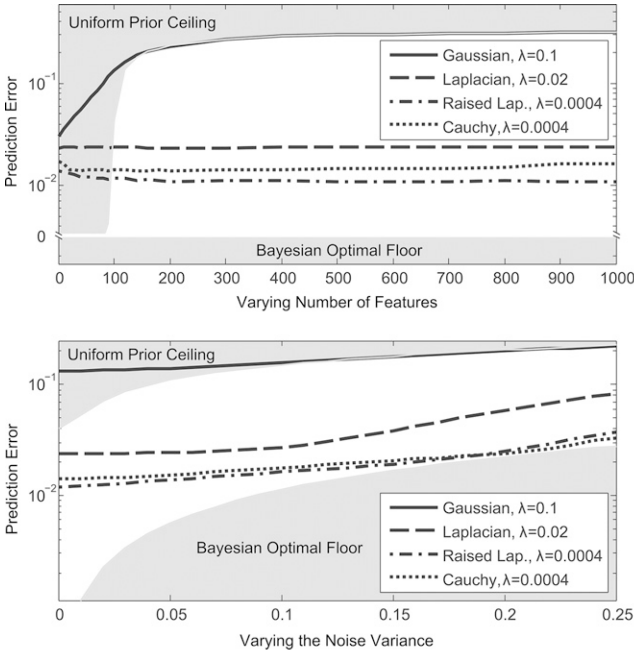

Fig. 3. Regression results with different priors. The above and below shaded regions bound useful results between the uniform prior's performance and the optimal Bayesian model's performance, respectively. The performance for each of the priors are shown for representative values of λ (Gaussian: $\lambda = 0.1$, Laplacian: $\lambda = 0.02$, raised Laplacian: $\lambda = 4\times 10^{-4}$, and Cauchy: $\lambda = 4\times 10^{-4}$). *Top Panel:* Prediction error as the number of features is varied. Representative λ values lower prediction errors for the Laplacian, raised Laplacian, and Cauchy priors, below that of the Gaussian prior. The raised Laplacian and Cauchy priors give slightly better performance than the Laplacian in general. *Bottom Panel:* The Gaussian prior only sees a marginal benefit over the uniform distribution at high noise levels. The raised Laplacian prior gives substantially better performance throughout the range than both the Gaussian and Laplacian priors, but comparable performance to the Cauchy prior.

point (i.e., the actual λ values in our equations are 103 times the reported λ values). This provides invariance to the number of training data points used, which will be especially important when we shift to online learning simulations in the next section. For all simulations, the learning rate was fixed as described earlier and further testing (not shown) revealed that changing the learning rate offered either none or only negligible benefit to any of the methods (i.e., no conclusions were affected). Additionally, we set the other parameters of the raised Laplacian to $\beta = 100$ and $z = 0.02$ and of the Cauchy prior to $\gamma = 5 \times 10^{-3}$. Exactly the same training and test data sets are used for all priors. Conveniently, this allows us to compare models using a matched-pairs type of statistical test to determine significance, namely the sign test. This testing was conducted between the representative curves (not shown in the figure for clarity's sake), which revealed that they are significantly different from one another ($p < 0.01$) except where curves cross one another. Model parameter values associated with the relevant input features will tend to be larger than other parameters. Their degree of prominence is displayed in their prediction errors — the lower the prediction error, the more prominent are the relevant feature parameters. In the Bayesian optimal case, only the relevant parameters have non-zero values. Thus, the proximity of a result to the Bayesian performance floor gives a sense of how clearly the relevant features dominate over the others.

In the top panel of Fig. 3 the noise is fixed at zero and the number of irrelevant inputs is varied between 0 and 1000 (giving between 2 and 1002 model inputs) for a training set of only 103 data points. LMS with a uniform prior (i.e., without the manual feature selection used to compute the optimal result) can solve a linear system of equations exactly, up until the number of inputs is greater than the number of data points (i.e., equations). Since we have 103 data points, 2 relevant features, and a bias parameter, this threshold is exactly set at 100 irrelevant inputs, where the prediction error due to LMS' uniform prior rises sharply in the figure. Here, the optimal λ value for the Gaussian prior is zero (i.e., a uniform distribution) because there is zero additive noise (Krogh & Hertz, 1992). Therefore we see that increasing λ only increases prediction errors. In contrast, the Laplacian and raised Laplacian priors completely eliminate the rapid climb seen for the uniform and Gaussian priors. Nevertheless, the λ values must be carefully chosen since larger λ values increase the base level of their prediction errors. The raised Laplacian prediction errors tend to be lower than the Laplacian errors. This is partly due to the fact that the λ values chosen for each approach were also selected to perform well in the case of varying noise. Finally, the Cauchy prior curves fall in about the same range as the raised Laplacian curves in this figure.

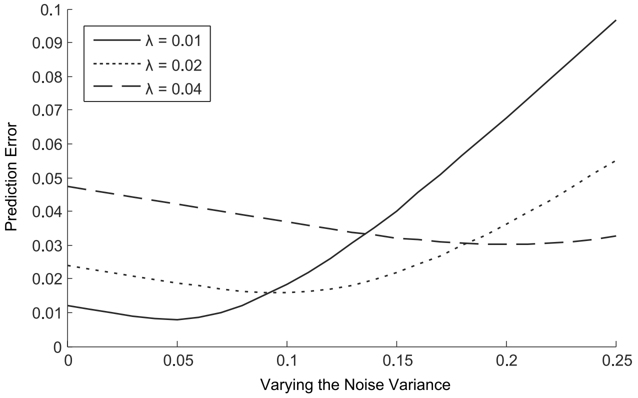

Fig. 4. Difference in prediction error between the Laplacian prior and the optimal Bayesian model as the value of λ is varied. From the graph, we see that the value of λ determines the amount of noise at which the closest approach is made to the optimal result. It demonstrates that for large amounts of noise, a larger λ is preferred, whereas for smaller amounts of noise, a smaller λ is preferred. This principle holds for the other priors as well, although it is not shown here.

In the bottom panel of Fig. 3 the number of irrelevant parameters is fixed at 100 (102 model inputs) and the variance of the additive Gaussian noise in the training data is varied between 0 and 0.25. Note that Gaussian noise is not added to the test sets to more clearly show the differences in performance. As noted earlier for the Gaussian prior, it has been shown that the optimal λ value depends on the true variance of the error (Krogh & Hertz, 1992). The same is true of the other priors. Fig. 4 shows the difference between the Bayesian optimal prediction error and the LASSO prediction error for various levels of additive Gaussian noise. The figure shows that the choice of λ changes the point at which the performance curve approaches the optimal curve, where larger λ values are preferred for high noise situations and smaller λ values are preferred for low noise. Our choice for the λ for each prior in Fig. 3 places its approach in the bottom panel within the selected noise range and gives good overall results. In the end, the representative Laplacian and raised Laplacian priors give smaller overall prediction errors in the bottom panel than the Gaussian prior. Furthermore, the raised Laplacian prior curve approaches the optimal curve more closely than the Laplacian curve (even at the various values of λ shown in Fig. 4). Again, the Cauchy prior is also very effective. In this task, it achieves comparable results to the raised Laplacian curve, approaching the optimal curve about as closely as the raised Laplacian does.

Is the gain of raising the Laplacian prior over the LASSO merely an artifact of the approximating the LASSO regularization process with our weight decay like implementation? In short, no. In the worst case, parameters that would be zero in exact LASSO implementations would be at most ±αλ. Given 100 irrelevant parameters (used in the bottom panel of Fig. 3), with a roughly equal probability of being positive or negative, the variance of such a binomial distribution is $100 * 0.5(1 - 0.5) = 25$. The prediction error due to LASSO approximation, within two standard deviations (about 95% of the time) would be at most $2\sqrt{25}\alpha\lambda = 2\sqrt{25}\frac{0.02}{102+2} = 1.9 \times 10^{-3}$. At the extreme of having 1000 irrelevant features (used in the top panel of Fig. 3), the error due to approximation would be $2\sqrt{250}\frac{0.02}{1002+2} = 6.3 \times 10^{-4}$. These are far less than the prediction error differences between the LASSO and the raised Laplacian. In fact, the raised Laplacian is also subject to prediction errors caused by the approximation process. Instead, the primary reason for the difference in performance is the LASSO's "shrinkage" of relevant parameter values or its *bias*. According to Eq. (8), when the total prediction error is reduced to λ, $\frac{\partial \log L(\phi)}{\partial \phi_j}$ will be zero and no further learning will occur. That is, the penalty term keeps relevant feature parameters from reaching their full predictive values. In contrast, the raised Laplacian does not decay relevant parameters when their values are outside of the decay zone. This is usually the case, allowing relevant parameters to reach full value. We will further consider this issue and its relevance to other LASSO variants in the discussion section.

3.2. Online learning and the eroding effect of certain priors during quiescent periods

The primary focus of this paper is to investigate how biological systems may perform Bayesian feature selection. In the preceding results, the priors were evaluated using batch learning, where a full set of data points is given and processed as a group. However, it seems likely that biological systems operate in an online learning mode to some degree, learning from their stream of experience in real-time instead of storing it all for subsequent processing. Furthermore, a biological agent may pass through a variety of environments and spend uneven blocks of time among them, whereas standard machine learning tasks for which regularization is normally employed tend to suppose a more uniform sampling. Thus, we turn our attention toward online learning tasks and how the forms of weight decay imposed by the priors behave therein. Let us consider scenarios where predictive cues and outcomes are only active over a certain duration but recur after a significant period, whether regularly (e.g., weekly or annually, etc.) or irregularly. Although the goal remains to learn to predict outcomes with as little experience/data as possible, it is important that during the quiescent periods, the predictive nature of the relevant cues is not lost due to the weight decay.

To think in terms of a concrete and biologically relevant example, wild strawberries ripen in many areas in the late spring, changing size and color indicating when they are ready for harvest. Eventually, the harvest period passes, and the fruits are either picked or rot so that these cues disappear until next year. With this scenario, we can test the priors in an online way by preparing a data set (2 relevant features, 100 irrelevant features, 0.1 noise variance) where we only allow each model to see the data once (i.e., one cycle only). The two relevant features represent berry size (larger for larger berries) and berry redness (larger for brighter redness) relative to average levels and the outcome is the quality or value of the fruit to the animal. The noisy features represent other features of the real-world unrelated to strawberry appearance. All parameters are updated after the presentation of each data point, which includes the effects of the weight decay.

For the first 103 data points, the inputs are chosen as per the previous simulations, representing the harvest period. After that, we clamp the relevant features' inputs to zero but choose other inputs in the same way. This growing number of additional data points represents a period of time between harvest periods where neither the relevant ripe strawberry features appear (i.e., zero-valued inputs) nor is a strawberry consumed (i.e., zero output plus noise).

For each time step throughout the harvest and post-harvest periods, we test the ability of the models to predict the values of a range of berries correctly (1000 test points, as per earlier simulations). Fig. 5 shows these results, comparing the effectiveness of the four forms of weight decay in this task. All priors lead to a reduction in prediction error during the first 100 or so time steps or data points. However, after this period, the Gaussian, and Laplacian prediction errors increase because the relevant parameter values are eroded by weight decay with the presentation of each additional data point. These eventually rise above the uniform prior results, even with different values of λ (not shown) than used in the representative batch learning results shown earlier. The slightly differently parameterized raised Laplacian prior shown in the figure ($\lambda = 2 \times 10^{-4}$, $\beta = 300$, $z = 0.02$), however, proves to be quite stable, maintaining a relatively low prediction error throughout (slightly lower than with the parameterization of the representative results shown earlier). Most of the relevant parameter weights were increased sufficiently in the early period such that they were prevented from decay during the period when relevant features were absent (clamped at zero) and therefore not reinforced. With enough data, the uniform distribution's prediction error drops, but not below that of the raised Laplacian (sign test, $p < 0.01$) because of the additive Gaussian noise. The thick Cauchy prior learning curve ($\gamma = 5 \times 10^{-3}$, $\lambda = 4 \times 10^{-4}$) decays much like the Gaussian and Laplacian curves in the post-harvest period. It would seem that one way to reduce the decay in the Cauchy prior would be to make the weight decay profile sharper, which can be done by reducing γ. The two thin dotted lines of Fig. 5 show the performance for two such Cauchy priors ($\gamma = 1 \times 10^{-3}$, $\lambda = 8 \times 10^{-5}$ and $\gamma = 2 \times 10^{-4}$, $\lambda = 1.6 \times 10^{-5}$) with adjusted values of λ to maintain the peak weight decay (0.08, as shown in the weight decay curve of Fig. 2). Their performance in this online task suggests that the slope of the increasing prediction error during the post-harvest period does indeed decline as the weight decay profile gets sharper, but the error is substantially higher in the harvest period (when accurate predictions are needed most). Also, the effective decay range for the prior is reduced and cannot overcome the effects of noise as well as before. Thus, the raised Laplacian prior appears to have some advantage over the other priors in this online task.

3.3. Decay in a multiple-session association task

To examine the online effects of decay further and to relate weight decay to experimental data, we simulate the perceptual expertise task of Krigolson et al. (2009), where the goal is to categorize unique visual shapes (blobs) into either class A or B based on their appearance. Making the association between blob features and the correct category is a challenging task, such that the subject pool in the original experiment became divided between whether one could or could not learn the task. In a replication of the experiment (unpublished data) subjects were trained for a period each day (1000 categorization trials) for five days, allowing time for decay to occur between days. During training, electroencephalography was used to acquire the reward positivity (Holroyd & Coles, 2002; Holroyd, Pakzad-Vaezi, & Krigolson, 2008; Krigolson, Hassall, & Handy, 2014). An examination of the first block (first 100 trials) of day one with regard to the event related potential (ERP) feedback averaged waveforms for correct and incorrect trials revealed a feedback error-related negativity (fERN: Miltner, Braun, & Coles, 1997). In subsequent blocks (and days) however, there were not enough incorrect trials to warrant an examination of the fERN without fear of frequency contamination (Holroyd & Krigolson, 2007) and, as a result, subsequent ERP analyses focused on the

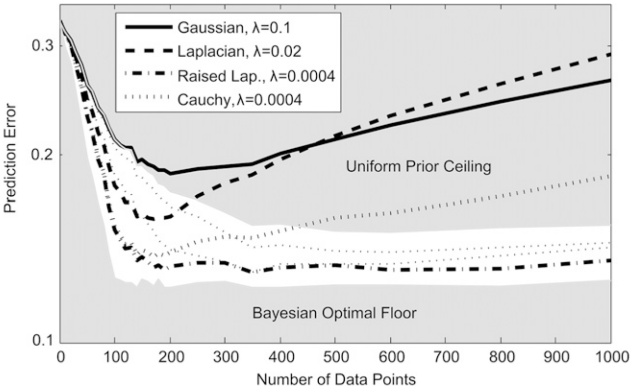

Fig. 5. Online learning prediction error as the number of data points is varied (100 irrelevant inputs and 0.1 noise variance). After the first 100 data points, the relevant feature inputs are clamped at zero. The weight decay from the Gaussian, Laplacian, and Cauchy priors (with the same parameterization as the earlier representative results) leads to an erasure of the relevant parameter values with increasing numbers of data points. The raised Laplacian ($\lambda = 2 \times 10^{-4}$, $\beta = 300$, $z = 0.02$), however, gives rise to a stable form of weight decay that erodes very little with time. Prediction errors for the uniform prior do not descend as quickly, but they are eventually lowered with enough data. The two thin dotted curves result from Cauchy priors that have a sharper weight decay profile ($\gamma = 1 \times 10^{-3}$, $\lambda = 8 \times 10^{-5}$ and $\gamma = 2 \times 10^{-4}$, $\lambda = 1.6 \times 10^{-5}$), but have the same maximum weight decay. The sharper profile is an attempt to reduce the rate of erasure of the relevant parameters with time. It does this but also increases prediction errors substantially during the harvest period.

reward positivity — the mean magnitude of the positive feedback average waveform for each subject in the time range of the fERN recorded above medial frontal cortex. In the top-left panel of Fig. 6, we see that the reward positivity descends over training, especially in the first two days, and is mostly flat on subsequent days. A comparison of the amplitude of the reward positivity between the last block of day one and the first block of day two revealed a small increase in amplitude — a restart cost (+1.4 uV; Cohen's $d = 0.28$, small effect: Cohen, 1988). A comparison of the last block of day one and the second block of day two revealed that the restart cost difference was diminished (+0.8 uV; Cohen's $d = 0.07$, no effect: Cohen, 1988). Thus, within the first block or two of training on the second day, the restart cost that had accrued during the intersession interval was overcome.

One possible explanation for the reward positivity increase between days is synaptic weight decay during the intersession period. Accordingly, Fig. 6 also contains learning curves showing the effect of each prior in a simulation of the experiment (see below for details). In our previous simulations, we saw that decay is good for generalization. Here, we have a real-world task (discrimination between two similar object types) where certain forms of decay, however, can have serious effects. Ultimately, it is desirable to keep the generalization benefit of decay during training without substantial intersession interval decay. The resulting simulation curves (one for each day of training and prior) show that the raised Laplacian ($z = 0.02$, $\beta = 100$, $\lambda = 4 \times 10^{-4}$) and Cauchy ($\gamma = 5 \times 10^{-3}$, $\lambda = 4 \times 10^{-4}$) priors provide a small intersession interval decay, matching the experimental data well, and have slightly steeper learning curves (i.e., better generalization) than the uniform prior. In contrast, the Gaussian ($\lambda = 0.1$) and Laplacian priors ($\lambda = 0.02$) appear to induce a learning reset during the intersession interval, despite a slight improvement in generalization over the uniform prior. The uniform prior naturally avoids decay during the intersession interval, but because there is some increase in reward positivity between days, it is not as good a model of the experimental data. In summary, for a decay explanation, the raised Laplacian and Cauchy priors fair best.

In the simulation fashioned after the task, there are 1000 training trials per simulated day separated into 100 trials per block and 5000 rest trials between the daily sessions. In every trial, regardless of type, the decay is active. Each blob class is represented uniquely by 10 of the 100 total inputs, where the relevance of each input is randomly chosen (uniformly between 0 and 1). In each trial, a blob's relative feature relevance remains constant, although their absolute value is scaled by a random value between 0.8 and 1.2 to allow for more or less generally salient blobs. The 80 irrelevant features, which are active 50% of the time with a random value (uniformly chosen between 0 and 1), represent that with the presentation of a given blob there will be some active features that are not unique to its class, which each reflect either an overlap with the other blob class or a common feature of the environment. The learning rate (set to 1×10^{-3}) was chosen to provide a learning curve on day one which descends at least half of the way toward the asymptote. The curves in Fig. 6 show the mean prediction error (not task accuracy) of 20 randomly initialized learners on each training trial, in an attempt to relate to the reward positivity decay curves above.

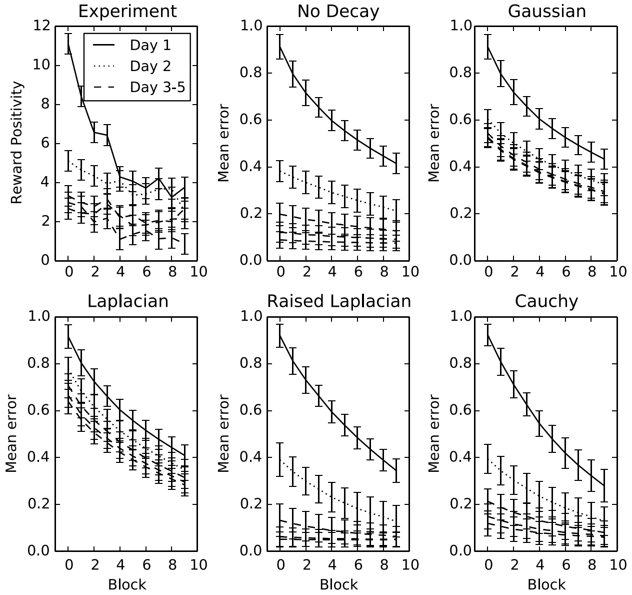

Fig. 6. Learning curves in the 2-blob association task (see text for details) using the same prior parameterizations as in the representative results of Fig. 3. The top-left panel shows the reward positivity data collected from subjects and the remaining panels provide simulation results of the task for various priors. The remaining panels show the mean and standard deviation of the error in a simulation of the task for each of the priors. The raised Laplacian and Cauchy appear to more closely resemble the real data, not decaying much between training sessions, and have slightly less error overall than does the uniform prior.

There are two primary variables that affect these decay curves: the number of rest trials representing the intersession interval and the number of unique features representing a specific blob class. Fig. 7 shows the mean prediction error after the first training block on the fifth day of training as these parameters are varied. As the number of rest trials increases, the Gaussian and Laplacian results rise (i.e., reset between days) sooner than the Cauchy, as shown in the top panel. The raised Laplacian always keeps a low prediction error. However, in the bottom panel, the Cauchy and raised Laplacian priors' results both begin to rise as the number of features representing a stimulus increases to the point that none of the associated weights escapes the priors' decay zones, and thus these priors erode the learned parameter values thoroughly during the intersession interval. For the same reason, similar curves for the raised Laplacian and Cauchy priors in the bottom panel would be generated if the salience of the prediction target (i.e., the value of the reward for predicting correctly) were reduced from 1 (used in our simulations) toward 0.

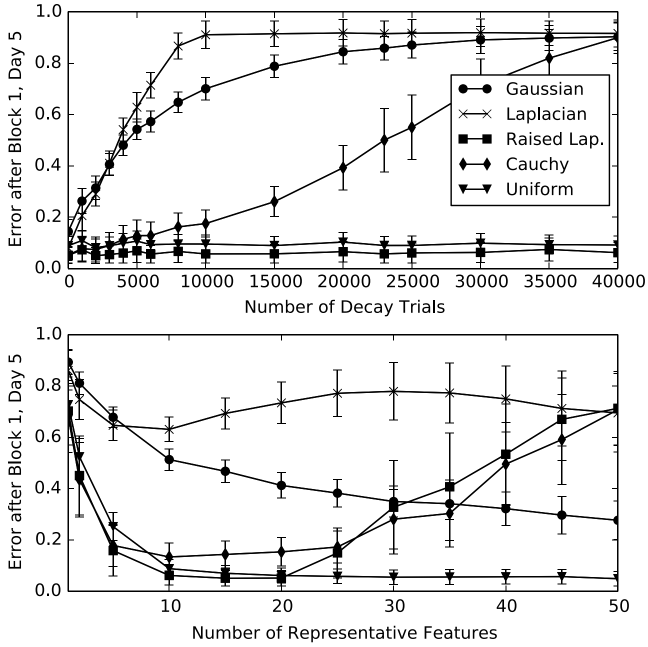

Fig. 7. The mean and standard deviation of the error after the first block on day 5 of training when varying the number of trials representing the intersession interval (top panel) and the number of features uniquely representing a blob class (bottom panel). The Cauchy and raised Laplacian priors tend to decay relatively little during the intersession interval unless the association is spread out among many representative features.

4. Discussion

It is a counter-intuitive notion that the decay of an association should assist learning. The principle of using an appropriate Bayesian prior and the results of the three foregoing simulations, however, unanimously affirm this notion. This is a reality because the decay reduces the influence of irrelevant features, allowing the association to be focused on the relevant features (i.e., decay leads to feature selection). Decay can also have a negative effect, however, as seen in the online learning simulations. Here, we have investigated how the form of decay affects prediction errors (i.e., generalization) and the strength of relevant feature associations over time, presumably both of which are relevant issues in biological processing.

It would seem that the "best" decay approach is one that reduces prediction error most while maintaining low decay of relevant parameters over extended periods of time. In our simulations, the weight-proportional (Gaussian) decay is least effective in lowering prediction error and badly decays relevant feature influence during non-training periods in online learning. Similarly, the constant (Laplacian or LASSO) decay is effective for generalization in batch learning but similarly leads to substantial decay in the online setting. The decay of primarily small valued weights (raised Laplacian and Cauchy) is effective in terms of achieving the lowest prediction error whether in the batch or online settings, and does not suffer nearly as much from the effects of decay following training periods. The raised Laplacian prior is somewhat more practical than the Cauchy prior in that its associations have a limited decay with time, since only weights within the decay zone will be lost. In contrast, the Cauchy prior leads to a small but significant decay for weights outside of the raised Laplacian's decay zone and thus will accrue substantial loss of relevant feature strength over intermediate and extended periods of time, as evidenced in the top panel of Fig. 7.

It has long been known that the LASSO produces a bias on prediction errors, and a number of LASSO variants have been designed to resolve it. For several of these and the priors discussed here, Fig. 8 shows their associated "thresholding functions", which illustrate the parameter bias as a y-axis deflection from the line of $y = x$. The adaptive LASSO (Zou, 2006) imposes a penalty that is very similar to the Cauchy prior presented here, having very little bias (or decay) for large parameter values, while having a large bias for small parameters. The smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001), also diagrammed in Fig. 8 resembles the raised Laplacian, bearing a configurable decay zone, but accomplishes this with a piece-wise continuous function. The capped or truncated LASSO (Zhang, 2013), which we call upon below as an approximation to the raised Laplacian, has the same bias as the LASSO for parameter values below some threshold and zero otherwise. All of these approaches suggest similar ways of mostly eliminating the bias on large parameter values. As reasonable approximations of the raised Laplacian and Cauchy priors, it would appear that such attempts at reducing the LASSO bias could be justified as accomplishing Bayesian feature selection following the general feature selection assumption.

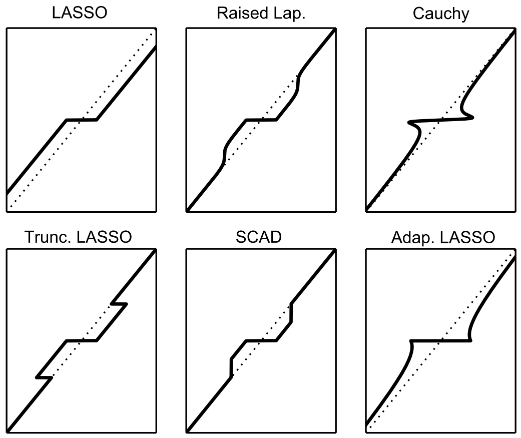

Fig. 8. Thresholding functions of the three priors and several LASSO variants that reduce or eliminate the LASSO's prediction error bias for large parameter values. The adaptive LASSO and Cauchy prior are very similar in their shape and the truncated LASSO and SCAD are both very similar to the raised Laplacian.

The raised Laplacian is non-convex, like several other LASSO variants, since there will be many local minima for the wide range of potential initial parameter values. To mitigate this, it is useful to initialize the parameter values to small or zero values where, within the decay zone and as in the LASSO generally, there is a single minimum. This is a reasonable constraint to expect of biological systems which must grow from scratch the hardware that implements their associative weights. It was also effective in our simulations, since the minima found gave lower prediction errors on average than the LASSO, which is convex. There are a number of algorithms that have been developed to solve the LASSO and its variants, which are devoted to improving computational efficiency (e.g., Efron, Hastie, Johnstone, & Tibshirani, 2004; Friedman, Hastie, Höfling, & Tibshirani, 2007; Friedman, Hastie, & Tibshirani, 2008). Although it is conceivable that the raised Laplacian and other priors may be similarly benefited for practical machine learning purposes, we have rather focused on the arguably less efficient weight decay approach because it is plausible and insightful in the primarily online, biological learning setting. This approach was sufficiently efficient to handle relatively large numbers of irrelevant features (1000). Additional work would be needed to determine the raised Laplacian's limits and to develop more computationally efficient implementations.

One side-effect of raising the Laplacian prior is that there are three additional parameters to be set. The parameter z represents the proportion of the probability that says features are drawn from a uniform distribution, which has a direct effect on the width of the decay zone (smaller z values give wider decay zones). The parameter β defines how wide the Laplacian prior should be, which has the direct effect of changing the slope of the decay zone boundary (smaller β values give gentler slopes and wider Laplacians), which affects the decay zone's width as well. Finally, the value of s or the width of the uniform prior must be defined to capture the maximum and minimum parameter values. The decay due to the raised Laplacian is a more expensive computation than the LASSO decay (compare Eqs. (11) and (13)). From Fig. 2, it would appear that the decaying effects of the raised Laplacian prior might be approximated by subtracting a constant value, as in the LASSO, for weights whose magnitudes fall below some threshold (i.e., within the decay zone), a less expensive computation. This is functionally the same as the truncated LASSO noted earlier. SCAD might also serve as an approximation, except that the width of its decay zone is intrinsically tied to its λ value and so will align with the raised Laplacian only under specific constraints. Fig. 9 compares the earlier performance of our raised Laplacian prior with a truncated LASSO approximation, where parameters less than ± 0.08 are reduced according to the LASSO decay in Eq. (11), with $\lambda = 0.04$. These results demonstrate that such an approximation performs comparably in our task.

When used in novel tasks, it will be necessary to find appropriate parameters for the priors. For the raised Laplacian, it makes sense to first find a truncation threshold and λ to achieve good results with the approximate model and then, if the proper raised Laplacian is needed, to adjust λ, β, z, and s to match the approximation. A good starting value for the truncation threshold is the average non-zero parameter value after training with the LASSO (estimating its λ according to conventional means). The truncation threshold and λ may be further refined by using a grid search within a small range around these values, evaluating each parameter pair using cross-validation in the novel task. Once a good approximation model has been found, plotting the derivative of the raised Laplacian (Eq. (13), as shown in Fig. 2) for various combinations of β and z can be used to find a close match to the approximation, from which a second cross-validation-based grid search could be used to further optimize as necessary. The s parameter is simply set to the width of the range of possible parameter values. For the Cauchy prior, the γ and λ parameters need tuning. The γ parameter dictates the sharpness of the Cauchy's gradient, as shown in Fig. 2, which indicates how aggressively it reduces larger values. For a starting value, we recommend choosing one that gives a gradient that is just short of enveloping the expected feature parameter values (again, the average non-zero parameter value after training with the LASSO). In our experiments, the Cauchy's effective λ value was either the same or very similar to λ for the raised Laplacian, which gave a peak decay in the derivative in Fig. 2 that was twice the peak value of the raised Laplacian, and four times the constant decay value of the Laplacian itself. Such a relative value would be a good start. The starting λ and γ parameters can be further optimized using a grid search and cross-validation as necessary. The parameter-finding processes for both the raised Laplacian and Cauchy priors may be used in both batch and online learning situations.

Aside from weight decay, there are two primary explanations of the intersession interval rebound of reward positivity in the experimental data to consider: warm-up decrement and retroactive inhibition. In warm-up decrement or loss of set (Ammons, 1947a, 1947b; Irion, 1948) the subject becomes efficient at performing the task over the course of training, adjusting his or her posture and attention to accomplish the task correctly and swiftly. During the intersession period, this skill set may be somewhat lost. When the subject is retrained after an intersession period, the task efficiency or skill takes some time to reestablish (i.e., the warm-up period). With a loss of set, one expects an initially steeper learning curve in a subsequent training session (Irion, 1948), because the set is recaptured rather quickly. Retroactive inhibition or interference and related theories of "forgetting" (McGeoch, 1942; Muller & Pilzecker, 1900; Robinson, 1927) would explain the data differently. It is possible for activities in the intersession period to decrease apparent retention, depending on the degree of similarity between these and the learning task. It appears that maximally similar tasks will cause the most interference. It has been hypothesized that intersession activities may either hinder memory consolidation of the task or reflect competition during recall (i.e., an impairment in retrieval). Loss of set and retroactive inhibition attribute different causes to the loss in retention than a mere decay. More finely resolved blocks (i.e., fewer trials per block) might have provided a better distinction between a decay in association and a loss of set by determining whether early learning on the second day had a sharper slope than uninterrupted learning at a comparable level of proficiency (say, at around block 8 of the previous day). At the 100 trials per block resolution, the fact that the reward positivity of day 2, block 2 is not below the day 1, block 10 value favors the decay explanation, since the first block (100 trials) of day 2 would seem sufficient to restore a loss of set in this task and allow the second block's reward positivity to descend below that of day 1, block 10. It is unclear how well retroactive inhibition might account for the decrement in the reward positivity on day 2 in this task. While, the blobs are very different from objects that subjects are likely to encounter during the intersession interval, there may be some overlap since object classification is part of daily routine. A version of the experiment that attempts to control for retroactive inhibition, by adding a strongly similar task during the intersession interval for one group, may be more effective at teasing this apart. In summary, a decay explanation of the experimental data would see steady (rather than steep) decreases in reward positivity on day two, and would see an intersession interval decay even when intersession periods involve minimally interfering activities (e.g., sleep).

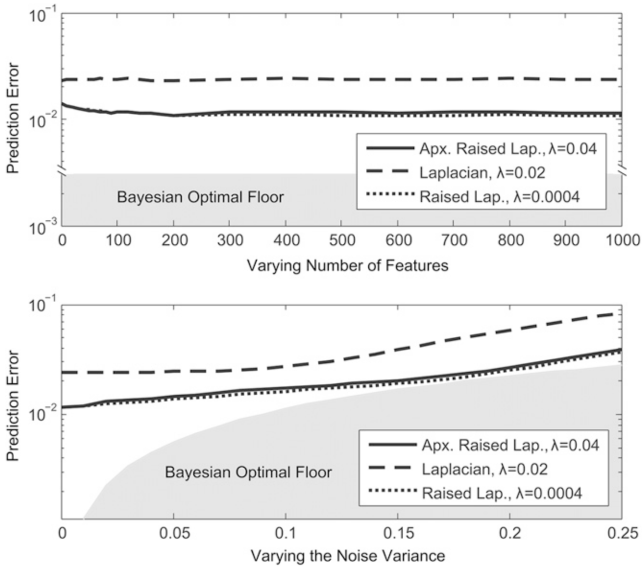

Fig. 9. Performance comparison between the raised Laplacian and a simple approximation referred to as the truncated LASSO. The approximation reduces parameters between −0.08 and 0.08 according to the LASSO decay in Eq. (11), where $\lambda = 0.04$. The approximation achieves very similar results to the raised Laplacian, maintaining significantly lower prediction errors than the LASSO.

5. Conclusion

In pursuit of a biologically plausible form of Bayesian feature selection, we have related various Bayesian priors in terms of their regularization or weight decay forms and have evaluated their influence in both batch and online learning tasks. Here, employing a raised Laplacian prior lowers prediction errors below those of the Gaussian and Laplacian priors in a batch learning regression task, since the raised Laplacian is a better model of the underlying parameter distribution. The Cauchy prior, with an unusual parameterization, can also achieve similar performance in this regression task. However, it along with the Gaussian and Laplacian priors, lead to the significant erosion of relevant parameters in online learning, a primary mode of learning in which biological systems are expected to engage. The raised Laplacian's "decay zone" instead limits the erosion of its relevant parameters, achieving persistent associations, unless the association is highly distributed among many features (or there is low target salience). We conclude that the exclusive decay of small weights, as imposed by the raised Laplacian prior, provides a biologically plausible implementation of Bayesian feature selection that is not only effective in batch learning, but sustains associations through quiescent periods in online learning and is therefore well suited to biological systems.

Acknowledgments

This work was supported by CIHR grant MOP-97741.

References

Ammons, R. B. (1947a). Acquisition of motor skill: I. quantitative analysis and theoretical formulation. *Psychological Review*, *54*, 263–281.

Ammons, R. B. (1947b). Acquisition of motor skill: II. rotary pursuit performance with continuous practice before and after a single rest. *Journal of Experimental Psychology*, *37*, 393–411.

Back, A. D., & Trappenberg, T. P. (2001). Selecting inputs for modeling using normalized higher order statistics and independent component analysis. *IEEE Transactions on Neural Networks*, *12*, 612–617.

Colombani, C., Legarra, A., Fritz, S., Guillaume, F., Croiseau, P., Ducrocq, V., et al. (2013). Application of Bayesian least absolute shrinkage and selection operator (LASSO) and BayesCπ methods for genomic selection in french holstein and montbéliarde breeds. *Journal of Dairy Science*, *96*, 575–591.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–499.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*, 1348–1360.

Fodor, I.K. (2002). A survey of dimension reduction techniques.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*, 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*, 432–441.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 1360–1383.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157–1182.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*, 679.

Holroyd, C. B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology*, *44*, 913–917.

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, *45*, 688–697.

Irion, A. L. (1948). The relation of 'set' to retention. *Psychological review*, *55*, 336–341.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, 730–773.

Ji, D., & Wilson, M. A. (2006). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, *10*, 100–107.

Krigolson, O. E., Hassall, C. D., & Handy, T. C. (2014). How we learn to make decisions: rapid propagation of reinforcement learning prediction errors in humans. *Journal of Cognitive Neuroscience*, *26*, 635–644.

Krigolson, O. E., Pierce, L. J., Holroyd, C. B., & Tanaka, J. W. (2009). Learning to become an expert: reinforcement learning and the acquisition of perceptual expertise. *Journal of Cognitive Neuroscience*, *21*, 1833–1840.

Krogh, A., & Hertz, J.A. (1992). A simple weight decay can improve generalization. In *Advances in neural information processing systems*, vol. 4 (pp. 950–957).

Lansink, C. S., Goltstein, P. M., Lankelma, J. V., McNaughton, B. L., & Pennartz, C. M. (2009). Hippocampus leads ventral striatum in replay of place-reward information. *PLoS Biology*, *7*, e1000173.

Lempers, F. B. (1971). *Posterior probabilities of alternative linear models: some theoretical considerations and empirical experiments*. (Ph.D. thesis), Universitaire Pers Rotterdam.

Loo, H. M., Cai, T., Gruber, M. J., Li, J., Jonge, P., Petukhova, M., et al. (2014). Major depressive disorder subtypes to predict long-term course. *Depression and Anxiety*, *31*, 765–777.

MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, *4*, 448–472.

McGeoch, J. A. (1942). *The psychology of human learning*. New York: Longmans, Green.

Miltner, W. H., Braun, C. H., & Coles, M. G. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a generic neural system for error detection. *Journal of Cognitive Neuroscience*, *9*, 788–798.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*, 1023–1032.

Muller, G., & Pilzecker, A. (1900). Experimentelle beitrage zur lehre vom gedachtnis. *Zeitschrift fur Psychologie und Physiologie der Sinnesorgane*, 1–300.

Robinson, E. S. (1927). The 'similarity' factor in retroaction. *The American Journal of Psychology*, 297–312.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*, 2507–2517.

Schmidt, M. (2005). Least squares optimization with L1-norm regularization.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 273–282.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 91–108.

Tikhonov, A. (1963). Solution of incorrectly formulated problems and the regularization method. In *Doklady Akademii Nauk SSSR*, vol. 151 (pp. 501–504).

Toiviainen, P., Alluri, V., Brattico, E., Wallentin, M., & Vuust, P. (2014). Capturing the musical brain with Lasso: dynamic decoding of musical features from fMRI data. *Neuroimage*, *88*, 170–180.

Vidaurre, D., Bielza, C., & Larrañaga, P. (2013). A survey of L1 regression. *International Statistical Review*, *81*, 361–387.

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

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., & Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computation*, *26*, 185–207.

Zhang, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli*, *19*, 2277–2293.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*, 1418–1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320.