MIT Open Access Articles

Risk and regret of hierarchical Bayesian learners

Risk and Regret of Hierarchical Bayesian Learners

Jonathan H. Huggins   JHUGGINS@MIT.EDU
Computer Science and Artificial Intelligence Laboratory, MIT

Joshua B. Tenenbaum   JBT@MIT.EDU
Brain and Cognitive Science Department, MIT

Abstract

Common statistical practice has shown that the full power of Bayesian methods is not realized until hierarchical priors are used, as these allow for greater "robustness" and the ability to "share statistical strength." Yet it is an ongoing challenge to provide a learning-theoretically sound formalism of such notions that: offers practical guidance concerning when and how best to utilize hierarchical models; provides insights into what makes for a good hierarchical prior; and, when the form of the prior has been chosen, can guide the choice of hyperparameter settings. We present a set of analytical tools for understanding hierarchical priors in both the online and batch learning settings. We provide regret bounds under log-loss, which show how certain hierarchical models compare, in retrospect, to the best single model in the model class. We also show how to convert a Bayesian log-loss regret bound into a Bayesian risk bound for any bounded loss, a result which may be of independent interest. Risk and regret bounds for Student's t and hierarchical Gaussian priors allow us to formalize the concepts of "robustness" and "sharing statistical strength." Priors for feature selection are investigated as well. Our results suggest that the learning-theoretic benefits of using hierarchical priors can often come at little cost on practical problems.

1. Introduction

There are two standard justifications for the use of hierarchical models. The first is that they allow for the representation of greater uncertainty by placing "hyperpriors" on

the hyperparameters of the prior distribution (Berger, 1985; Bernardo & Smith, 2000; Gelman et al., 2013). By explicitly modeling the additional uncertainty, there is greater "robustness" to misspecification and unexpected data. The second is that hierarchical models permit the "sharing of statistical strength" between related observations or cohorts (Gelman et al., 2013). For example, take the recent "Big Bayes Stories" special issue of the journal *Statistical Science*, which was comprised of short articles describing successful applications of Bayesian models to a diverse range of problems, including political science, astronomy, and public health (Mengersen & Robert, 2014). Most of the Bayesian models were hierarchical, and the need for robustness and sharing of statistical strength because of limited data were commonly cited reasons by the practitioners for choosing a hierarchical Bayesian approach. Gelman & Hill (2006) and Gelman et al. (2013) both contain further examples of problems in which hierarchical modeling is critical to obtaining high-quality inferences.

Within the machine learning and vision literature, Salakhutdinov et al. (2011) offers an illustrative case study in the benefits and the pitfalls of employing a hierarchical model. The motivation of Salakhutdinov et al. (2011) was that, for image classification tasks, some categories of objects (e.g., "car" or "dog") have many labeled positive and negative examples while other, visually related, categories (e.g., "bus" or "anteater") have only a few labeled examples. Fig. 1(right, a) shows the distribution of training examples for the 200 object categories used while Fig. 1(right, b) shows the same distribution, but now objects are grouped with those with similar appearances. In both cases, the distributions are fat-tailed: there are a few categories with many training examples and many categories with a few training examples. It was hypothesized that by using a hierarchical Bayesian model, the classes with large amounts of labeled data could be used to construct better classifiers for the classes with small amounts of labeled data.

The model used by Salakhutdinov et al. (2011), which

we will analyze in Section 4.2, consisted of a hierarchical Gaussian prior with a logistic regression likelihood. Two-level, one-level, and flat priors were all tested. The purpose of using the two-level prior was that it was able to encode information about which object classes had visually similar objects (e.g., car and track, dog and horse). Fig. 1(left) compares the predictive performance of this two-level hierarchical prior with the two more impoverished priors. Observe that the one-level and two-level priors both improve performance on most object classes compared to the flat prior, but not all. Furthermore, the two-level prior always leads to greater improvement than the one-level prior on object classes where a hierarchical model helps, but also almost always leads to a greater degradation in performance on object classes where the hierarchical models decrease performance. Why the different performance characteristics for the two hierarchical models? Why do some categories have improved accuracy while others decreased accuracy? In a post-hoc analysis, Salakhutdinov et al. (2011) note that the "objects with the largest improvement...borrow visual appearance from other frequent objects" while "objects with the largest decrease [such as 'umbrella' and 'merchandise'] are abstract, and their visual appearance is very different from other object categories."

The results just described lead to numerous theoretical questions of practical consequence:

Q1 Can we formalize why for some object classes there was a beneficial sharing of statistical strength, while for other classes the sharing was detrimental?

Q2 Can we understand when a flat model should be preferred to a hierarchical one to avoid unfavorable sharing?

Q3 More generally, can we obtain guidance on the best type of prior for the problem at hand? Perhaps a different hierarchical prior would have been better suited to learning the image classifiers. For example, could placing hyperpriors on the variance parameters lead to greater "robustness" for object categories such as 'umbrella' and 'merchandise,' whose visual appearance differs from other object categories?

Q4 Once the form of the prior has been chosen, how should hyperparameters be set to maximize learning? The settings of the variance hyperparameters was left unspecified by Salakhutdinov et al. (2011), and it is not clear a priori how they should be set, or how much effect their choice will have on learning.

While we have primarily framed these questions in terms of a single model from one paper, this focus was simply for concreteness. Similar results leading to the same types of questions can be found in the numerous articles that make
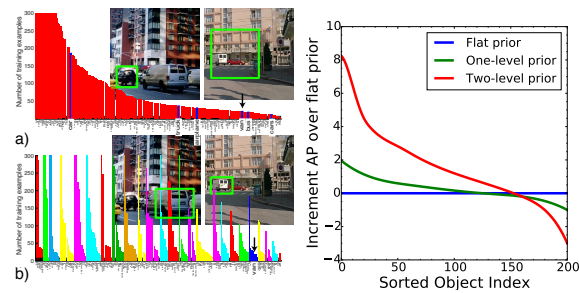

Figure 1. **Right: a)** Distribution of training examples per object class. **b)** Same as a), but with objects grouped by visual appearance. **Left:** Improvement in classification accuracy of hierarchical models compared to flat model. Object categories are sorted by improvement. Reproduced and reconstructed from Salakhutdinov et al. (2011).

use of hierarchical Bayesian methods. For example, one might instead consider the hierarchical models have been used in political science for analyzing polling and census data to predict election outcomes (Ghitza & Gelman, 2013) and in demography for predicting population growth, life expectancy, and fertility rates (Raftery et al., 2013; 2012; Alkema et al., 2011).

In this paper we seek to answer the questions just posed in terms of two learning-theoretic quantities: regret (in online learning) and statistical risk (in batch learning). The online learning setting applies, for example, to the demography applications and election prediction while the batch setting is relevant to the image classification problem as well as election prediction (whether the online or batch analysis applies to election prediction depends on how the problem is formulated).

In the online learning framework (Dawid & Vovk, 1999; Cesa-Bianchi & Lugosi, 2006), no assumptions are made about the data-generating mechanism. Inputs are presented to the learner one by one. After receiving each input the learner predicts the output, then suffers a loss after observing the true output. The goal of the learner is to not do much worse (i.e., have large regret) compared to a fixed class of predictors. Online learning guarantees are attractive for the analysis of hierarchical Bayesian models because such models are so often used in exactly those circumstances when orthodox Bayesian justifications do not apply: typically the modeler does not think that her model reflects the true data generating process, but is instead employing hierarchical methods either to increase robustness against a poor choice of hyperparameters or to speed learning by allowing for the sharing of statistical strength between populations.

Regret bounds, however, do not themselves give any generalization guarantees about how the learner will perform on future data. Statistical risk bounds provide guarantees

about the learner's expected loss on unseen examples by making assumptions about how the data are generated — for example, from an i.i.d. or strongly mixing process. Although there is a stochastic assumption, risk bounds also do not assume that the data is generated according to the model used. We derive a general result for transferring Bayesian regret bounds to risk bounds for bounded losses.

Regret bound for a number of Bayesian models have previously been developed (Vovk, 2001; Kakade & Ng, 2004; Kakade et al., 2005; Banerjee, 2006; 2007; Seeger et al., 2008), with a particular focus on regression and simple priors such as independent Gaussian distributions for each regression coefficient. For a discussion of more general (but asymptotic) Bayesian regret bounds for exponential families and other sufficiently "regular" model classes, see Grünwald (2007, Chapter 8). We follow the approach originally taken in Kakade & Ng (2004), and further explored in Kakade et al. (2005), Banerjee (2006), and Seeger et al. (2008), which applies to a large class of Bayesian generalized linear models (GLMs). We extend the technique to apply to certain non-GLM likelihoods as well, including to multi-class logistic regression. All proofs are deferred to the Appendices.

We answer Questions Q1-Q4 in some important cases by deriving regret bounds for three types of hierarchical priors. First, we consider the use of an inverse gamma hyperprior for the Gaussian prior's variance parameter and demonstrate that the hyperprior leads to greater robustness to data that is well-explained by setting the GLM parameter vector to have very large ℓ^2 norm. Next, we analyze hierarchical Gaussian models that allow for the sharing of statistical strength. Our results, which complement existing work on transfer and multitask learning theory (Baxter, 1997; Ben-David & Schuller, 2003; Pentina & Lampert, 2014), show that when the parameters with small regret for a collection of related tasks are either (a) similar or (b) not unexpected under the prior, then the hierarchical model has a smaller regret bound than assuming the tasks are independent. Finally, we show that spike-and-slab priors can exploit sparse parameters with small regret.

2. Bayesian Online Learning

In online learning, the learner must predict (a distribution over) $y \in \mathcal{Y} \subseteq \mathbb{R}$ after observing $x \in \mathcal{X} \subseteq \mathbb{R}^n$. In this paper, we assume the prediction is made according to a generalized linear model (GLM) $p(y \mid x, \theta) = p(y \mid \theta \cdot x)$, where $\theta \in \Theta \subseteq \mathbb{R}^n$ is a parameter vector to be chosen. GLMs provide significant modeling flexibility, and the class of GLM models and priors we analyze include a range of models used in real-world scientific applications (Gelman & Hill, 2006; Gelman et al., 2013). Two widely

used GLMs are the logistic regression likelihood

$$p(y \mid \theta, x) = \frac{1}{1 + \exp(y\theta \cdot x)}, \quad y \in \{-1, 1\}, \quad (1)$$

and the Gaussian linear regression likelihood $p(y \mid \theta, x) = \mathcal{N}(y \mid \theta \cdot x, \sigma^2)$, $y \in \mathbb{R}$. Since we are taking a Bayesian approach, we place a prior density $p_0(\theta)$ on Θ, with corresponding distribution P_0.[1] At time step t, the learner observes x_t, outputs a distribution over $\mathcal{Y}$, then observes $y_t \in \mathcal{Y}$. The Bayesian (model average) learner predicts $p(y \mid x_t, Z_{t-1})$, where $Z_t \triangleq \{(x_1, y_1), \ldots, (x_t, y_t)\}$, and then suffers the log-loss $-\ln p(y_t \mid x_t, Z_{t-1})$. Hence, the cumulative loss incurred is

$$L_{Bayes}(Z_T) \triangleq \sum_{t=1}^{T} -\ln p(y_t \mid x_t, Z_{t-1}).$$

If Q is a distribution over θ, then using Q for prediction leads to loss on example t of $\ell_t(Q) \triangleq \mathbb{E}_Q[-\ln p(y_t \mid x_t, \theta)]$ and hence cumulative loss

$$L_Q(Z_T) \triangleq \mathbb{E}_Q \left[\sum_{t=1}^{T} -\ln p(y_t \mid x_t, \theta) \right].$$

If $Q = \delta_{\theta}$, then we write L_{θ} instead of L_Q, so $L_Q(Z_T) = \mathbb{E}_Q[L_{\theta}(Z_T)]$. Our objective is to derive regret bounds of the form

$$\mathcal{R}(Z_T, \theta) \triangleq L_{Bayes}(Z_T) - L_{\theta}(Z_T) \le B(\theta) + C(T), \quad (2)$$

where $\mathcal{R}(Z_T, \theta)$ is the *regret* and $B(\theta) + C(T)$ is a *regret bound* depending on the choice of prior P_0. We aim for $C(T) = o(T)$, so that for a fixed θ, the average loss $T^{-1}L_{Bayes}(Z_T)$ is bounded by $T^{-1}L_{\theta}(Z_T) + o(1)$.

Our approach to bounding $L_{Bayes}(Z_T)$ follows that of previous work on Bayesian GLM regret bounds with log-loss (Kakade & Ng, 2004; Kakade et al., 2005; Seeger et al., 2008), relying on the following well-known result:

Proposition 2.1 (Kakade & Ng (2004); Banerjee (2006)). *The Bayesian cumulative loss is bounded as*

$$L_{Bayes}(Z_T) \le L_Q(Z_T) + \mathrm{KL}(Q||P_0). \quad (3)$$

For the GLM model $p(y \mid \theta \cdot x)$, define $f_y(z) \triangleq -\ln p(y \mid \theta \cdot x = z)$. We make two assumptions throughout the remainder of the paper (they will usually not be stated explicitly):

$$|f_y''(z)| \le c \quad \text{for all } y, z \quad \text{(A1)}$$

$$\|x_t\|_2 \le 1 \quad \text{for all } t. \quad \text{(A2)}$$

The first assumption can be understood as requiring the likelihood to be sufficiently smooth. The second assumption sets the scale of the problem, which is necessary since

---

[1]Throughout, we use lowercase letters for densities and uppercase letters to denote the corresponding measures.

scaling $\boldsymbol{x}_t$ up requires scaling $\boldsymbol{\theta}$ down, and vice-versa: $p(y \,|\, C^{-1}\boldsymbol{\theta} \cdot C\boldsymbol{x}) = p(y \,|\, \boldsymbol{\theta} \cdot \boldsymbol{x})$ for any $C \neq 0$. Note that for the Gaussian linear regression model with variance σ^2 and the logistic regression model, (A1) holds with $c = 1/\sigma^2$ and $c = 1/2$, respectively. Proposition 2.1 leads to the following theorem for obtaining regret bounds for the Bayesian model average learner.

Theorem 2.2 (Bayesian regret meta-theorem). *Let $Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}}$ be a distribution with parameter $\boldsymbol{\phi} \in \Phi \subseteq \mathbb{R}^d$ (written ϕ if $d = 1$) and mean $\boldsymbol{\theta}^*$. If (A1) and (A2) hold, then for all ϕ,*

$$\mathcal{R}(Z, \boldsymbol{\theta}^*) \leq \frac{Tc}{2} \| \operatorname{Var}_{Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}}}[\boldsymbol{\theta}] \| + \operatorname{KL}(Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}} || P_0),$$

where $\| \cdot \|$ is the spectral norm. In particular, if the components of $\boldsymbol{\theta}^*$ are uncorrelated, then $\| \operatorname{Var}_{Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}}}[\boldsymbol{\theta}] \| = \sup_i \operatorname{Var}_{Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}}}[\theta_i]$

Theorem 2.2 is our first main result and will be repeatedly applied in Section 4 by choosing an appropriate $Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}}$ and then optimizing ϕ. Although the bound appears to be linear in T, typically ϕ can be chosen such that $\| \operatorname{Var}_{Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}}}[\boldsymbol{\theta}] \| = \Theta(T^{-1})$ and $\operatorname{KL}(Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}} || P_0) = \Theta(\ln T)$, leading to a logarithmic regret bound. Theorem 2.2 provides an attractive approach to deriving regret bounds because there is no need to work directly with the posterior, which is often analytically intractable. For example, there is no closed-form expression for the posterior of the Bayesian logistic regression model. The theorem generalizes the approach originally taken in Kakade & Ng (2004), in which a Gaussian prior for $\boldsymbol{\theta}$ was considered:

Theorem 2.3 (Gaussian regret (Kakade & Ng, 2004)). *If $\boldsymbol{\theta} \sim \mathcal{N}(0, \sigma^2 I)$, then $\mathcal{R}(Z, \boldsymbol{\theta}^*)$ is bounded by*

$$R_{Bayes}^G(Z, \boldsymbol{\theta}^*) \triangleq \frac{1}{2\sigma^2} \|\boldsymbol{\theta}^*\|^2 + \frac{n}{2} \ln\left(1 + \frac{Tc\sigma^2}{n}\right).$$

2.1. Beyond GLMs

Theorem 2.2 follows from a more general result, Theorem 2.4, which allows for non-GLM likelihoods. Specifically, instead of the likelihood being a GLM, we assume the likelihood can be written in the form $p(y \,|\, \boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\psi}) = p(y \,|\, \boldsymbol{\xi}\boldsymbol{x}, \boldsymbol{\psi})$, where $\boldsymbol{\xi} \in \mathbb{R}^{n' \times n}$ is a matrix and $\boldsymbol{\psi} \in \mathbb{R}^{n''}$. The full parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\xi}, \boldsymbol{\psi}) \in \mathbb{R}^N$, $N \triangleq nn' + n''$ (implicitly flattening the matrix $\boldsymbol{\xi}$). Let $f_y(\boldsymbol{z}) \triangleq -\ln p(y \,|\, (\boldsymbol{\xi}\boldsymbol{x}, \boldsymbol{\psi}) = \boldsymbol{z})$. We require the following assumption in place of (A1):

$$\|f_y''(\boldsymbol{z})\| \leq c \quad \text{for all } y, \boldsymbol{z}, \tag{A1'}$$

where $f_y''(\boldsymbol{z})$ denotes the matrix of second partial derivatives (Hessian).

Theorem 2.4 (Generalized Bayesian regret meta-theorem). *Let $Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}}$ be a distribution with parameter $\boldsymbol{\phi} \in \Phi \subseteq \mathbb{R}^d$*

and mean $\boldsymbol{\theta}^*$. If (A1') and (A2) hold, then for all ϕ,

$$\mathcal{R}(Z, \boldsymbol{\theta}) \leq \frac{Tc(n' + n'')}{2} \| \operatorname{Var}_{Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}}}[\boldsymbol{\theta}] \| + \operatorname{KL}(Q_{\boldsymbol{\theta}^*,\boldsymbol{\phi}} || P_0),$$

Of particular interest is that Theorem 2.4 can handle multi-class logistic regression (MLR). In multi-class regression, each example $\boldsymbol{x}_t$ has one of K labels $y_t \in \{1, \ldots, K\}$, indicating which class the example belongs to. For MLR, each class k has an associated parameter $\boldsymbol{\theta}^{(k)}$. The parameters are combined into a single likelihood:

$$p(y_t \,|\, \boldsymbol{\theta}, \boldsymbol{x}_t) = \frac{\exp(\boldsymbol{\theta}^{(y_t)} \cdot \boldsymbol{x})}{\sum_{k=1}^K \exp(\boldsymbol{\theta}^{(k)} \cdot \boldsymbol{x}_t)}. \tag{4}$$

Theorem 2.5 (MLR Gaussian regret). *If $\boldsymbol{\theta}^{(k)} \sim \mathcal{N}(0, \sigma^2 I)$, $k = 1, \ldots, K$, then using the MLR likelihood guarantees that $\mathcal{R}(Z, \boldsymbol{\theta}^*)$ is bounded by*

$$R_{Bayes}^{mlr-G}(\boldsymbol{\theta}^*, Z) \triangleq \frac{1}{2\sigma^2} \|\boldsymbol{\theta}^*\|^2 + \frac{nK}{2} \ln\left(1 + \frac{TKc\sigma^2}{n}\right).$$

3. Risk Bounds

While online regret bounds are attractive because they make no assumptions about the data-generating process, it is also desirable to have risk bounds in the batch setting since risk bounds provide generalization guarantees for unseen data. We now develop a connection between regret and risk bounds via a PAC-Bayesian analysis (McAllester, 2003; Audibert & Bousquet, 2007; Catoni, 2007). Such bounds also have the benefit of applying to any bounded loss (e.g., the 0-1 loss for binary classification), which may be more task-relevant than the log-loss. In the batch setting, the data Z_T are received all at once by the learner and are assumed to be distributed i.i.d. according to some distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$: $(\boldsymbol{x}_t, y_t) \overset{\text{i.i.d.}}{\sim} \mathcal{D}, t = 1, \ldots, T$. Let ℓ be a bounded loss function taking a probability distribution over $\mathcal{Y}$ and an element of $\mathcal{Y}$ as arguments. Without loss of generality assume $\ell \in [0, 1]$. Writing $\ell_{\boldsymbol{\theta}}(\boldsymbol{x}, y) \triangleq \ell(p(\cdot \,|\, \boldsymbol{x}, \boldsymbol{\theta}), y)$, for any distribution Q over Θ, let

$$\mathcal{L}(Q) \triangleq \mathbb{E}_{(\boldsymbol{x},y) \sim \mathcal{D}} \mathbb{E}_{\boldsymbol{\theta} \sim Q}[\ell_{\boldsymbol{\theta}}(\boldsymbol{x}, y)]$$
$$\hat{\mathcal{L}}(Q, Z_T) \triangleq T^{-1} \sum_{t=1}^T \mathbb{E}_{\boldsymbol{\theta} \sim Q}[\ell_{\boldsymbol{\theta}}(\boldsymbol{x}_t, y_t)]$$

be, respectively, the expected and empirical losses under Q. PAC-Bayesian analyses consider the risk of the Gibbs predictor for the distribution Q (i.e., sample $\boldsymbol{\theta} \sim Q$, predict with $p(\cdot \,|\, \boldsymbol{x}, \boldsymbol{\theta})$), not the model average over Q (i.e., predict with $\int p(\cdot \,|\, \boldsymbol{x}, \boldsymbol{\theta}) Q(\mathrm{d}\boldsymbol{\theta})$). A typical bound (specialized to the Bayesian setting) is the following (here $p_T(\boldsymbol{\theta}) \triangleq p(\boldsymbol{\theta} \,|\, Z_T)$):

Theorem 3.1 (Audibert & Bousquet (2007)). *Fix $\kappa > 1/2$ and write $\kappa' \triangleq 2\kappa/(2\kappa - 1)$. For any distribution $\mathcal{D}$, with*

probability at least $1 - \delta$ *over samples* $(\boldsymbol{x}_t, y_t) \overset{i.i.d.}{\sim} \mathcal{D}$,

$$
\begin{aligned}
&|\mathcal{L}(P_T) - \hat{\mathcal{L}}(P_T, Z_T)| \\
&\leq T^{-1/2} \sqrt{\kappa} \sqrt{\mathrm{KL}(P_T \| P_0) + \ln \kappa'/\delta}.
\end{aligned}
\tag{5}
$$

Combining the PAC-Bayesian risk bound with Bayesian regret bounds leads to our second main result:

Theorem 3.2. *Assume that* (2) *holds and fix* $\kappa > 1/2$. *For any distribution* $\mathcal{D}$, *with probability at least* $1 - \delta$ *over samples* $(\boldsymbol{x}_t, y_t) \overset{i.i.d.}{\sim} \mathcal{D}$,

$$
\begin{aligned}
&|\mathcal{L}(P_T) - \hat{\mathcal{L}}(P_T, Z_T)| \\
&\leq T^{-1/2} \sqrt{\kappa} \sqrt{B(\hat{\boldsymbol{\theta}}) + C(T) + \ln \kappa'/\delta},
\end{aligned}
\tag{6}
$$

where $\hat{\boldsymbol{\theta}} \triangleq \arg\min_{\boldsymbol{\theta}} L_{\boldsymbol{\theta}}(Z_T)$.

An attractive feature of Theorem 3.2 is that the bound does not rely on understanding the posterior P_T, as is required by a direct application of a PAC-Bayesian bound such as that given in Theorem 3.1, which requires calculating $\mathrm{KL}(P_T \| P_0)$. Yet the PAC-Bayesian regret bound remains data-dependent due to its dependence on the empirical risk minimizer (ERM) $p(y \mid \boldsymbol{x}, \hat{\boldsymbol{\theta}})$.

Examining the proof of Theorem 3.2, it is easily seen that in fact any $\tilde{\boldsymbol{\theta}}$ such that $L_{\tilde{\boldsymbol{\theta}}}(Z_T) < L_{P_T}(Z_T)$ can be chosen in place of $\hat{\boldsymbol{\theta}}$. Such alternative choices may lead to significantly tighter bounds and are particularly important, for example, in the application of the theorem to the spike-and-slab prior (cf. Section 4.3), as the ERM parameter will in almost all circumstances satisfy $\|\hat{\boldsymbol{\theta}}\|_0 = n$, which would lead to a poor generalization bound when n is large.

In words, Theorem 3.2 can be understood as stating that if the Bayesian (model average) learner has small log-loss regret compared to the ERM, then with high probability the Bayesian Gibbs predictor will generalize well if the loss function is bounded. Or, as a slogan, the theorem shows that "small regret in the online learning setting implies good generalization bounds in the batch setting." The theorem thus connects PAC-Bayesian bounds, Bayesian regret bounds, and empirical risk minimization.

4. Applications

We now use Theorem 2.2 to analyze hierarchical priors for robustness, sharing of statistical strength, and feature selection.

4.1. Hierarchical Priors for Robustness

In this section we answer questions Q2-Q4 as they relate to hierarchical priors for robust inference, demonstrating how, with a proper choice of hyperparameters, a hierarchical prior can lead to increased robustness compared to a flat prior. Specifically, we analyze a canonical use of a hierarchical prior — to capture greater uncertainty in the value of a parameter by placing a hyperprior on the variance of the Gaussian prior on that parameter (Berger, 1985; Bishop, 2006; Gelman et al., 2013):

$$
\sigma_0^2 \mid \alpha, \beta \sim \Gamma^{-1}(\alpha, \beta) \quad \text{and} \quad \theta_i \mid \mu_0, \sigma_0^2 \sim \mathcal{N}(\mu_0, \sigma_0^2),
$$

where $\Gamma^{-1}(\alpha, \beta)$ is the inverse gamma distribution with shape α and scale β. Let $\nu \triangleq 2\alpha$ and $\sigma^2 \triangleq \beta/\alpha$. Then the marginal distribution of $\boldsymbol{\theta}$ follows the multivariate t-distribution with location $\mu_0 \boldsymbol{1}$, scale matrix $\sigma^2 I$, and ν degrees of freedom:

$$
\boldsymbol{\theta} \mid \mu_0, \sigma^2, \nu \sim \mathcal{T}_\nu(\mu_0 \boldsymbol{1}, \sigma^2 I),
$$

where $\boldsymbol{1}$ is the all-ones vector. The multivariate t-distribution density is

$$
\begin{aligned}
&p_{\mathcal{T}}(\boldsymbol{\theta} \mid \boldsymbol{\mu}, \Sigma, \nu) \\
&= \frac{\Gamma(\frac{\nu+n}{2}) \left(1 + \frac{1}{\nu}(\boldsymbol{\theta} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu})\right)^{-\frac{\nu+n}{2}}}{\Gamma(\frac{\nu}{2}) \pi^{n/2} \nu^{n/2} |\Sigma|^{1/2}}.
\end{aligned}
$$

When ν is finite, the multivariate t-distribution is heavy-tailed: the probability of $\|\boldsymbol{\theta}\|$ decreases at a polynomial rate as $\|\boldsymbol{\theta}\| \to \infty$, compared to the exponential rate for a multivariate Gaussian. For example $\nu = \sigma^2 = 1$ and $\Sigma = \sigma^2$ gives the multivariate Cauchy distribution. A multivariate Gaussian with covariance matrix Σ is recovered by taking $\nu \to \infty$. Placing a multivariate t-distribution prior on $\boldsymbol{\theta}$ yields the following regret bound:

Theorem 4.1 (Multivariate t-distribution regret). *If* $\boldsymbol{\theta} \sim \mathcal{T}_\nu(\boldsymbol{0}, \sigma^2 I)$, *then* $\mathcal{R}(Z, \boldsymbol{\theta}^*)$ *is bounded by*

$$
\begin{aligned}
R_{Bayes}^{mvt}(Z, \boldsymbol{\theta}^*) &\triangleq \frac{\nu + n}{2} \ln\left(1 + \frac{\|\boldsymbol{\theta}^*\|^2}{\nu\sigma^2}\right) \\
&+ \frac{n}{2} \ln\left(\frac{(\nu+1)(\nu+n)}{\nu^2} + \frac{Tc(\nu+1)\sigma^2}{\nu n}\right).
\end{aligned}
\tag{7}
$$

Theorem 2.3 can be obtained as a special case of Theorem 4.1 by taking $\nu \to \infty$.

Assume $\nu \geq 1$. If $\frac{\|\boldsymbol{\theta}^*\|^2}{\nu\sigma^2}$ is small, then

$$
F(\|\boldsymbol{\theta}\|) \triangleq \frac{\nu + n}{2} \ln\left(1 + \frac{\|\boldsymbol{\theta}^*\|^2}{\nu\sigma^2}\right) \approx \frac{n + \nu}{\nu} \frac{\|\boldsymbol{\theta}^*\|^2}{\sigma^2},
$$

so for "small" values of $\|\boldsymbol{\theta}^*\|^2$ (relative to $\nu\sigma^2$) the regret bound behaves similarly to having a Gaussian prior on $\boldsymbol{\theta}$. However, if $\frac{\|\boldsymbol{\theta}^*\|^2}{\nu\sigma^2} \gg 1$, then the regret bound grows only logarithmically with $\|\boldsymbol{\theta}\|$, as we would expect given that the multivariate t-distribution has heavy tails. Roughly speaking, $F(x)$ can be thought of as switching from quadratic to logarithmic behavior when $x^2 = \nu\sigma^2$, since this is the value at which F switches from being convex to concave.

In general, the regret bound is large when the choice of $\boldsymbol{\theta}^*$ with small loss has large magnitude. If a Gaussian prior is used, the possibility of $\|\boldsymbol{\theta}^*\|$ being large can be ameliorated by choosing σ^2 large, since there is only a logarithmic regret penalty in σ. However, without a priori knowledge of how large the optimal $\boldsymbol{\theta}^*$ might be, choosing a multivariate t-distribution prior with a small value for ν and a moderate value for σ^2 allows for guaranteed logarithmic regret in the magnitude of $\boldsymbol{\theta}^*$ no matter how large $\|\boldsymbol{\theta}^*\|$ is. Hence, the use of the hierarchical (multivariate t-distribution) prior does in fact yield greater robustness than the non-hierarchical (Gaussian) prior.[2]

We can, in fact, develop more specific guidance on the choice of hyperparameters for the t-distribution. Our goal is to choose ν such that we obtain a t-distribution regret bound that is essentially as good as the Gaussian prior regret bound $R_{Bayes}^G = \frac{\|\boldsymbol{\theta}^*\|^2}{2\sigma^2} + \frac{n}{2} \ln\left(1 + \frac{Tc\sigma^2}{n}\right)$ for small $\|\boldsymbol{\theta}^*\|$ and better when $\|\boldsymbol{\theta}^*\|$ is large. If we choose ν equal to a constant, then for n much larger than ν, we have $R_{Bayes}^{mvt} \approx \frac{n}{2} \ln\left(1 + \frac{\|\boldsymbol{\theta}^*\|^2}{\nu\sigma^2}\right) + \frac{n}{2} \ln\left(\frac{n}{\nu} + \frac{Tc\sigma^2}{n}\right)$. In the case of $\|\boldsymbol{\theta}^*\|$ small, we therefore have that the first term of R_{Bayes}^{mvt} is approximately $\frac{n}{\nu} \frac{\|\boldsymbol{\theta}^*\|^2}{2\sigma^2}$, and thus larger than the first term of R_{Bayes}^G by a factor of n/ν. Furthermore, for small $T \ll \frac{n}{c\sigma^2}$ and any $\boldsymbol{\theta}^*$, the second term of R_{Bayes}^{mvt} is approximately $\frac{n}{2} \ln(n/\nu)$ whereas the second term of R_{Bayes}^G is approximately $Tc\sigma^2 \ll n$. Thus, R_{Bayes}^{mvt} with constant ν is not competitive with R_{Bayes}^G in the large n and small T regimes. Instead consider the choice $\nu = Cn$ for constant $C > 0$, so

$$
R_{Bayes}^{mvt}
$$
$$
\approx \frac{(C+1)n}{2} \ln\left(1 + \frac{\|\boldsymbol{\theta}^*\|^2}{Cn\sigma^2}\right) + \frac{n}{2} \ln\left(\frac{C+1}{C} + \frac{Tc\sigma^2}{n}\right)
$$
$$
\leq \frac{C+1}{C} \frac{\|\boldsymbol{\theta}^*\|^2}{2\sigma^2} + \frac{n}{2} \ln\left(\frac{C+1}{C} + \frac{Tc\sigma^2}{n}\right).
$$

In this case, by choosing a moderate value of C, we see that a multivariate t-distribution prior with $\nu = Cn$ has a competitive regret bound with a Gaussian prior in the small $\|\boldsymbol{\theta}^*\|$ regime, and exponentially smaller regret bound as $\|\boldsymbol{\theta}^*\|$ becomes large. Furthermore, the t-distribution prior remains competitive with the Gaussian prior when T is small.

---

[2]A more rigorous version of this statement can be obtained for the Gaussian regression likelihood by using the fact that there is a matching lower bound on the regret for the Gaussian prior/Gaussian regression model (Kakade & Ng, 2004).

4.2. Hierarchical Priors for Sharing Statistical Strength

4.2.1. BACKGROUND

We next consider hierarchical priors that allow for the sharing of statistical strength, providing answers to Q1 and Q2: we specify some conditions under which sharing of statistical strength can be achieved and others in which a non-hierarchical prior is preferable.[3] In the machine learning literature, the goal of "sharing statistical strength" has been formalized via multitask learning (MTL) and "learning-to-learn" (LTL) frameworks. A number of theoretical investigations of MTL and LTL haven been carried out, beginning with a series of papers by Baxter (cf. Baxter, 1997; 2000). Generically, such MTL and LTL frameworks involve two or more learning problems that are related to each other in some manner. The learning properties are investigated as the number of tasks and/or the number of examples from each task is increased. Baxter (2000) and Ben-David & Schuller (2003) give sample complexity bounds based on classical ideas from statistical and PAC learning theory. Baxter (1997) examines the asymptotic learning properties of hierarchical Bayesian models. Pentina & Lampert (2014) take a PAC-Bayesian approach while Hassan Mahmud & Ray (2007); Hassan Mahmud (2009), and Juba (2006) develop notions of task-relatedness from an (algorithmic) information-theoretic perspective.

Typically, tasks are equated with probability distributions over examples (e.g., $(\boldsymbol{x}, y)$ pairs). It is assumed that the tasks are drawn i.i.d. from an unknown task distribution. The goal is to learn the individual tasks and learn about the task distribution. Alternatively, a notion of similarity can be used to relate the tasks: the more similar the tasks, the greater the advantage of learning them using multitask algorithms. In the online learning framework no assumptions are made about the distribution of examples, so we consider two MTL scenarios in line with the latter setting. In the first scenario, one example from one task is received at each time step. In the second, which is described in the Appendix E.1, at each time step an example for each task is received simultaneously.

4.2.2. SEQUENTIAL OBSERVATIONS FROM MULTIPLE SOURCES

The sequential observation setting is relevant to the image classification example given in the introduction, in which there are many observations from some data sources and only a small number of observations from numerous other data sources. To model this situation, at time step t, an input $\boldsymbol{x}_t$ from source z_t is observed, where $z_t \in$

---

[3]For simplicity results are for Gaussian priors, though the extension to multivariate t-distribution priors is straightforward.

$\{1, \ldots, K\}$. The learner predicts y_t according to the posterior of $\boldsymbol{\theta}^{(z_t)}$ given Z_{t-1}. An equivalent formulation is that the Bayesian learner observes $\boldsymbol{x}_t = (\mathbf{0}, \ldots, \boldsymbol{x}_t^{(k)}, \mathbf{0}, \ldots)$ (if $z_t = k$) at each time, then receives y_t. Instead of using independent Gaussian priors on $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}$, place a prior over the means of the K priors. For each dimension $j = 1, \ldots, n$, let $\mu_j \,|\, \sigma_0^2 \sim \mathcal{N}(0, \sigma_0^2)$ and $\theta_j^{(k)} \,|\, \mu_j, \sigma^2 \sim \mathcal{N}(\mu_j, \sigma^2)$, $k = 1, \ldots, K$, and write $\boldsymbol{\theta}_j^{(1:K)} \triangleq (\theta_j^{(1)}, \ldots, \theta_j^{(K)})$. Integrating out μ_j yields

$$\boldsymbol{\theta}_j^{(1:K)} \,|\, \sigma_0^2, \sigma^2 \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{8}$$

where, with 1_K denoting the $K \times K$ all-ones matrix,

$$\Sigma \triangleq s^2 \rho 1_K + s^2 (1 - \rho) I \tag{9}$$
$$s^2 \triangleq \sigma_0^2 + \sigma^2, \qquad \rho \triangleq \sigma_0^2 / (\sigma_0^2 + \sigma^2), \tag{10}$$

This prior corresponds to the one-level prior in Salakhutdinov et al. (2011). Similar results, which will be discussed qualitatively below, can be obtained for the two-level prior at the cost of a significantly more complicated bound. Define $T^{(k)} \triangleq \sum_{t=1}^T \delta_k(z_t)$, $Z^{(k)} \triangleq \{(x, y) \in Z | z_t = k\}$, and $\gamma^2 \triangleq K\sigma_0^2 + \sigma^2$

Theorem 4.2 (Hierarchical Gaussian regret, sequential observations). *If* $\boldsymbol{\theta}_j^{(1:K)} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $j = 1, \ldots, n$, *then* $\mathcal{R}(Z, \boldsymbol{\theta}^*)$ *is bounded by*

$$R_{Bayes}^{HG-seq}(Z, \boldsymbol{\theta}^*) \triangleq \frac{1}{2\gamma^2} \sum_{k=1}^K \|\boldsymbol{\theta}^{*(k)}\|^2$$
$$+ \frac{\sigma_0^2}{\sigma^2 \gamma^2} \sum_{k < \ell} \|\boldsymbol{\theta}^{*(k)} - \boldsymbol{\theta}^{*(\ell)}\|^2 + \frac{n}{2} \ln\left(1 + \frac{K\sigma_0^2}{\sigma^2}\right)$$
$$+ \frac{n}{2} \sum_{k=1}^K \ln\left(1 - \frac{\sigma_0^2}{\gamma^2} + \frac{T^{(k)} c\sigma^2}{n}\right). \tag{11}$$

It is instructive to compare the upper bound given in (11) to $\sum_k R_{Bayes}^G(Z_{(k)}, \boldsymbol{\theta}^{*(k)})$ with prior variance $s^2 = \sigma_0^2 + \sigma^2$. Setting $\sigma_0 = \sigma$ yields a condition for the hierarchical model to have smaller regret bound than the non-hierarchical model:

$$4\|\boldsymbol{\theta}^{*(1)} - \boldsymbol{\theta}^{*(2)}\|^2 + 3s^2 n \sum_{k=1}^2 \ln\left(\frac{\frac{4}{3}n + T^{(i)} cs^2}{n + T^{(k)} cs^2}\right)$$
$$\leq \|\boldsymbol{\theta}^{*(1)}\|^2 + \|\boldsymbol{\theta}^{*(2)}\|^2 + 0.863 s^2 n. \tag{12}$$

Of particular interest is the "one-shot learning" scenario, in which only one observation (or a small number of observations) from a source are made while many observations are made from some other sources. This setting is exactly that of the image classification problem of Salakhutdinov et al. (2011). For concreteness, consider a "large data" task with $T^{(1)} \gg \frac{n}{cs^2}$ and a "small data" task $T^{(2)} = 2$, so that

$\ln\left(\frac{\frac{4}{3}n + T^{(1)} cs^2}{n + T^{(1)} cs^2}\right) \approx 0$ and (12) becomes (approximately)

$$4\|\boldsymbol{\theta}^{*(1)} - \boldsymbol{\theta}^{*(2)}\|^2 + 3s^2 n \ln\left(\frac{4n + 6cs^2}{3n + 6cs^2}\right)$$
$$\leq \|\boldsymbol{\theta}^{*(1)}\|^2 + \|\boldsymbol{\theta}^{*(2)}\|^2 + 0.863 s^2 n.$$

But even for $n = 1$, $3\ln\left(\frac{4n + 6cs^2}{3n + 6cs^2}\right) < 0.863$, so the hierarchical model has smaller regret bound as long as $4\|\boldsymbol{\theta}^{*(1)} - \boldsymbol{\theta}^{*(2)}\|^2 \leq \|\boldsymbol{\theta}^{*(1)}\|^2 + \|\boldsymbol{\theta}^{*(2)}\|^2 + Cs^2 n$ for some constant $0 < C < 0.863$.

Hierarchical models for one-shot learning are designed with the goal of providing good predictive power on the new problem (the second data source) even with a small number of examples from that problem. To see if this is in fact the case for the hierarchical prior considered here, we can investigate how much greater the regret bound is for $T^{(2)} > 0$ than for $T^{(2)} = 0$ with $T^{(1)} \gg \frac{n}{cs^2}$ fixed. With $\sigma_0^2 = \sigma^2$, (11) is greater in the former ($K = 2$) than the latter ($K = 1$) scenario by at most

$$-\frac{\|\boldsymbol{\theta}^{*(1)}\|^2}{6s^2} + \frac{\|\boldsymbol{\theta}^{*(2)}\|^2}{3s^2} + \frac{2\|\boldsymbol{\theta}^{*(1)} - \boldsymbol{\theta}^{*(2)}\|^2}{3s^2} + \frac{3T^{(2)} cs^2}{8}.$$

So if $\|\boldsymbol{\theta}^{*(2)}\|$ is small and $\boldsymbol{\theta}^{*(2)}$ and $\boldsymbol{\theta}^{*(1)}$ are close in ℓ^2 distance, the regret bound for the second source is small.

The regret bound for the two-level prior in Salakhutdinov et al. (2011) is quite similar to that for the one-level prior. Let $s_k \in \{1, \ldots, S\}$ denote the superclass of class k. In the case of image classification, object classes that have similar visual appearance would have a common superclass. The two-level prior consistes of an overall parameter prior $\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_0^2 I)$, superclass parameter priors $\boldsymbol{\mu}^{(s)} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma_1^2 I)$, and class parameter priors $\boldsymbol{\theta}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}^{(s_k)}, \sigma_2^2 I)$. The regret bound for the two-level prior is

$$c_0 \sum_{k=1}^K \|\boldsymbol{\theta}^{*(k)}\|^2 + \sum_{k < \ell} c_{k\ell} \|\boldsymbol{\theta}^{*(k)} - \boldsymbol{\theta}^{*(\ell)}\|^2$$
$$+ \frac{n}{2} \sum_{k=1}^K O(\ln(c_1 + c_2 T^{(k)})) + O(1), \tag{13}$$

where $c_0, c_1, c_2 > 0$ are constants, $c_{k\ell} = \tilde{c}_{s_k}$ if $s_k = s_\ell$ and $c_{k\ell} = \tilde{c}_{s_k s_\ell}$ if $s_k \neq s_\ell$. Furthermore, $\tilde{c}_s > \tilde{c}_{s's''}$ for all $s, s', s'' \in \{1, \ldots, S\}$. Hence, the regret bound's being small depends more on the parameter vectors in the same superclass being close to each other than on parameter vectors from different superclasses being close to each other. See Appendix E.2 for details. The two-level regret bound well-explains the results of Salakhutdinov et al. (2011). The poor performance on image classes with very different visual appearance from the other classes is unsurprising since the parameter vectors that predict these classes well are going to have large ℓ^2 distance from the parameter vectors of other object classes.

4.3. Hierarchical Priors for Feature Selection

A shortcoming of the priors investigated so far is the poor dependence on the feature space dimension n. For example, the Gaussian prior regret bound is (approximately)

$$L_{Bayes}(Z) \leq \inf_{\boldsymbol{\theta}^*} L_{\boldsymbol{\theta}^*}(Z) + \frac{1}{2\sigma^2}\|\boldsymbol{\theta}^*\|^2 + \frac{Tc\sigma^2}{2},$$

when $Tc\sigma^2 \ll n$, so the regret may grow linearly in this regime.[4] In the infinite-dimensional case, Gaussian processes can be used while still obtaining meaningful regret bounds (Kakade et al., 2005; Seeger et al., 2008). However, methods that are applicable to high-dimensional problems for which $n \gg T$ but n is still finite are of great general interest. For example, in the image classification example from the introduction, the feature vector has $n \approx 5000$, whereas most object classes have fewer than 200 training examples. In high-dimensional problems it is desirable to use feature selection or sparse methods to reduce the effective dimension of the problem, with the aim of achieving better generalization performance and increasing interpretability of the model. A popular non-Bayesian approach for inducing sparsity is ℓ_1 regularization, such as the lasso for linear regression (Tibshirani, 1996). A Bayesian approach is the Bayesian lasso: the ℓ_1 regularizer of the lasso is converted into a prior, which amounts to placing a Laplace prior on $\boldsymbol{\theta}$ (Park & Casella, 2008). However, the Bayesian lasso still seems to lead to a linear dependence on the dimension because the model puts zero prior mass on a component being exactly zero. A regret bound for the Bayesian lasso can be found in the Appendix F.1 (we suspect that our bound is essentially tight, though we have been unable to obtain a matching lower bound).

Another common Bayesian approach to inducing sparsity is to use a hierarchical "spike and slab" prior, which places positive probability on a component being exactly zero (Ishwaran & Rao, 2005; Narisetty & He, 2014). One version of the spike and slab prior is

$$z_i \,|\, p \sim \mathsf{Bern}(p) \quad \text{and} \quad \theta_i \,|\, z_i \sim z_i\delta_0 + (1-z_i)\mathcal{N}(0, \sigma^2).$$

So with probability p component i is zero and with probability $1-p$ it is Gaussian-distributed. Integrating out z_i yields prior density $p_0(\theta_i) = p\delta_0(\theta_i) + (1-p)\mathcal{N}(\theta_i \,|\, 0, \sigma^2)$. Let $\|\boldsymbol{v}\|_0$ denote the ℓ_0 norm of the vector $\boldsymbol{v}$.

Theorem 4.3. *For the spike-and-slab prior, if $m = \|\boldsymbol{\theta}^*\|_0$, then $\mathcal{R}(Z, \boldsymbol{\theta}^*)$ is bounded by*

$$R_{Bayes}^{SS}(Z, \boldsymbol{\theta}^*) \triangleq \frac{1}{2\sigma^2}\|\boldsymbol{\theta}^*\|^2 + m\ln\frac{1}{1-p} \qquad (14)$$

$$+ (n-m)\ln\frac{1}{p} + \frac{m}{2}\ln\left(1 + \frac{Tc\sigma^2}{m}\right).$$

---

[4]Dimension-independent regret bounds for the priors already considered can be obtained, but at the price of a constant greater than one front of the $L_{\boldsymbol{\theta}^*}(Z)$ term. See, e.g., Banerjee (2007).

In particular, if $p \triangleq q^{1/n}$ for some constant $0 < q < 1$, then $R_{Bayes}^{SS}(Z, \boldsymbol{\theta}^*)$ is at most

$$\frac{\|\boldsymbol{\theta}^*\|^2}{2\sigma^2} + m\ln\frac{n}{1-q} + \ln\frac{1}{q} + \frac{m}{2}\ln\left(1 + \frac{Tc\sigma^2}{m}\right). \quad (15)$$

The theorem shows the importance of properly scaling p with the dimension of the problem. If p is kept fixed, then the regret has linear dependence on n. However, by scaling p to be $q^{1/n}$, we increase the probability of a component being zero as the dimension increases and thus are able to ensure that the regret is only logarithmic in n while simultaneously maintaining the appropriate linear dependence on m. The constant q turns out to be the prior probability that all of the components are set to zero. More generally, $\binom{n}{k}q^{(n-k)/n}(1 - q^{1/n})^k \overset{n\to\infty}{\to} \frac{q\ln^k q^{-1}}{k!}$, the limiting prior probability of choosing exactly k components to be non-zero. Hence, as $n \to \infty$, the prior over the number of non-zero components converges to a Poisson distribution with rate parameter $\ln q^{-1}$. So when n is large the expected number of non-zero components is $\approx \ln q^{-1}$. The choice of p close to 1 for large n is in notable contrast to the common practice of setting $p = 1/2$ or some other constant independent of n (Schneider & Corcoran, 2004; Ishwaran & Rao, 2005). Our results strongly recommend against this practice. See Narisetty & He (2014) for a discussion of purely statistical reasons to scale p with the dimension.

5. Conclusion

In this paper we set out to understand and quantify the learning-theoretic benefits of Bayesian hierarchical modeling. In Section 4, we used first our main result, Theorem 2.2, to analyze three specific hierarchical priors that, particularly when combined with a logistic or Gaussian regression likelihood, are widely used in practice. Indeed, these prior-likelihood combinations have often been used with substantial success even in situations when they are known to be rather poor models for the data generating mechanism. Our analysis offers an explanation for this success. The priors we analyzed are representative of the variety of ways in which hierarchical models are employed: representing uncertainty in hyperparameters, tying together related groups of observations, and creating more complicated distributions from simpler ones. Thus, our results answer Questions Q1-Q4 in some important cases and exemplify a learning-theoretic analysis technique that can be applied to other hierarchical models. In addition, using our second main result, Theorem 3.2, all of the insights gained in Section 4 for the log-loss regret setting apply equally well to the batch setting of statistical risk with bounded loss, further extending the applicability of our conclusions.

Acknowledgments

References

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., and Heilig, G. K. Probabilistic Projections of the Total Fertility Rate for All Countries. *Demography*, 48(3):815–839, July 2011.

Audibert, J.-Y. and Bousquet, O. Combining PAC-Bayesian and generic chaining bounds. *The Journal of Machine Learning Research*, 8:863–889, 2007.

Banerjee, A. On Bayesian Bounds. In *International Conference on Machine Learning*, pp. 81–88. ACM, 2006.

Banerjee, A. An Analysis of Logistic Models: Exponential Family Connections and Online Performance. In *International Conference on Data Mining*, pp. 204–215, 2007.

Baxter, J. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine learning*, 28(1):7–39, 1997.

Baxter, J. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

Ben-David, S. and Schuller, R. Exploiting task relatedness for multiple task learning. In *Conference on Learning Theory*, pp. 567–580. Springer, 2003.

Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edition, 1985.

Bernardo, J. M. and Smith, A. F. M. *Bayesian Theory*. Wiley, New York, 2000.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Catoni, O. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture Notes - Monograph Series*. Institute of Mathematical Statistics, 2007.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.

Dawid, A. P. and Vovk, V. G. Prequential probability: Principles and properties. *Bernoulli*, 5(1):125–162, 1999.

Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition, 2013.

Ghitza, Y. and Gelman, A. Deep Interactions with MRP: Election Turnout and Voting Patterns Among Small Electoral Subgroups. *American Journal of Political Science*, 57(3):762–776, February 2013.

Grünwald, P. D. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

Hassan Mahmud, M. M. On universal transfer learning. *Theoretical Computer Science*, 410(19):1826–1846, 2009.

Hassan Mahmud, M. M. and Ray, S. R. Transfer Learning using Kolmogorov Complexity: Basic Theory and Empirical Evaluations. In *Advances in Neural Information Processing Systems*, pp. 361–368, 2007.

Ishwaran, H. and Rao, J. S. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, 33(2):730–773, April 2005.

Juba, B. Estimating Relatedness via Data Compression. In *International Conference on Machine Learning*, pp. 441–448, 2006.

Kakade, S. M. and Ng, A. Y. Online Bounds for Bayesian Algorithms. In *Advances in Neural Information Processing Systems*, 2004.

Kakade, S. M., Seeger, M., and Foster, D. P. Worst-case bounds for Gaussian process models. In *Advances in Neural Information Processing Systems*, pp. 193–200, 2005.

McAllester, D. A. Simplified PAC-Bayesian Margin Bounds. In *Conference on Learning Theory*, pp. 203–215, 2003.

Mengersen, K. L. and Robert, C. P. Big Bayes Stories—Foreword. *Statistical Science*, 29(1):1–1, 2014.

Narisetty, N. N. and He, X. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, April 2014.

Park, T. and Casella, G. The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686, June 2008.

Pentina, A. and Lampert, C. H. A PAC-Bayesian bound for Lifelong Learning. In *International Conference on Machine Learning*, 2014.

Raftery, A. E., Li, N., Ševčíková, H., Gerland, P., and Heilig, G. K. Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, 109(35):13915–13921, 2012.

Raftery, A. E., Chunn, J. L., Gerland, P., and Ševčíková, H. Bayesian Probabilistic Projections of Life Expectancy for All Countries. *Demography*, 50(3):777–801, March 2013.

Salakhutdinov, R., Torralba, A., and Tenenbaum, J. B. Learning to share visual appearance for multiclass object detection. In *Conference on Computer Vision and Pattern Recognition*, pp. 1481–1488. IEEE, 2011.

Schneider, U. and Corcoran, J. N. Perfect sampling for Bayesian variable selection in a linear regression model. *Journal of Statistical Planning and Inference*, 126(1): 153–171, November 2004.

Seeger, M. W., Kakade, S. M., and Foster, D. P. Information Consistency of Nonparametric Gaussian Process Methods. *Information Theory, IEEE Transactions on*, 54 (5):2376–2382, 2008.

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

Vovk, V. Competitive Online Statistics. *International Statistical Review*, 69(2):213–248, 2001.

A. Regret Bounds for Non-GLM Likelihoods

Recall Proposition 2.1, restated here for convenience:

Proposition. *The Bayesian cumulative loss is bounded as*

$$L_{Bayes}(Z_T) \leq L_Q(Z_T) + \mathrm{KL}(Q\|P_0). \tag{A.1}$$

Proof of Theorem 2.4. Fix a choice of $\boldsymbol{\theta}^*$ and $\boldsymbol{\phi}$ and write $Q = Q_{\boldsymbol{\theta}^*, \boldsymbol{\phi}}$. Take a second-order Taylor expansion of f_y about $\boldsymbol{z}^*$, yielding

$$f_y(\boldsymbol{z}) = f_y(\boldsymbol{z}^*) + f_y'(\boldsymbol{z}^*)^\top (\boldsymbol{z} - \boldsymbol{z}^*) + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}^*)^\top f_y''(\boldsymbol{\zeta}(\boldsymbol{z}))(\boldsymbol{z} - \boldsymbol{z}^*),$$

for some function $\boldsymbol{\zeta}$. Let $\boldsymbol{z} = (\boldsymbol{\xi}\boldsymbol{x}, \boldsymbol{\psi})$ with $\boldsymbol{\theta} \sim Q$ and let $\boldsymbol{z}^* = \mathbb{E}[\boldsymbol{z}] = (\boldsymbol{\xi}^*\boldsymbol{x}, \boldsymbol{\psi}^*)$. Hence,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{z}}[f_y(\boldsymbol{z})] &= f_y(\boldsymbol{z}^*) + f_y'(\boldsymbol{z}^*)^\top \boldsymbol{0} + \frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\left[(\boldsymbol{z} - \boldsymbol{z}^*)^\top f_y''(\boldsymbol{\zeta}(\boldsymbol{z}))(\boldsymbol{z} - \boldsymbol{z}^*)\right] \\
&\leq f_y(\boldsymbol{z}^*) + \frac{c}{2}\mathbb{E}_{\boldsymbol{z}}\left[(\boldsymbol{z} - \boldsymbol{z}^*)^\top (\boldsymbol{z} - \boldsymbol{z}^*)\right].
\end{aligned}
$$

Defining

$$\boldsymbol{\omega} \triangleq (\underbrace{\boldsymbol{x}, \ldots, \boldsymbol{x}}_{n' \text{ times}}, \underbrace{1, \ldots, 1}_{n'' \text{ times}}),$$

we next observe that

$$(\boldsymbol{z} - \boldsymbol{z}^*)^\top (\boldsymbol{z} - \boldsymbol{z}^*) = \boldsymbol{\omega}^\top (\boldsymbol{\theta} - \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \boldsymbol{\omega}. \tag{A.2}$$

Letting $\Sigma = \mathrm{Var}[\boldsymbol{\theta}]$, we thus have

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{z}}\left[(\boldsymbol{z} - \boldsymbol{z}^*)^\top (\boldsymbol{z} - \boldsymbol{z}^*)\right] &= \boldsymbol{\omega}^\top \mathbb{E}_{\boldsymbol{\theta}}[(\boldsymbol{\theta} - \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top]\boldsymbol{\omega} \\
&\leq \|\boldsymbol{\omega}\|_2^2 \|\mathbb{E}_{\boldsymbol{\theta}}[(\boldsymbol{\theta} - \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top]\| \\
&= (n'\|\boldsymbol{x}\|_2^2 + n'')\|\Sigma\| \\
&\leq (n' + n'')\|\Sigma\|
\end{aligned}
$$

since it is assumed that $\|\boldsymbol{x}\|_2 \leq 1$. Noting that $L_Q(Z_T) = \sum_t \mathbb{E}_Q[f_{y_t}(\boldsymbol{\xi}\boldsymbol{x}_t, \boldsymbol{\psi})]$ and $L_{\boldsymbol{\theta}^*}(Z_T) = \sum_t f_{y_t}(\boldsymbol{\xi}^*\boldsymbol{x}_t, \boldsymbol{\psi}^*)$, we have

$$L_Q(Z_T) \leq L_{\boldsymbol{\theta}^*}(Z_T) + \frac{Tc(n' + n'')\|\Sigma\|}{2}. \tag{A.3}$$

Combining (A.1) and (A.3) yields the theorem. $\square$

Proof of Theorem 2.2. Follows as a special case of Theorem 2.4 by choosing $n' = 1$ and $n'' = 0$. $\square$

A.1. Application to Multi-class Logistic Regression

For multi-class logistic regression (MLR) $y \in \{1, \ldots, K\}$ is one of K classes, the parameters are $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(k)}\}_{k=1}^K$, and the likelihood is

$$p(y \mid \boldsymbol{\theta}, \boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}^{(y)} \cdot \boldsymbol{x})}{\sum_{k=1}^K \exp(\boldsymbol{\theta}^{(k)} \cdot \boldsymbol{x})}. \tag{A.4}$$

In order to apply Theorem 2.4, we require the following result:

Proposition A.1. *Assumption* (A1') *holds for the MLR likelihood with $c = 1/2$.*

Proof. First note that

$$f_y(\boldsymbol{z}) = -z_y + \ln \sum_{k=1}^{K} e^{z_i}, \tag{A.5}$$

where $z_i = \boldsymbol{\theta}^{(k)} \cdot \boldsymbol{x}$, and hence the Hessian of $f_y(\boldsymbol{z})$ is independent of y:

$$f_y''(\boldsymbol{z}) = \frac{1}{(\sum_{k=1}^{K} e^{z_i})^2} \begin{pmatrix} \sum_{i \neq 1} e^{z_1 + z_i} & -e^{z_1 + z_2} & \cdots & -e^{z_1 + z_K} \\ -e^{z_2 + z_1} & \sum_{i \neq 2} e^{z_2 + z_i} & \cdots & -e^{z_2 + z_K} \\ \vdots & & \ddots & \end{pmatrix} \tag{A.6}$$

Applying Gershgorin's circle theorem, we find that

$$\|f_y''(\boldsymbol{z})\| \le \frac{2 e^{z_1} \sum_{i \neq 1} e^{z_i}}{(\sum_{k=1}^{K} e^{z_k})^2}, \tag{A.7}$$

where with loss of generality we have applied the theorem to the first row of the Hessian. Defining $a \triangleq e^{z_1} \ge 0$ and $b \triangleq \sum_{i \neq 1} e^{z_i} \ge 0$, we have $\|f_y''(\boldsymbol{z})\| \le \frac{2ab}{(a+b)^2}$. Maximization over the positive orthant occurs at $a = b > 0$, so $\|f_y''(\boldsymbol{z})\| \le 1/2$. $\qquad\square$

Reasoning similarly to Theorem E.1, one can easily prove:

Theorem A.2 (Hierarchical Gaussian regret, multi-class regression). *If $\boldsymbol{\theta}_j^{(1:K)} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $j = 1, \dots, n$, then using the MLR likelihood guarantees that $\mathcal{R}(Z, \boldsymbol{\theta}^*)$ is bounded by*

$$\begin{aligned} R_{Bayes}^{mlr-HG}(Z, \boldsymbol{\theta}^*) \triangleq {}& \frac{1}{2\gamma^2} \sum_{k=1}^{K} \|\boldsymbol{\theta}^{*(k)}\|^2 + \frac{\sigma_0^2}{\sigma^2 \gamma^2} \sum_{k < \ell} \|\boldsymbol{\theta}^{*(k)} - \boldsymbol{\theta}^{*(\ell)}\|^2 \\ & + \frac{n}{2} \ln\left(1 + \frac{K\sigma_0^2}{\sigma^2}\right) + \frac{nK}{2} \ln\left(1 - \frac{\sigma_0^2}{\gamma^2} + \frac{T\sigma^2}{2n}\right), \end{aligned} \tag{A.8}$$

where $\gamma^2 \triangleq K\sigma_0^2 + \sigma^2$.

Theorem 2.5 follows as a special case of Theorem A.2 by taking $\sigma_0^2 = 0$.

B. Proof of Theorem 3.2

Since $p_T(\boldsymbol{\theta}) = \frac{p(Y \mid X, \boldsymbol{\theta}) p_0(\boldsymbol{\theta})}{p(Y \mid X)}$,

$$\begin{aligned} \mathrm{KL}(P_T \| P_0) &= \mathbb{E}_{P_T}\left[\ln \frac{p_T(\boldsymbol{\theta})}{p_0(\boldsymbol{\theta})}\right] \\ &= \mathbb{E}_{P_T}\left[\ln \frac{p(Y \mid X, \boldsymbol{\theta})}{p(Y \mid X)}\right] \\ &= L_{Bayes}(Z_T) - L_{P_T}(Z_T). \end{aligned} \tag{B.1}$$

Combining (2) and (B.1) with Theorem 3.1 implies that with probability $1 - \delta$, for all $\boldsymbol{\theta}$,

$$|\mathcal{L}(P_T) - \hat{\mathcal{L}}(P_T, Z_T)| \le \sqrt{\kappa} \sqrt{\frac{L_{\boldsymbol{\theta}}(Z_T) - L_{P_T}(Z_T) + B(\boldsymbol{\theta}) + C(T) + \ln \kappa'/\delta}{T}}.$$

Observing that $L_{\boldsymbol{\theta}^*}(Z_T) < L_{P_T}(Z_T)$, so $L_{\boldsymbol{\theta}^*}(Z_T) - L_{P_T}(Z_T) < 0$, completes the proof.

C. KL Divergence Derivations

C.1. Multivariate Gaussians

Let $D_i = \mathcal{N}(\mu_i, \Sigma_i), i = 1, 2$, where $\dim(\mu_i) = n$. Then

$$
\begin{aligned}
\mathrm{KL}(D_1 \| D_2) &= \frac{1}{2} \mathbb{E}_{D_1} \left[\ln \frac{|\Sigma_2|}{|\Sigma_1|} - (x - \mu_1)^\top \Sigma_1^{-1}(x - \mu_1) + (x - \mu_2)^\top \Sigma_2^{-1}(x - \mu_2) \right] \\
&= \frac{1}{2} \left\{ \ln \frac{|\Sigma_2|}{|\Sigma_1|} + \mathbb{E}_{D_1} \left[-\mathrm{Tr}(\Sigma_1^{-1}(x - \mu_1)^\top(x - \mu_1)) + \mathrm{Tr}(\Sigma_2^{-1}(x - \mu_2)^\top(x - \mu_2)) \right] \right\} \\
&= \frac{1}{2} \left\{ \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \mathrm{Tr}(\Sigma_1^{-1}\Sigma_1) + \mathbb{E}_{D_1} \left[\mathrm{Tr}(\Sigma_2^{-1}(x^\top x - 2x^\top \mu_2 + \mu_2^\top \mu_2)) \right] \right\} \\
&= \frac{1}{2} \left\{ \ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathbb{E}_{D_1} \left[\mathrm{Tr}(\Sigma_2^{-1}(x^\top x - 2x^\top \mu_2 + \mu_2^\top \mu_2)) \right] \right\} \\
&= \frac{1}{2} \left\{ \ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathrm{Tr}(\Sigma_2^{-1}(\Sigma_1 + \mu_1^\top \mu_1 - 2\mu_1^\top \mu_2 + \mu_2^\top \mu_2)) \right\} \\
&= \frac{1}{2} \left\{ \ln \frac{|\Sigma_2|}{|\Sigma_1|} - n + \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^\top \Sigma_2^{-1}(\mu_1 - \mu_2) \right\}.
\end{aligned}
$$

C.2. Gaussian and t-Distribution

Let $D_1 = \mathcal{N}(\mu_1, \Sigma_1)$ and $D_2 = \mathcal{T}_\nu(\mu_2, \Sigma_2)$, where $\dim(\mu_i) = k$. Then

$$
\begin{aligned}
\mathrm{KL}(D_1 \| D_2) &= \ln \left(\frac{\Gamma(\frac{\nu}{2})\nu^{k/2}}{\Gamma(\frac{\nu+k}{2})} \right) + \frac{k}{2} \ln \pi + \frac{1}{2} \ln |\Sigma_2| - \frac{k}{2} \ln 2\pi e - \frac{1}{2} \ln |\Sigma_1| \\
&\quad + \frac{\nu + k}{2} \mathbb{E}_{D_1} \left[\ln \left(1 + \frac{1}{\nu}(x - \mu_2)^\top \Sigma_2^{-1}(x - \mu_2) \right) \right] \\
&= \ln \left(\frac{\Gamma(\frac{\nu}{2})\nu^{k/2}}{\Gamma(\frac{\nu+k}{2})} \right) + \frac{1}{2} \ln \frac{|\Sigma_2|}{|\Sigma_1|} - \frac{k}{2} \ln 2e \\
&\quad + \frac{\nu + k}{2} \mathbb{E}_{D_1} \left[\ln \left(1 + \frac{1}{\nu}(x - \mu_2)^\top \Sigma_2^{-1}(x - \mu_2) \right) \right].
\end{aligned}
$$

For the first term, if k is even, then

$$
\frac{\Gamma(\frac{\nu}{2})\nu^{k/2}}{\Gamma(\frac{\nu+k}{2})} = \frac{\nu^{k/2}}{(\frac{\nu+k}{2})^{\underline{k/2}}},
$$

where $y^{\underline{n}} = y(y - 1) \dots (y - n + 1)$ is the descending factorial. Now assume k is odd. By Gautschi's inequality, $\frac{\Gamma(a)}{\Gamma(a+1/2)} \leq \left(\frac{2a+1}{2a^2} \right)^{1/2}$. Choosing $a = \nu/2$ yields

$$
\frac{\Gamma(\frac{\nu}{2})\nu^{k/2}}{\Gamma(\frac{\nu+k}{2})} = \frac{\Gamma(\frac{\nu}{2})\nu^{1/2}\nu^{(k-1)/2}}{\Gamma(\frac{\nu+1}{2})(\frac{\nu+k}{2})^{\underline{(k-1)/2}}} \leq \frac{(\nu + 1)^{1/2}\nu^{(k-1)/2}}{(\frac{\nu}{2})^{1/2}(\frac{\nu+k}{2})^{\underline{(k-1)/2}}}.
$$

Now, bounding the expectation gives

$$
\begin{aligned}
&\mathbb{E}_{D_1} \left[\ln \left(1 + \frac{1}{\nu}(x - \mu_2)^\top \Sigma_2^{-1}(x - \mu_2) \right) \right] \\
&\leq \ln \left(1 + \frac{1}{\nu} \mathbb{E}_{D_1} \left[(x - \mu_2)^\top \Sigma_2^{-1}(x - \mu_2) \right] \right) \\
&= \ln \left(1 + \frac{1}{\nu} \mathrm{Tr}(\Sigma_2^{-1}\Sigma_1) + \frac{1}{\nu}(\mu_1 - \mu_2)^\top \Sigma_2^{-1}(\mu_1 - \mu_2) \right)
\end{aligned}
$$

$$\leq \ln\left(1 + \frac{1}{\nu}(\mu_1 - \mu_2)^\top \Sigma_2^{-1}(\mu_1 - \mu_2)\right) + \frac{\mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)}{\nu + (\mu_1 - \mu_2)^\top \Sigma_2^{-1}(\mu_1 - \mu_2)}$$

$$\leq \ln\left(1 + \frac{1}{\nu}(\mu_1 - \mu_2)^\top \Sigma_2^{-1}(\mu_1 - \mu_2)\right) + \frac{1}{\nu}\mathrm{Tr}(\Sigma_2^{-1}\Sigma_1),$$

where the second inequality follows from the fact that $\ln(a + b) \leq \ln(a) + b/a$. Combining everything yields

$$\mathrm{KL}(D_1\|D_2) \leq \ln\Lambda_{\nu,k} + \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{k}{2}\ln 2e + \frac{\nu + k}{2\nu}\mathrm{Tr}(\Sigma_2^{-1}\Sigma_1)$$

$$+ \frac{\nu + k}{2}\ln\left(1 + \frac{1}{\nu}(\mu_1 - \mu_2)^\top \Sigma_2^{-1}(\mu_1 - \mu_2)\right),$$

where

$$\Lambda_{\nu,k} = \begin{cases} \frac{\nu^{k/2}}{(\frac{\nu+k}{2})^{k/2}} & \text{if } k \text{ is even} \\ \frac{(\nu+1)^{1/2}\nu^{(k-1)/2}}{(\frac{\nu}{2})^{1/2}(\frac{\nu+k}{2})^{(k-1)/2}} & \text{if } k \text{ is odd.} \end{cases}$$

C.3. Gaussian and Laplace

Let $D_1 = \mathcal{N}(\mu, \sigma^2)$ and $D_2 = \mathsf{Lap}(\beta)$. Then

$$\mathrm{KL}(D_1\|D_2) = \ln(2\beta) + \frac{1}{\beta}\mathbb{E}_{D_1}[|x|] - \frac{1}{2}\ln(2\pi e\sigma^2)$$

$$= \ln(2\beta) + \frac{1}{2\beta}\left[\mu\mathrm{Erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right) + \frac{2\sqrt{2}\sigma}{\sqrt{\pi}}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\right] - \frac{1}{2}\ln(2\pi e\sigma^2)$$

$$\leq \frac{1}{2}\ln\frac{2\beta^2}{\sigma^2} + \frac{1}{2\beta}\left[|\mu|\sqrt{1 - \exp\left\{-\frac{2\mu^2}{\pi\sigma^2}\right\}} + \frac{2\sqrt{2}\sigma}{\sqrt{\pi}}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\right] - \frac{1}{2}\ln(\pi e).$$

D. Proof of Theorem 4.1

Choose $Q_{\boldsymbol{\theta}^*,\phi} = \mathcal{N}(\boldsymbol{\theta}^*, \phi^2 I)$. With $P_0 = \mathcal{T}_\nu(\mathbf{0}, \sigma^2 I)$, we have (Appendix C.2)

$$\mathrm{KL}(Q_{\boldsymbol{\theta}^*,\phi}\|P_0) \leq \ln\Lambda_{\nu,n} + \frac{n}{2}\ln\frac{\sigma^2}{\phi^2} - \frac{n}{2}\ln 2e + \frac{n(\nu + n)}{2\nu}\frac{\phi^2}{\sigma^2} + \frac{\nu + n}{2}\ln\left(1 + \frac{1}{\nu\sigma^2}\|\boldsymbol{\theta}^*\|^2\right),$$

where

$$\Lambda_{\nu,n} = \begin{cases} \frac{\nu^{n/2}}{(\frac{\nu+n}{2})^{n/2}} & \text{if } n \text{ is even} \\ \frac{(\nu+1)^{1/2}\nu^{(n-1)/2}}{(\frac{\nu}{2})^{1/2}(\frac{\nu+n}{2})^{(n-1)/2}} & \text{if } n \text{ is odd.} \end{cases}$$

Note that if n is even then $\frac{\Lambda_{\nu,n}}{2^{n/2}} \leq 1$ and if n is odd then $\frac{\Lambda_{\nu,n}}{2^{n/2}} \leq \frac{\nu+1}{\nu}$. Since $\mathrm{Var}_{Q_{\boldsymbol{\theta}^*,\phi}}[\theta_i] = \phi^2$, we have

$$L_{Bayes}(Z) \leq \inf_{\boldsymbol{\theta}^*} L_{\boldsymbol{\theta}^*}(Z) + \frac{Tc\phi^2}{2} + \frac{n}{2}\ln\frac{\nu+1}{\nu} + \frac{n}{2}\ln\frac{\sigma^2}{\phi^2} - \frac{n}{2} + \frac{n(\nu+n)}{2\nu}\frac{\phi^2}{\sigma^2} + \frac{\nu+n}{2}\ln\left(1 + \frac{1}{\nu\sigma^2}\|\boldsymbol{\theta}^*\|^2\right)$$

Choosing $\phi^2 = \frac{\nu\sigma^2 n}{Tc\nu\sigma^2 + (\nu+n)n}$ yields the theorem.

E. More on Hierarchical Priors for Sharing Statistical Strength

E.1. Multiple Simultaneous Observations

The Bayesian learner receives K input-output pairs $\{(\boldsymbol{x}_t^{(k)}, y_t^{(k)})\}_{k=1}^K$ at each time step. Each output is predicted using a separate weight vector $\boldsymbol{\theta}^{(k)}$, so the k-th likelihood is $p(y \mid \boldsymbol{\theta}^{(k)} \cdot \boldsymbol{x})$, $k = 1, \ldots, K$. Write $Z^{(k)} \triangleq \{(\boldsymbol{x}_t^{(k)}, y_t^{(k)})\}_{t=1}^T$.

Instead of using independent Gaussian priors on $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(K)}$, place a prior over the means of the K priors. For each dimension $j = 1, \ldots, n$, let

$$\mu_j \mid \sigma_0^2 \sim \mathcal{N}(0, \sigma_0^2) \tag{E.1}$$

and

$$\theta_j^{(k)} \mid \mu_j, \sigma^2 \sim \mathcal{N}(\mu_j, \sigma^2), \quad k = 1, \ldots, K, \tag{E.2}$$

and write $\boldsymbol{\theta}_j^{(1:K)} \triangleq (\theta_j^{(1)}, \ldots, \theta_j^{(K)})$. Integrating out μ_j yields

$$\boldsymbol{\theta}_j^{(1:K)} \mid \sigma_0^2, \sigma^2 \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{E.3}$$

where, with 1_K denoting the $K \times K$ all-ones matrix,

$$\Sigma \triangleq s^2 \rho 1_K + s^2 (1 - \rho) I \qquad s^2 \triangleq \sigma_0^2 + \sigma^2 \qquad \rho \triangleq \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}, \tag{E.4}$$

The Bayesian learner uses this hierarchical prior to simultaneously predict $y_t^{(1)}, \ldots, y_t^{(K)}$. For the following theorem, we must replace (A2) with an appropriately modified assumption for the simultaneous prediction task:

$$\|\boldsymbol{x}_t^{(k)}\|_2 \leq 1 \quad \text{for all } t, k. \tag{A2'}$$

Theorem E.1 (Hierarchical Gaussian regret, simultaneous observations). *If $\boldsymbol{\theta}_j^{(1:K)} \sim \mathcal{N}(\mathbf{0}, \Sigma)$, $j = 1, \ldots, n$, and* (A2') *holds in lieu of* (A2)*, then $\mathcal{R}(Z, \boldsymbol{\theta}^*)$ is bounded by*

$$\begin{aligned}
R_{Bayes}^{HG-sim}(Z, \boldsymbol{\theta}^*) &\triangleq \frac{1}{2\gamma^2} \sum_{k=1}^K \|\boldsymbol{\theta}^{*(k)}\|^2 + \frac{\sigma_0^2}{\sigma^2 \gamma^2} \sum_{k < \ell} \|\boldsymbol{\theta}^{*(k)} - \boldsymbol{\theta}^{*(\ell)}\|^2 \\
&\quad + \frac{n}{2} \ln\left(1 + \frac{K\sigma_0^2}{\sigma^2}\right) + \frac{nK}{2} \ln\left(1 - \frac{\sigma_0^2}{\gamma^2} + \frac{Tc\sigma^2}{n}\right),
\end{aligned} \tag{E.5}$$

where $\gamma^2 \triangleq K\sigma_0^2 + \sigma^2$.

It is instructive to compare the upper bound given in (E.5) to $\sum_k R_{Bayes}^G(Z_{(k)}, \boldsymbol{\theta}^{*(k)})$ with prior variance $s^2 = \sigma_0^2 + \sigma^2$. To do so, we find $\Delta(\boldsymbol{\theta}^*) \triangleq \sum_k R_{Bayes}^G(Z_{(k)}, \boldsymbol{\theta}^{*(k)}) - R_{Bayes}^{HG}(Z, \boldsymbol{\theta}^*)$:

$$\begin{aligned}
\Delta(\boldsymbol{\theta}^*) &= \frac{(K-1)\sigma_0^2}{2\gamma^2 s^2} \sum_{k=1}^K \|\boldsymbol{\theta}^{*(k)}\|^2 - \frac{\sigma_0^2}{\sigma^2 \gamma^2} \sum_{k<\ell} \|\boldsymbol{\theta}^{*(k)} - \boldsymbol{\theta}^{*(\ell)}\|^2 \\
&\quad - \frac{nK}{2} \ln\left(\frac{n\frac{s^2}{\sigma^2}(1 - \frac{\sigma_0^2}{\gamma^2}) + Tcs^2}{n + Tcs^2}\right) - \frac{n}{2} \ln\left(\left[1 + \frac{K\sigma_0^2}{\sigma^2}\right] \frac{\sigma^{2K}}{s^{2K}}\right)
\end{aligned}$$

For example, setting $\sigma_0 = \sigma$, so the correlation ρ is $1/2$, and $K = 2$, we find that if

$$4\|\boldsymbol{\theta}^{*(1)} - \boldsymbol{\theta}^{*(2)}\|^2 + 6s^2 n \ln\left(\frac{\frac{4}{3}n + Tcs^2}{n + Tcs^2}\right) \leq \|\boldsymbol{\theta}^{*(1)}\|^2 + \|\boldsymbol{\theta}^{*(2)}\|^2 + 0.863 s^2 n,$$

then the hierarchical model has a smaller regret bound than the non-hierarchical model.[5] As long as $Tcs^2 > 2n$, the condition becomes $4\|\boldsymbol{\theta}^{*(1)} - \boldsymbol{\theta}^{*(2)}\|^2 \leq \|\boldsymbol{\theta}^{*(1)}\|^2 + \|\boldsymbol{\theta}^{*(2)}\|^2 + Cs^2 n$ for some $0 < C < 0.863$. In this case there are two important observations about the benefits of the hierarchical model. First, noting that the expected magnitude of $\|\boldsymbol{\theta}^{*(1)}\|^2$ and $\|\boldsymbol{\theta}^{*(2)}\|^2$ is $\sigma^2 n$, as long as $\|\boldsymbol{\theta}^{*(1)}\|^2$ and $\|\boldsymbol{\theta}^{*(2)}\|^2$ are only a constant fraction $C/4$ of their expected magnitudes, the hierarchical model will always have smaller regret bound. Second, even if the previous condition does not hold, the difference in $\|\boldsymbol{\theta}^{*(1)} - \boldsymbol{\theta}^{*(2)}\|$ must be significantly larger than the expected magnitudes of $\|\boldsymbol{\theta}^{*(1)}\|^2$ and $\|\boldsymbol{\theta}^{*(2)}\|^2$ for the hierarchical model to have a larger regret bound than the non-hierarchical model. Thus, the use of the hierarchical model has potentially significantly reduced regret compared to the non-hierarchical model.

---

[5] For clarity, we have replaced $3\ln(4/3)$ with the bound 0.863.

E.2. Two-level Prior

In this section we derive bounds for the two-level prior in the case of sequential observations. Recall that the prior is

$$\boldsymbol{\beta} \sim \mathcal{N}(0, \sigma_0^2 I) \tag{E.6}$$

$$\boldsymbol{\mu}^{(s)} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma_1^2 I) \qquad\qquad s = 1, \ldots, S \tag{E.7}$$

$$\boldsymbol{\theta}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}^{(s_k)}, \sigma_2^2 I) \qquad\qquad k = 1, \ldots, K. \tag{E.8}$$

Integrating out $\boldsymbol{\beta}$, we immediately obtain:

$$\boldsymbol{\mu}_i^{(1:S)} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mu), \tag{E.9}$$

where $\Sigma_\mu \triangleq \sigma_0^2 1_S + \sigma_1^2 I$. Writing $\boldsymbol{\mu}_i = \boldsymbol{\mu}_i^{(1:S)}$ and $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^{(1:K)}$, we have

$$\begin{pmatrix} \boldsymbol{\mu}_i \\ \boldsymbol{\theta}_i \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \Sigma\right), \qquad\qquad \Sigma \triangleq \begin{pmatrix} \Sigma_\mu & \Sigma_{\mu\theta} \\ \Sigma_{\mu\theta}^\top & \Sigma_\theta \end{pmatrix}. \tag{E.10}$$

Hence,

$$\boldsymbol{\theta}_i \,|\, \boldsymbol{\mu}_i \sim \mathcal{N}(\Sigma_{\mu\theta}^\top \Sigma_\mu^{-1} \boldsymbol{\mu}_i, \Sigma_\theta - \Sigma_{\mu\theta}^\top \Sigma_\mu^{-1} \Sigma_{\mu\theta}). \tag{E.11}$$

Define the matrix P such that $P_{ks} = \mathbb{1}\{s = s_k\}$. We therefore have $\Sigma_{\mu\theta}^\top \Sigma_\mu^{-1} \boldsymbol{\mu}_i = P\boldsymbol{\mu}_i$ and hence $\Sigma_{\mu\theta}^\top = P\Sigma_\mu$, and furthermore $\Sigma_\theta - \Sigma_{\mu\theta}^\top \Sigma_\mu^{-1} \Sigma_{\mu\theta} = \sigma_2^2 I$ and hence $\Sigma_\theta = \sigma_2^2 I + P\Sigma_\mu P^\top$.

Hence, the prior on $\boldsymbol{\theta}_i$ is $P_0 = \mathcal{N}(\mathbf{0}, \Sigma_\theta)$. Choose $Q_{\boldsymbol{\theta}_i^*, \boldsymbol{\phi}} = \mathcal{N}(\boldsymbol{\theta}_i^*, \operatorname{diag} \boldsymbol{\phi})$, yielding

$$\mathrm{KL}(Q_{\boldsymbol{\theta}_i^*, \boldsymbol{\phi}} || P_0) = \frac{1}{2}\left\{ \ln \frac{|\Sigma_\theta|}{\prod_k \phi_k^2} - k - \mathrm{Tr}(\Sigma_\theta^{-1}) \sum_k \phi_k^2 + (\boldsymbol{\theta}_i^*)^\top \Sigma_\theta^{-1} \boldsymbol{\theta}_i^* \right\}. \tag{E.12}$$

Straightforward calculations show that the regret is bounded by

$$\sum_{i=1}^n (\boldsymbol{\theta}_i^*)^\top \Sigma_\theta^{-1} \boldsymbol{\theta}_i^* + \sum_{k=1}^K \frac{n}{2} \ln\left(2 \mathrm{Tr}(\Sigma_\theta^{-1}) + \frac{cT^{(k)}}{n} \right) + \frac{n}{2} \ln |\Sigma_\theta|. \tag{E.13}$$

E.3. Proof of Theorem E.1

First take $n = 1$, which will later generalize to arbitrary n. Choose $Q_{\boldsymbol{\theta}^{*(1:K)}, \phi} = \mathcal{N}(\boldsymbol{\theta}^{*(1:K)}, \phi^2 I)$ and note that

$$|\Sigma| = \sigma^{2K-2}(K\sigma_0^2 + \sigma^2) = \sigma^{2K-2}\gamma^2 \quad \text{and} \quad \Sigma^{-1} = -\frac{\sigma_0^2}{\sigma^2\gamma^2} 1_K + \frac{1}{\sigma^2} I.$$

Thus (Appendix C.1)

$$\begin{aligned}
\mathrm{KL}(Q_{\boldsymbol{\theta}^{*(1:K)}, \phi} || P_0) &= \frac{1}{2}\left\{ \ln \frac{|\Sigma|}{|\phi^2 I|} - K + \phi^2 \mathrm{Tr}(\Sigma^{-1}) + (\boldsymbol{\theta}^{*(1:K)})^\top \Sigma^{-1} \boldsymbol{\theta}^{*(1:K)} \right\} \\
&= \frac{K}{2} \ln \frac{\sigma^2 \gamma^{2/K}}{\phi^2 \sigma^{2/K}} - \frac{K}{2} + \frac{K(\gamma^2 - \sigma_0^2)}{2\sigma^2\gamma^2} \phi^2 \\
&\quad + \frac{1}{2\gamma^2} \sum_{k=1}^K (\theta^{*(k)})^2 + \frac{\sigma_0^2}{\sigma^2\gamma^2} \sum_{k<\ell} (\theta^{*(k)} - \theta^{*(\ell)})^2.
\end{aligned}$$

Moving to the case of general n, since $\mathrm{Var}_{Q_{\boldsymbol{\theta}^*, \phi}}[\sum_k \theta_j^{(k)}] = K\phi^2$ for all $j = 1, \ldots, n$, applying Theorem 2.2 gives

$$\begin{aligned}
L_{Bayes}(Z) &\le \sum_{k=1}^K L_{\boldsymbol{\theta}^{*(k)}}(Z^{(k)}) + \frac{TKc\phi^2}{2} + \frac{nK}{2} \ln \frac{\sigma^2\gamma^{2/K}}{\phi^2\sigma^{2/K}} - \frac{nK}{2} \\
&\quad \frac{nK(\gamma^2 - \sigma_0^2)}{2\sigma^2\gamma^2} \phi^2 + \frac{1}{2\gamma^2} \sum_{k=1}^K \|\boldsymbol{\theta}^{*(k)}\|^2 + \frac{\sigma_0^2}{\sigma^2\gamma^2} \sum_{k<\ell} \|\boldsymbol{\theta}^{*(k)} - \boldsymbol{\theta}^{*(\ell)}\|^2.
\end{aligned}$$

Choosing $\phi^2 = \frac{n\sigma^2\gamma^2}{n(\gamma^2 - \sigma_0^2) + Tc\sigma^2\gamma^2}$ yields the theorem.

E.4. Proof of Theorem 4.2

The proof is similar to that for Theorem E.1. However, use separate variances for each source:

$$Q_{\boldsymbol{\theta}^{*(1:K)}, \boldsymbol{\phi}} = \prod_k Q_{\boldsymbol{\theta}^{*(k)}, \phi_k} = \prod_k \mathcal{N}(\theta^{*(k)}, \phi_k^2).$$

The error term from the Taylor expansion used in Theorem 2.2 is $\sum_k \frac{T^{(k)} c \phi_k^2}{2}$, so

$$L_{Bayes}(Z) \leq \sum_{k=1}^{K} L_{\boldsymbol{\theta}^{*(k)}}(Z^{(k)}) + \sum_k \frac{T^{(k)} c \phi_k^2}{2} + \frac{n}{2} \ln \frac{\sigma^{2K} \gamma^2}{\sigma^2 \prod_k \phi_k^2} - \frac{nK}{2}$$

$$\frac{n(\gamma^2 - \sigma_0^2)}{2\sigma^2 \gamma^2} \sum_k \phi_k^2 + \frac{1}{2\gamma^2} \sum_{k=1}^{K} \|\boldsymbol{\theta}^{*(k)}\|^2 + \frac{\sigma_0^2}{\sigma^2 \gamma^2} \sum_{k<\ell} \|\boldsymbol{\theta}^{*(k)} - \boldsymbol{\theta}^{*(\ell)}\|^2.$$

Choosing $\phi_k^2 = \frac{n\sigma^2 \gamma^2}{n(\gamma^2 - \sigma_0^2) + T^{(k)} c \sigma^2 \gamma^2}$ yields the theorem.

F. More on Feature Selection

F.1. The Bayesian Lasso

For Bayesian model average learner we have:

Theorem F.1 (GLM Bayesian lasso regret). *If* $\theta_i \sim \mathsf{Lap}(\theta_i, \beta)$, $i = 1, \ldots, n$, *then*

$$\mathcal{R}(Z, \boldsymbol{\theta}^*) \leq \frac{1}{2\beta} \sum_i \min \left\{ \sqrt{\frac{2}{\pi\phi^2}} (\theta_i^*)^2, |\theta_i^*| \right\}$$

$$+ \frac{n}{2} \ln \left(\frac{2T^2 c^2 \beta^4}{\left(\sqrt{2n^2 + Tcn\beta^2 \pi} - \sqrt{2n^2} \right)^2} \right). \tag{F.1}$$

In the regime of $Tc\beta^2 \ll n$, (F.1) becomes (approximately)

$$\mathcal{R}(Z, \boldsymbol{\theta}^*) \leq \frac{1}{2\beta} \sum_i \min \left\{ \sqrt{\frac{2}{\pi\phi^2}} (\theta_i^*)^2, |\theta_i^*| \right\} + Cn$$

for some constant C independent of β and c. Hence, even for sparse $\boldsymbol{\theta}^*$, the regret bound is $\Theta(n)$. The inequalities used to prove the regret bound are all quite tight, so we conjecture that, up to constant factors, there is a matching lower bound, as least in the Gaussian regression case.

F.2. Proof of Theorem F.1

Apply Theorem 2.2 with $Q_{\boldsymbol{\theta}^*, \phi} = \mathcal{N}(\boldsymbol{\theta}^*, \phi^2 I)$. Since $p_0(\boldsymbol{\theta}) = \prod_i \mathsf{Lap}(\theta_i, \beta)$, we have (see Appendix C.3)

$$\mathrm{KL}(Q_{\boldsymbol{\theta}^*, \phi} \| P_0) \leq \frac{n}{2} \ln \frac{2\beta^2}{\phi^2} - \frac{n}{2} \ln(\pi e) + \frac{1}{2\beta} \sum_i \left[|\theta_i^*| \sqrt{1 - \exp\left\{ -\frac{2(\theta_i^*)^2}{\pi\phi^2} \right\}} + \frac{2\sqrt{2}\phi}{\sqrt{\pi}} \exp\left\{ -\frac{(\theta_i^*)^2}{2\phi^2} \right\} \right]$$

$$\leq \frac{n}{2} \ln \frac{2\beta^2}{\phi^2} - \frac{n}{2} \ln(\pi e) + \frac{\sqrt{2} n\phi}{\sqrt{\pi}\beta} + \frac{1}{2\beta} \sum_i \min \left\{ \sqrt{\frac{2}{\pi\phi^2}} (\theta_i^*)^2, |\theta_i^*| \right\}.$$

Since $\mathrm{Var}_{Q_{\boldsymbol{\theta}^*, \phi}}[\theta_i] = \phi^2$,

$$L_{Bayes}(Z) \leq \inf_{\boldsymbol{\theta}^*} L_{\boldsymbol{\theta}^*}(Z) + \frac{Tc\phi^2}{2} - \frac{n}{2} \ln(\pi e) + \frac{\sqrt{2} n\phi}{\sqrt{\pi}\beta} + \frac{n}{2} \ln \frac{2\beta^2}{\phi^2} + \frac{1}{2\beta} \sum_i \min \left\{ \sqrt{\frac{2}{\pi\phi^2}} (\theta_i^*)^2, |\theta_i^*| \right\}.$$

Choosing $\phi^2 = \frac{\left(\sqrt{2n^2 + Tcn\beta^2 \pi} - \sqrt{2n^2} \right)^2}{T^2 c^2 \beta^2 \pi}$ gives the desired result.

F.3. Proof of Theorem 4.3

Fix some $\boldsymbol{\theta}^*$. If $\theta_i^* = 0$, then let $Q_{\theta_i^*,\phi^2} = \delta_0$, so $\mathrm{KL}(Q_{\theta_i^*,\phi^2}||P_0) = \ln\frac{1}{p}$. If $\theta_i^* = 0$, then let $Q_{\theta_i^*,\phi^2} = \mathcal{N}(\theta_i^*, \phi^2)$, so

$$\mathrm{KL}(Q_{\theta_i^*,\phi^2}||P_0) = \mathrm{KL}(Q_{\theta_i^*,\phi^2}||\mathcal{N}(0, \sigma^2)) + \ln\frac{1}{1-p}.$$

The rest of the proof of (14) then closely follows earlier ones. To obtain (15), we observe that if $p = q^{1/n}$, then

$$m\ln\frac{1}{1-p} = m\ln\frac{1}{1-q^{1/n}} \le m\ln\frac{n}{1-q}$$

and

$$(n-m)\ln\frac{1}{p} = \frac{n-m}{n}\ln\frac{1}{q} \le \ln\frac{1}{q}.$$