SUPPLEMENTARY MATERIAL FOR *RISK AND REGRET OF HIERARCHICAL BAYESIAN LEARNERS*

JONATHAN H. HUGGINS AND JOSHUA B. TENENBAUM

Appendix A. Regret Bounds for Non-GLM Likelihoods

Recall Proposition 2.1, restated here for convenience:

Proposition. *The Bayesian cumulative loss is bounded as*

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

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

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

Defining

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

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

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

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

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

Combining (A.1) and (A.3) yields the theorem. □

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

A.1. **Application to Multi-class Logistic Regression.** For multi-class logistic regression (MLR) $y \in \{1, \dots, K\}$ is one of K classes, the parameters are $\boldsymbol{\theta} = \{\boldsymbol{\theta}^{(k)}\}_{k=1}^K$, and the likelihood is

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

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

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

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

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

Applying Gershgorin's circle theorem, we find that

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

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

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*

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

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

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

Appendix B. Proof of Theorem 3.2

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

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

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

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

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.

Appendix 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

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

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, $\frac{\Gamma(a)}{\Gamma(a+1/2)} \le \left(\frac{2a+1}{2a^2}\right)^{1/2}$. Choosing $a = \nu/2$ yields

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

Now, bounding the expectation gives

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

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

$$
\begin{aligned}
\mathrm{KL}(D_1||D_2) &\le \ln\Lambda_{\nu,k} + \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{k}{2}\ln 2e + \frac{\nu+k}{2\nu}\operatorname{Tr}(\Sigma_2^{-1}\Sigma_1) \\
&\quad + \frac{\nu+k}{2}\ln\left(1 + \frac{1}{\nu}(\mu_1-\mu_2)^\top\Sigma_2^{-1}(\mu_1-\mu_2)\right),
\end{aligned}
$$

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

$$\begin{aligned} \mathrm{KL}(D_1||D_2) &= \ln(2\beta) + \frac{1}{\beta}\mathbb{E}_{D_1}[|x|] - \frac{1}{2}\ln(2\pi e\sigma^2) \\ &= \ln(2\beta) + \frac{1}{2\beta}\left[\mu \mathrm{Erf}\left(\frac{\mu}{\sqrt{2}\sigma}\right) + \frac{2\sqrt{2}\sigma}{\sqrt{\pi}}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\right] - \frac{1}{2}\ln(2\pi e\sigma^2) \\ &\le \frac{1}{2}\ln\frac{2\beta^2}{\sigma^2} + \frac{1}{2\beta}\left[|\mu|\sqrt{1-\exp\left\{-\frac{2\mu^2}{\pi\sigma^2}\right\}} + \frac{2\sqrt{2}\sigma}{\sqrt{\pi}}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\right] - \frac{1}{2}\ln(\pi e). \end{aligned}$$

Appendix 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)

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

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

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

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

Appendix 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 $p(y \,|\, \boldsymbol{\theta}^{(k)} \cdot \boldsymbol{x})$, $k = 1, \ldots, K$. 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

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

and

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

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

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

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:

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

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*

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

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 $\Delta(\boldsymbol{\theta}^*) \triangleq \sum_k R_{Bayes}^G(Z_{(k)}, \boldsymbol{\theta}^{*(k)}) - R_{Bayes}^{HG}(Z, \boldsymbol{\theta}^*)$:

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

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

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

then the hierarchical model has a smaller regret bound than the non-hierarchical model.[1] As long as $Tcs^2 > 2n$, the condition becomes $4\|\boldsymbol{\theta}^{*(1)} - \boldsymbol{\theta}^{*(2)}\|^2 \le \|\boldsymbol{\theta}^{*(1)}\|^2 + \|\boldsymbol{\theta}^{*(2)}\|^2 + Cs^2 n$ 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.

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

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

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

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

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

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

Hence,

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

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

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

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

Straightforward calculations show that the regret is bounded by

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

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)

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

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

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

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

$$\begin{aligned}L_{Bayes}(Z) \le \sum_{k=1}^{K} L_{\boldsymbol{\theta}^{*(k)}}(Z^{(k)}) + \sum_k \frac{T^{(k)}c\phi_k^2}{2} + \frac{n}{2}\ln\frac{\sigma^{2K}\gamma^2}{\sigma^2\prod_k\phi_k^2} - \frac{nK}{2} \\ \frac{n(\gamma^2-\sigma_0^2)}{2\sigma^2\gamma^2}\sum_k\phi_k^2 + \frac{1}{2\gamma^2}\sum_{k=1}^{K}\|\boldsymbol{\theta}^{*(k)}\|^2 + \frac{\sigma_0^2}{\sigma^2\gamma^2}\sum_{k<\ell}\|\boldsymbol{\theta}^{*(k)} - \boldsymbol{\theta}^{*(\ell)}\|^2.\end{aligned}$$

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.

Appendix 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*

$$\begin{aligned}\mathcal{R}(Z, \boldsymbol{\theta}^*) \le \frac{1}{2\beta}\sum_i \min\left\{\sqrt{\frac{2}{\pi\phi^2}}(\theta_i^*)^2, |\theta_i^*|\right\} \\ + \frac{n}{2}\ln\left(\frac{2T^2c^2\beta^4}{\left(\sqrt{2n^2+Tcn\beta^2\pi} - \sqrt{2n^2}\right)^2}\right).\end{aligned} \tag{F.1}$$

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

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

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)

$$\begin{aligned}
\mathrm{KL}(Q_{\theta^*,\phi}||P_0) &\le \frac{n}{2}\ln\frac{2\beta^2}{\phi^2} - \frac{n}{2}\ln(\pi e) + \frac{1}{2\beta}\sum_i \left[|\theta_i^*| \sqrt{1 - \exp\left\{-\frac{2(\theta_i^*)^2}{\pi\phi^2}\right\}} + \frac{2\sqrt{2}\phi}{\sqrt{\pi}} \exp\left\{-\frac{(\theta_i^*)^2}{2\phi^2}\right\}\right] \\
&\le \frac{n}{2}\ln\frac{2\beta^2}{\phi^2} - \frac{n}{2}\ln(\pi e) + \frac{\sqrt{2}n\phi}{\sqrt{\pi}\beta} + \frac{1}{2\beta}\sum_i \min\left\{ \sqrt{\frac{2}{\pi\phi^2}} (\theta_i^*)^2, |\theta_i^*| \right\}.
\end{aligned}$$

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

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

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

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

Computer Science and Artificial Intelligence Laboratory, MIT
URL: http://jhhuggins.org/
E-mail address: jhuggins@mit.edu

Brain and Cognitive Science Department, MIT
URL: http://web.mit.edu/cocosci/josh.html/
E-mail address: jbt@mit.edu