LETTER Communicated by Ilya M. Nemenman

Bayesian Feature Selection with Strongly Regularizing Priors Maps to the Ising Model

Charles K. Fisher
charleskennethfisher@gmail.com
Pankaj Mehta
pankajm@bu.edu
Department of Physics, Boston University, Boston, MA 02215, U.S.A.

Identifying small subsets of features that are relevant for prediction and classification tasks is a central problem in machine learning and statistics. The feature selection task is especially important, and computationally difficult, for modern data sets where the number of features can be comparable to or even exceed the number of samples. Here, we show that feature selection with Bayesian inference takes a universal form and reduces to calculating the magnetizations of an Ising model under some mild conditions. Our results exploit the observation that the evidence takes a universal form for strongly regularizing priors—priors that have a large effect on the posterior probability even in the infinite data limit. We derive explicit expressions for feature selection for generalized linear models, a large class of statistical techniques that includes linear and logistic regression. We illustrate the power of our approach by analyzing feature selection in a logistic regression-based classifier trained to distinguish between the letters B and D in the notMNIST data set.

1 Introduction

Modern technological advances have fueled the growth of large data sets where thousands, or even millions, of features can be measured simultaneously. Dealing with such a large number of features is often impractical, however, especially when the sample size is limited. Feature selection is one promising approach for dealing with such complex data sets (Bishop, 2006; Mackay, 2003). The goal of feature selection is to identify a subset of features relevant for statistical tasks such as prediction or classification. Feature selection is a difficult problem because features are often redundant and strongly correlated with each other. Furthermore, the relevance of a feature depends not only on the data set being analyzed, but also the statistical techniques being employed. While powerful methods for feature selection exist for a handful of techniques such as linear regression (Tibshirani, 1996; Zou & Hastie, 2005), there is no unified framework for performing feature selection in a computationally tractable manner. To address this shortcoming,

Neural Computation **27, 2411–2422** (2015) © 2015 Massachusetts Institute of Technology
doi:10.1162/NECO_a_00780

we present a unified approach for Bayesian feature selection that is applicable to a large class of commonly used statistical models.

In principle, Bayesian inference provides a unified framework for feature selection (O'Hagan, Forster, & Kendall, 2004; Mackay, 2003). Unfortunately, Bayesian methods often require extensive Monte Carlo simulations that become computationally intractable for very high-dimensional problems. In the Bayesian framework, a statistical model is defined by its likelihood function, $p_x(y|\boldsymbol{\theta})$, which describes the observed data, y, as a function of some features, $\boldsymbol{x}$, and model parameters, $\boldsymbol{\theta}$. This likelihood function is supplemented by a prior, $p(\boldsymbol{\theta})$, that encodes our belief about the model parameters in the absence of data. Using Bayes' rule, one can define the posterior probability, $p_x(\boldsymbol{\theta}|y) \propto p_x(y|\boldsymbol{\theta})p(\boldsymbol{\theta})$, that describes our belief about the parameter $\boldsymbol{\theta}$ given the observed data y. Notice that for a flat (constant) prior, maximizing the posterior distribution is equivalent to maximizing the likelihood, and Bayesian inference reduces to the usual maximum likelihood framework. More generally, the log of the prior can be interpreted as a penalty function that "regularizes" the model parameters. For example, the choice of a gaussian and Laplace prior corresponds to imposing an $L2$ and $L1$ penalty on model parameters, respectively (Bishop, 2006).

Since the ultimate goal of feature selection is to identify a subset of features, it is helpful to introduce a set of indicator variables, $\boldsymbol{s}$, that indicates whether a feature is included in the statistical model, with $s_i = 1$ if feature i is included and $s_i = -1$ if it is not. The posterior distribution for $\boldsymbol{s}$ can be computed as (MacKay, 1992)

$$p_x(\boldsymbol{s}|y) \propto p_x(y|\boldsymbol{s})p_0(\boldsymbol{s}), \tag{1.1}$$

where the $p_0(\boldsymbol{s})$ is a prior distribution describing any a priori information we have about which variables are relevant, and $p_x(y|\boldsymbol{s})$, often referred to as the evidence, is given by

$$p_x(y|\boldsymbol{s}) = \int d\boldsymbol{\theta}\, p_x(y|\boldsymbol{\theta}, \boldsymbol{s})p(\boldsymbol{\theta}|\boldsymbol{s}). \tag{1.2}$$

For simplicity, we will assume that $p_0(\boldsymbol{s}) \propto 1$ is a flat prior for the rest of this letter; it is easy to extend our results to other cases. Feature selection can be performed by choosing features with large posterior probabilities.

Our central result is that the logarithm of the evidence maps to the energy of an Ising model for a large class of priors (i.e., $p(\boldsymbol{\theta})$) we term strongly regularizing priors. Thus, computing the marginal posterior probabilities for the relevance of each feature reduces to computing the magnetizations of an Ising model. The key property of strongly regularizing priors is that they affect the posterior probability even when the sample size goes to infinity (Fisher & Mehta, 2015). This should be contrasted with the usual assumption of Bayesian inference that the effect of the prior on posterior

probabilities diminishes inversely with the number of samples (Mackay, 2003). We expect that strongly regularizing priors are especially useful for feature selection problems where the number of potential features is greater than, or comparable to, the number of data points. Surprisingly, our results do not depend on the specific choice of prior or likelihood function, under some mild conditions, suggesting that Bayesian feature selection has universal properties for strongly regularizing priors.

Practically, our Bayesian Ising approximation (BIA) provides an efficient method for computing the posterior probabilities necessary for feature selection with many commonly employed statistical procedures such as generalized linear models (GLMs). We envision using the BIA as part of a two-stage procedure where it is applied to rapidly screen out irrelevant candidates—those features that have low rank in posterior probability—before applying a more computationally intensive cross-validation procedure to infer the model parameters with the remaining features.

2 General Formulas

For concreteness, consider a series of n independent observations of some dependent variable y, $\boldsymbol{y} = (y_1, \ldots, y_n)$, that we want to explain using a set of p potential features $\boldsymbol{x} = (x_1, \ldots, x_p)$. Furthermore, assume that to each of these potential features, x_i, we associate a model parameter θ_i. Let S denote an index set specifying the positions j for which $s_j = +1$. The cardinality of S, that is, the number of indicator variables with $s_j = +1$, is denoted M. Also, we define a vector $\tilde{\boldsymbol{\theta}}$ of length M that contains the model parameters corresponding to the active features, S. With these definitions, we define the log likelihood for the observed data as a function of the active features as $\mathcal{L}_X(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}) = -\log p_X(\boldsymbol{y}|\tilde{\boldsymbol{\theta}})$. Throughout, we assume that the log likelihood is twice differentiable with respect to $\boldsymbol{\theta}$.

In addition to the log likelihood, we need to specify prior distributions for the parameters. Here, we will work with factorized priors of the form

$$P(\tilde{\boldsymbol{\theta}}) = \prod_{j \in S} \frac{1}{Z(\lambda)} e^{-\lambda f(\bar{\theta}_j)}, \tag{2.1}$$

where λ is a parameter that controls the strength of the regularization, $f(\theta_j)$ is a twice differentiable convex function minimized at the point $\bar{\theta}_j$ with $f(\bar{\theta}_j) = 0$, and $Z(\lambda)$ is the normalization constant. As the function is convex, we are assuming that the second derivative evaluated at $\bar{\theta}_j$, denoted $\partial^2 f$, is positive. Many commonly used priors take this form, including gaussian and hyperbolic priors. Plugging these expressions into equation 1.2, yields an expression for the evidence where only the subset of features, S, is

included:

$$p_X(\boldsymbol{y}|S) = Z^{-M}(\lambda) \int d\tilde{\boldsymbol{\theta}} e^{\mathcal{L}_X(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}) - \lambda \sum_{j\in S} f(\tilde{\theta}_j)}. \tag{2.2}$$

By definition, this integral is dominated by the second term for strongly regularizing priors. Since the log likelihood is extensive in the number of data points, n, this generally requires that the regularization strength be much larger than the number of data points, $\lambda \gg n$. For such strongly regularizing priors, we can perform a saddle-point approximation for the evidence. This yields, up to an irrelevant constant,

$$\begin{aligned}\mathcal{L}_X(\boldsymbol{y}|S) &= \log p_X(\boldsymbol{y}|S) \\ &= -\frac{1}{2}\log|I - \Lambda^{-1}H| + \frac{\Lambda^{-1}}{2}\boldsymbol{b}'(I - \Lambda^{-1}H)^{-1}\boldsymbol{b},\end{aligned} \tag{2.3}$$

where $\Lambda = \lambda(\partial^2 f)$ is a renormalized regularization strength,

$$\boldsymbol{b} = \nabla \mathcal{L}_X(\tilde{\boldsymbol{\theta}}|\boldsymbol{y}) \tag{2.4}$$

is the gradient of the log likelihood, and

$$H_{ij} = \frac{\partial^2 \mathcal{L}_X(\tilde{\boldsymbol{\theta}}|\boldsymbol{y})}{\partial\tilde{\theta}_i \partial\tilde{\theta}_j} \tag{2.5}$$

is the Hessian of the log likelihood, with all derivatives evaluated at $\tilde{\theta}_j = \bar{\theta}_j$. We emphasize that the large parameter in our saddle-point approximation is the regularization strength $\Lambda \gg n$. This differs from previous statistical physics approaches to Bayesian feature selection that commonly assume that $n \gg \Lambda$. In that case, one can take n as the large parameter in the saddle-point approximation, which leads to an approximation for the evidence related to the Bayesian information criterion (Kinney & Atwal, 2014; Balasubramanian, 1997; Nemenman & Bialek, 2002). The fundamental reason for this difference is that we work in the strongly regularizing regime where the prior is assumed to be important even in the infinite data limit.

Since for strongly regularizing priors, Λ is the largest scale in the problem, we can expand the log in equation 2.3 in a power series expansion in $\epsilon = \Lambda^{-1}$ to order $O(\epsilon^3)$ to get

$$\mathcal{L}_X(\boldsymbol{y}|S) \simeq \frac{\epsilon}{2}(\mathrm{Tr}[H] + \boldsymbol{b}'\boldsymbol{b}) + \frac{\epsilon^2}{2}\left(\frac{1}{2}\mathrm{Tr}[H^2] + \boldsymbol{b}'H\boldsymbol{b}\right). \tag{2.6}$$

One of the most striking aspects of this expression is that it is independent of the detailed form of the prior function, $f(\theta)$. All that is required is that the prior is a twice differential convex function with a global minimum at $f(\bar{\theta}_j) = 0$. Different choices of prior simply "renormalize" the effective regularization strength Λ.

Rewriting this expression in terms of the spin variables, $\boldsymbol{s}$, we see that the evidence takes the Ising form with

$$\mathcal{L}_X(\boldsymbol{y}|\boldsymbol{s}) = \epsilon \left(\sum_i s_i h_i + \sum_{ij} J_{ij} s_i s_j \right) \tag{2.7}$$

and couplings given by

$$h_i = \frac{1}{4}(H_{ii} + b_i^2) + 2\sum_j J_{ij}, \tag{2.8}$$

$$J_{ij} = \frac{\epsilon}{16}(H_{ij}^2 + 2b_i H_{ij} b_j). \tag{2.9}$$

Notice that the couplings, which are proportional to the small parameter ϵ, are weak. According to Bayes' rule, $\mathcal{L}(\boldsymbol{s}|\boldsymbol{y}) = \mathcal{L}(\boldsymbol{y}z\ \boldsymbol{s}) + constant$ if the prior on $\boldsymbol{s}$ is flat, so the posterior probability of a set of features $\boldsymbol{s}$ is described by an Ising model of the form

$$p(\boldsymbol{s}|\boldsymbol{y}) = \mathcal{Z}^{-1} e^{\sum_i s_i h_i + \sum_{ij} s_i s_j J_{ij}}, \tag{2.10}$$

where $\mathcal{Z}$ is the partition function that normalizes the probability distribution. We term this the Bayesian Ising approximation (BIA) (Fisher & Mehta, 2015). Finally, for future use, it is useful to define a scale Λ^* for which the Ising approximation breaks down. This scale Λ^* can be computed by requiring that the power series expansion used to derive equation 2.6 converges (Fisher & Mehta, 2015).

We demonstrated the utility of the BIA for feature selection in the specific context of linear regression in a recent paper and can adapt that machinery to the current problem (Fisher & Mehta, 2015). It is useful to explicitly indicate the dependence of various expression on the regularization strength Λ. We want to compute the marginal posterior probability that a feature, j, is relevant:

$$p_\Lambda(s_j = 1|y) \simeq (1 + m_j(\Lambda))/2, \tag{2.11}$$

where we have defined the magnetizations $m_j(\Lambda) = \langle s_j \rangle$. While there are many techniques for calculating the magnetizations of an Ising model, we

focus on the mean field approximation, which leads to a self-consistent equation (Opper & Winther, 2001):

$$m_i(\Lambda) = \tanh\left[\frac{n^2}{4\Lambda}\left(h_i(\Lambda) + \frac{1}{2}\sum_{j\neq i} J_{ij}(\Lambda) m_j(\Lambda)\right)\right] \tag{2.12}$$

Our expressions depend on a free parameter (Λ) that determines the strength of the prior distribution. As it is usually difficult, in practice, to choose a specific value of Λ ahead of time, it is often helpful to compute the feature selection path—that is, compute $m_j(\Lambda)$ over a wide range of Λ's. Indeed, computing the variable selection path is a common practice when applying other feature selection techniques such as Lasso regression (Tibshirani, 1996). To obtain the mean field variable selection path as a function of $\epsilon = 1/\Lambda$, we notice that $\lim_{\epsilon\to 0} m_j(\epsilon) = 0$ and so define the recursive formula,

$$m_i(\epsilon + \delta_\epsilon) = \tanh\left[\frac{(\epsilon+\delta_\epsilon)n^2}{4}\left(h_i(\epsilon+\delta_\epsilon) + \frac{1}{2}\sum_{j\neq i} J_{ij}(\epsilon+\delta_\epsilon) m_j(\epsilon)\right)\right], \tag{2.13}$$

with a small step size $\delta_\epsilon \ll 1/\Lambda^*$. We have set $\delta_\epsilon = 0.05/\Lambda^*$ in all of the examples we present.

3 Examples

Logistic regression is a commonly used statistical method for modeling categorical data (Bishop, 2006). To simplify notation, it is also useful to define an extra feature variable $x_0 = 1$ that is always equal to 1 and a $p+1$-dimensional vector of feature, $\boldsymbol{x} = (x_0, x_1, \ldots, x_p)$ with corresponding parameters, $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)$. In terms of these vectors, the likelihood function for logistic regression takes the compact form

$$P_x(y = 1|\boldsymbol{\theta}) = 1 - P_x(y = 1|\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\theta}'\boldsymbol{x}}}{1 + e^{\boldsymbol{\theta}'\boldsymbol{x}}}, \tag{3.1}$$

where $\boldsymbol{\theta}'$ is the transpose of $\boldsymbol{\theta}$. If we have n independent observations of the data (labeled by index $l = 1 \ldots n$), the log likelihood can be written as

$$\mathcal{L}_X(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_l y_l(\boldsymbol{\theta}'\boldsymbol{x}^l) - \log\left(1 + e^{\boldsymbol{\theta}'\boldsymbol{x}^l}\right). \tag{3.2}$$

We supplement this likelihood function with an L^2 norm on the parameters of the form

$$p(\boldsymbol{\theta}) = \prod_{j=1}^{p} \sqrt{\frac{\lambda}{2\pi}} e^{-\lambda(\theta_j - \bar{\theta}_j)^2/2} \tag{3.3}$$

with $\bar{\boldsymbol{\theta}} = (\bar{\theta}_0, 0, \ldots, 0)$ and where $\bar{\theta}_0$ is chosen to match the observed probability that $y = 1$ in the data:

$$P_{obs}(y). = \frac{e^{\bar{\theta}_0}}{1 + e^{\bar{\theta}_0}}. \tag{3.4}$$

Using these expressions, we can calculate the gradient and the Hessian of the log likelihood:

$$b_i \equiv \frac{\partial \mathcal{L}_X(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \theta_i}|_{\bar{\boldsymbol{\theta}}} = \sum_l x_i^l (y^l - P_{obs}(y)) \tag{3.5}$$

and

$$H_{ij} = -\left(P_{obs}(y)(1 - P_{obs}(y))\right) \cdot \sum_l x_i^l x_j^l. \tag{3.6}$$

Plugging these into equation 2.9 yields the fields and couplings for the Ising model in equation 2.10:

$$h_i = \frac{1}{4}\left(\left(\sum_l x_i^l \Delta y^l\right)^2 - P_{obs}(y)(1 - P_{obs}(y)) \sum_l x_i^l x_j^l\right) + 2\sum_j J_{ij}, \tag{3.7}$$

$$J_{ij} = \frac{\epsilon}{16}\left(P_{obs}(y)(1 - P_{obs}(y)) \sum_l x_i^l x_j^l\right)^2 - \frac{\epsilon}{8}\left(P_{obs}(y)(1 - P_{obs}(y)) \sum_l x_i^l x_j^l\right)\left(\sum_{l'} x_i^{l'} \Delta y^{l'}\right)\left(\sum_{l''} x_j^{l''} \Delta y^{l''}\right), \tag{3.8}$$

where we have defined $\Delta y^l = y^l - P_{obs}(y)$.

Notice that the gradient, b_i, is proportional to the correlations between the y and x_i. Furthermore, except for a multiplicative constant reflecting the variance of y, H_{ij} is just the correlation between x_i and x_j. Thus, as in linear regression (Fisher & Mehta, 2015), the coefficients of the Ising model are related to the correlations between variables or the data, or both. In fact,

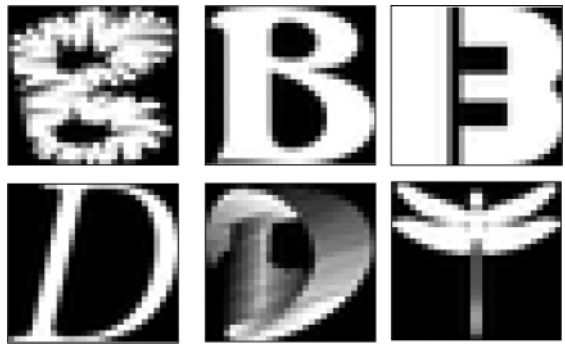

Figure 1: Three randomly chosen examples of Bs (top) and Ds (bottom) from the notMNIST data set. Each letter in the notMNIST data set is represented by a 28×28 pixel grayscale image.

it is easy to show this is the case for all generalized linear models (see the appendix).

To illustrate our approach, we used the BIA for a logistic regression-based classifier designed to classify Bs and Ds in the notMNIST data set, which consists of diverse images of the letters A to J composed from publicly available computer fonts. Each letter is represented as a 28×28 grayscale image where pixel intensities vary between 0 and 255. Figure 1 shows three randomly chosen examples of the letters B and D from the notMNIST data set. The BIA was performed using 500 randomly chosen examples of each letter. Notice that the number of examples (500 or each letter) is comparable to the number of pixels (784) in an image, suggesting that strongly regularizing the problem is appropriate.

Using expressions 3.7 and 3.8, we calculated the couplings for the Ising model describing our logistic-regression-based classifier and calculated the feature selection path as a function of Λ/Λ^* using the mean field approximation. As in linear regression, we used $\Lambda^* = n(1 + pr)$, where $n = 1000$ is the number of samples, $p = 784$ is the number of potential features, and r is the root-mean-squared correlation between pixels (see Figure 2A). Figure 2B shows the posterior probability of all 784 pixels when $\Lambda = \Lambda^*$. To better visualize this, we have labeled the pixels with the highest posterior probabilities in red in the feature selection path in Figure 2A and in the sample images shown in Figure 2C. The results agree well with our basic intuition about which pixels are important for distinguishing the letters B and D.

4 Discussion and Conclusion

We have presented a general framework for Bayesian feature selection in a large class of statistical models. In contrast to previous Bayesian approaches

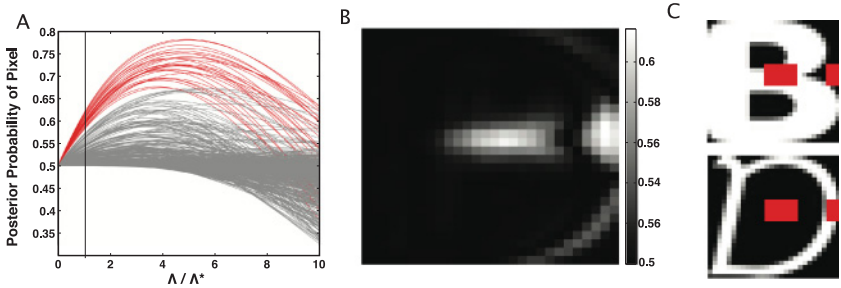

Figure 2: Identifying the most informative pixels for distinguishing Bs and Ds. (A) We used the Bayesian Ising approximation (BIA) to perform feature selection for a logistic regression–based classifier designed to classify Bs and Ds. The model was trained using 500 randomly chosen examples of each letter from the notMNIST data set. The resulting feature selection path is shown as a function of the regularization strength, Λ. Each line corresponds to one of 784 pixels. We expect the BIA to break down when $\Lambda < \Lambda^*$. Pixels in red have the highest posterior probability and hence can be used for distinguishing between Bs and Ds. (B) Posterior probabilities of all 784 pixels when $\Lambda/\Lambda^* = 1$. (C) Visualization of the most informative pixels labeled in red in panel A.

that assume that the effect of the prior vanishes inversely with the amount of data, we have used strongly regularizing priors that have large effects on the posterior distribution even in the infinite data limit. We have shown that in the strongly regularized limit, Bayesian feature selection takes the universal form of an Ising model. Thus, the marginal posterior probabilities that each feature is relevant can be efficiently computed using a mean field approximation. Furthermore, for generalized linear models, we have shown the coefficients of the Ising model can be calculated directly from correlations between the data and features.

Methods for feature selection, including the BIA, generally attempt to identify a subset of features in which each feature is strongly correlated to the response but only weakly correlated to the other selected features. Most of these feature selection algorithms, such as the Lasso, identify a subset of features that minimizes a chosen objective function (Tibshirani, 1996; Zou & Hastie, 2005). This can also be viewed as finding a subset of features with maximum posterior probability with a suitably chosen likelihood and prior. It is typically difficult, however, to assess the significance of the selected features obtained from these maximum a posteriori methods (Lockhart, Taylor, Tibshirani, & Tibshirani, 2014). By contrast, the BIA directly computes the marginal posterior probabilities that each of the features is relevant. Thus, selecting a subset of features with the BIA is equivalent to choosing a threshold for statistical significance in a Bayesian framework in which significance is measured by posterior probabilities.

Like other approaches to feature selection, the BIA has a hyperparameter (Λ) that must be specified. Here, we have set $\Lambda = \Lambda^*$ because it is the weakest prior for which the mapping to the Ising model still holds. In practical applications, however, it will be worthwhile to explore other approaches to specifying a value for Λ by maximizing the evidence, or through cross-validation. Moreover, we have used a flat prior on $\boldsymbol{s}$ throughout this work, but it is trivial to extend our results to other cases. Including a prior on $\boldsymbol{s}$ would introduce other hyperparameters, which also have to be specified.

Surprisingly, aside from some mild regularity conditions, our approach is independent of the choice of prior or likelihood. This suggests that it may be possible to obtain general results about strongly regularizing priors. It will be interesting to explore Bayesian inference in this new limit. Our approach also gives a practical algorithm for quickly performing feature selection in many commonly employed statistical and machine learning approaches. The methods outlined here are especially well suited for modern data sets where the number of potential features can vastly exceed the number of independent samples. We envision using the Bayesian feature selection algorithm outlined here as part of a two-stage procedure. One can use the BIA to rapidly screen out irrelevant candidates and reduce the complexity of the data set before applying a more comprehensive cross-valid procedure. More generally, it will be interesting to develop statistical physics–based approaches for the analysis of complex data.

Appendix: Formulas for Generalized Linear Models

The gradient, $\boldsymbol{b}$, and Hessian, H, of the log likelihood have particularly simple definitions for generalized linear models (GLMs) that extend the exponential family of distributions. In the exponential family, we can write the distribution in the form

$$p(y|\theta) = g(y)e^{\boldsymbol{\eta}(\theta)\boldsymbol{T}(y)+F(\eta)}, \tag{A.1}$$

where $\boldsymbol{T}(y)$ is a vector of sufficient statistics and the $\boldsymbol{\eta}$ are called natural parameters. Notice that for these distributions, we have that

$$\frac{\partial F}{\partial \eta_i} = -\langle T_i(y)\rangle \tag{A.2}$$

and

$$\frac{\partial^2 F}{\partial \eta_i \partial \eta_j} = \mathrm{Cov}[T_i(y), T_j(y)] \tag{A.3}$$

where Cov denotes the covariance (connected correlation function).

In a GLM, we restrict ourselves to distribution to scalar quantities where $T(y) = y$ and say that $\eta = \boldsymbol{\theta}'\boldsymbol{x}$. Then we can write the likelihood as

$$p(y|\boldsymbol{\theta}, \boldsymbol{x}) = g(y)e^{(\boldsymbol{\theta}'\boldsymbol{x})y + F(\boldsymbol{\theta}'\boldsymbol{x})}. \tag{A.4}$$

If we have n independent data points with $l = 1, \ldots, n$, then we can write the log likelihood for such a distribution as

$$\mathcal{L}_X(\boldsymbol{y}|\boldsymbol{\theta}) = \sum_l \log g(y^l) + (\boldsymbol{\theta}'\boldsymbol{x}^l)y^l + F(\boldsymbol{\theta}'\boldsymbol{x}^l). \tag{A.5}$$

Using the expressions above for the exponential family and equation 2.4, we have that

$$b_i = \frac{\partial \mathcal{L}_X(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \theta_i}|_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}} = \sum_l y^l x_i^l - x_i^l \langle y \rangle_{\bar{\boldsymbol{\theta}}}, \tag{A.6}$$

where $\langle y \rangle_{\bar{\boldsymbol{\theta}}}$ is the expectation value of y for choice of parameter $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$. If we choose $\bar{\theta}$ to reproduce the empirical probability, we get

$$b_i \approx n\mathrm{Cov}[y, x_i]. \tag{A.7}$$

Moreover, the entries of the Hessian are given by

$$H_{ij} = \frac{\partial^2 \mathcal{L}_X(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = -\sum_l x_i^l x_j^l \mathrm{Var}[y]_{\boldsymbol{\theta}=\bar{\boldsymbol{\theta}}}. \tag{A.8}$$

If we consider standardized variables, $\boldsymbol{x}$, then we can write

$$H_{ij} \approx -n\mathrm{Cov}[x_i, x_j]\mathrm{Var}[y]. \tag{A.9}$$

Acknowledgments

We thank Alex Lang, Javad Noorbakhsh, and David Schwab for helpful discussions. This work was supported by a Sloan Research Fellowship and a Simons Investigator Award (to P.M.), and internal funding from Boston University.

References

Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, *9*(2), 349–368.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.
Fisher, C. K., & Mehta, P. (2015). Bayesian feature selection for high dimensional linear regression via the Ising approximation with applications to genomics. *Bioinformatics*, btv037.
Kinney, J. B., & Atwal, G. S. (2014). Parametric inference in the large data limit using maximally informative models. *Neural Computation*, *26*(4), 637–653.
Lockhart, R., Taylor, J., Tibshirani, R. J., & Tibshirani, R. (2014). A significance test for the Lasso. *Annals of Statistics*, *42*(2), 413–468.
MacKay, D. J. (1992). Bayesian interpolation. *Neural Computation*, *4*(3), 415–447.
Mackay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.
Nemenman, I., & Bialek, W. (2002). Occam factors and model independent Bayesian learning of continuous distributions. *Physical Review E*, *65*(2), 026137.
O'Hagan, A., Forster, J., & Kendall, M. G. (2004). *Bayesian inference*. London: Arnold.
Opper, M., & Winther, O. (2001). From naive mean field theory to the TAP equations. In M. Opper & D. Saad (Eds.), *Advanced mean field methods: Theory and practice*. Cambridge, MA: MIT Press.
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Received April 17, 2015; accepted June 29, 2015.