We next observe that expansion of \( f \) yields:

\[
E_x [f_y(z)] = f_y(z^*) + f_y'(z^*)^T (z - z^*) + \frac{1}{2} (z - z^*)^T f_y''(\zeta(z))(z - z^*),
\]

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

\[
\mathbb{E}_z [f_y(z)] = f_y(z^*) + f_y'(z^*)^T 0 + \frac{1}{2} \mathbb{E}_z [(z - z^*)^T f_y''(\zeta(z))(z - z^*)] 
\leq f_y(z^*) + \frac{c}{2} \mathbb{E}_z [(z - z^*)^T (z - z^*)].
\]

Defining

\[
\omega \triangleq (x, \ldots, x, 1, \ldots, 1),
\]

we next observe that

\[
(z - z^*)^T (z - z^*) = \omega^T (\theta - \theta^*) (\theta - \theta^*)^T \omega.
\]

Letting \( \Sigma = \text{Var}[\theta] \), we thus have

\[
\mathbb{E}_z [(z - z^*)^T (z - z^*)] = \omega^T \mathbb{E}_\theta [(\theta - \theta^*) (\theta - \theta^*)^T] \omega 
\leq \|\omega\|_2^2 \|\mathbb{E}_\theta [(\theta - \theta^*) (\theta - \theta^*)^T]\| 
= (n'\|x\|_2^2 + n''\|\Sigma\|) 
\leq (n' + n'')\|\Sigma\|
\]

since it is assumed that \( \|x\|_2 \leq 1 \). Noting that \( L_Q(Z_T) = \sum_i \mathbb{E}_Q [f_y(\xi x_i, \psi)] \) and \( L_{\theta^*}(Z_T) = \sum_i f_y(\xi^* x_i, \psi^*) \), we have

\[
L_Q(Z_T) \leq L_{\theta^*}(Z_T) + \frac{T c (n' + n'')\|\Sigma\|}{2}.
\]

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

**Proof of Theorem 2.4** 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 \( \theta = \{\theta^{(k)}\}_{k=1}^K \), and the likelihood is

\[
p(y | \theta, x) = \frac{\exp(\theta^{(y)} \cdot x)}{\sum_{k=1}^K \exp(\theta^{(k)} \cdot x)}.
\]

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

**Proposition A.1**. Assumption [A1] holds for the MLR likelihood with \( c = 1/2 \).
Applying Gershgorin’s circle theorem, we find that
\[ f_y(z) = -z_y + \ln \sum_{k=1}^{K} e^{z_i}, \]  
where \( z_i = \theta^{(k)} \cdot x_i \), and hence the Hessian of \( f_y(z) \) is independent of \( y \):
\[ f''_y(z) = \frac{1}{(\sum_{k=1}^{K} e^{z_i})^2} \begin{pmatrix} \sum_{i \neq 1} e^{z_1+z_i} & \cdots & \sum_{i \neq K} e^{z_K+z_i} \\ \vdots & \ddots & \vdots \\ \sum_{i \neq 1} e^{z_1+z_i} & \cdots & \sum_{i \neq K} e^{z_K+z_i} \end{pmatrix}. \] (A.6)

Applying Gershgorin’s circle theorem, we find that
\[ \|f''_y(z)\| \leq \frac{2e^{z_i} \sum_{i \neq 1} e^{z_i}}{(\sum_{k=1}^{K} e^{z_k})^2}, \] (A.7)
where with loss of generality we have applied the theorem to the first row of the Hessian. Defining \( a = e^{z_1} \geq 0 \) and \( b = \sum_{i \neq 1} e^{z_i} \geq 0 \), we have \( \|f''_y(z)\| \leq \frac{2ab}{(a+b)^2} \). Maximization over the positive orthant occurs at \( a = b > 0 \), so \( \|f''_y(z)\| \leq \frac{1}{2} \).

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

**Theorem A.2** (Hierarchical Gaussian regret, multi-class regression). If \( \theta^{(1:K)} \sim N(0, \Sigma) \), \( j = 1, \ldots, n \), then using the MLR likelihood guarantees that \( R(Z, \theta^*) \) is bounded by
\[ R^{\text{mlr-HG}}_{\text{Bayes}}(Z, \theta^*) \leq \frac{1}{2} \sum_{k=1}^{K} \|\theta^{(k)}\|^2 + \frac{\sigma^2}{\gamma^2} \sum_{k=1}^{K} \|\theta^{(k)} - \theta^*\|^2 + \frac{nK\sigma_0^2 + 2T\sigma^2}{2n} \], (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 \( p_T(\theta) = \frac{p(Y|X, \theta)p_0(\theta)}{p(Y|X)} \),
\[ \text{KL}(P_T||P_0) = \mathbb{E}_{P_T} \left[ \ln \frac{p_T(\theta)}{p_0(\theta)} \right] = \mathbb{E}_{P_T} \left[ \ln \frac{p(Y|X, \theta)}{p(Y|X)} \right] = L_{\text{Bayes}}(Z_T) - L_{P_T}(Z_T). \]
(B.1)

Combining (2) and (B.1) with Theorem 3.1 implies that with probability \( 1 - \delta \), for all \( \theta \),
\[ |\mathcal{L}(P_T) - \hat{\mathcal{L}}(P_T, Z_T)| \leq \sqrt{C} \sqrt{\frac{L_{\theta}(Z_T) - L_{P_T}(Z_T) + B(\theta) + C(T) + \ln \kappa'/\delta}{T}}. \]

Observing that \( L_{\theta^*}(Z_T) < L_{P_T}(Z_T) \), so \( L_{\theta^*}(Z_T) - L_{P_T}(Z_T) < 0 \), completes the proof.
C.1. Multivariate Gaussians. Let $D_i = N(\mu_i, \Sigma_i), i = 1, 2$, where $\dim(\mu_i) = n$. Then

$$\text{KL}(D_1 || D_2) = \frac{1}{2} \mathbb{E}_{D_1} \left[ \ln \frac{\left| \Sigma_2 \right|}{\left| \Sigma_1 \right|} - (x - \mu_1)\Sigma_1^{-1}(x - \mu_1) + (x - \mu_2)\Sigma_2^{-1}(x - \mu_2) \right]$$

$$= \frac{1}{2} \left\{ \ln \frac{\left| \Sigma_2 \right|}{\left| \Sigma_1 \right|} + \mathbb{E}_{D_1} \left[ -\text{Tr}(\Sigma_1^{-1}(x - \mu_1)\Sigma_1^{-1}(x - \mu_1)) + \text{Tr}(\Sigma_2^{-1}(x - \mu_2)\Sigma_2^{-1}(x - \mu_2)) \right] \right\}$$

$$= \frac{1}{2} \left\{ \ln \frac{\left| \Sigma_2 \right|}{\left| \Sigma_1 \right|} - \text{Tr}(\Sigma_1^{-1}\Sigma_1) + \mathbb{E}_{D_1} \left[ \text{Tr}(\Sigma_2^{-1}(x\Sigma_2\Sigma_2^{-1}x - 2x\Sigma_2\Sigma_2^{-1}x)\Sigma_2) \right] \right\}$$

$$= \frac{1}{2} \left\{ \ln \frac{\left| \Sigma_2 \right|}{\left| \Sigma_1 \right|} - n + \mathbb{E}_{D_1} \left[ \text{Tr}(\Sigma_2^{-1}(x\Sigma_2\Sigma_2^{-1}x - 2x\Sigma_2\Sigma_2^{-1}x)) \right] \right\}$$

$$= \frac{1}{2} \left\{ \ln \frac{\left| \Sigma_2 \right|}{\left| \Sigma_1 \right|} - n + \text{Tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)\Sigma_2^{-1}(\mu_1 - \mu_2) \right\}.$$

C.2. Gaussian and $t$-Distribution. Let $D_1 = N(\mu_1, \Sigma_1)$ and $D_2 = T_\nu(\mu_2, \Sigma_2)$, where $\dim(\mu_i) = k$. Then

$$\text{KL}(D_1 || D_2) = \ln \left( \frac{\Gamma\left(\frac{\nu}{2}\right)\nu^{k/2}}{\Gamma\left(\frac{\nu+k}{2}\right)} \right) + \frac{k}{2} \ln \pi + \frac{1}{2} \ln \left| \Sigma_2 \right| - \frac{k}{2} \ln 2\pi e - \frac{1}{2} \ln \left| \Sigma_1 \right|$$

$$\quad + \frac{\nu + k}{2} \mathbb{E}_{D_1} \left[ \ln \left( 1 + \frac{1}{\nu}(x - \mu_2)\Sigma_2^{-1}(x - \mu_2) \right) \right]$$

$$= \ln \left( \frac{\Gamma\left(\frac{\nu}{2}\right)\nu^{k/2}}{\Gamma\left(\frac{\nu+k}{2}\right)} \right) + \frac{1}{2} \ln \left| \Sigma_2 \right| - \frac{k}{4} \ln 2e$$

$$\quad + \frac{\nu + k}{2} \mathbb{E}_{D_1} \left[ \ln \left( 1 + \frac{1}{\nu}(x - \mu_2)\Sigma_2^{-1}(x - \mu_2) \right) \right].$$

For the first term, if $k$ is even, then

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

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

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

Now, bounding the expectation gives

$$\mathbb{E}_{D_1} \left[ \ln \left( 1 + \frac{1}{\nu}(x - \mu_2)\Sigma_2^{-1}(x - \mu_2) \right) \right]$$

$$\leq \ln \left( 1 + \frac{1}{\nu} \mathbb{E}_{D_1} \left[ (x - \mu_2)\Sigma_2^{-1}(x - \mu_2) \right] \right)$$

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

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

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

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

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

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

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

C.3. Gaussian and Laplace. Let \( D_1 = N(\mu, \sigma^2) \) and \( D_2 = \text{Lap}(\beta) \). Then

\[
\text{KL}(D_1 \| D_2) = \ln(2\beta) + \frac{1}{\beta} \mathbb{E}_{D_1} [\| x \|] - \frac{1}{2} \ln(2\pi \sigma^2)
= \ln(2\beta) + \frac{1}{\beta} \left[ \mu \text{Erf} \left( \frac{\mu}{\sqrt{2}\sigma} \right) + \frac{2\sqrt{2}\sigma}{\sqrt{\pi}} \exp \left\{ -\frac{\mu^2}{2\sigma^2} \right\} \right] - \frac{1}{2} \ln(2\pi \sigma^2)
\leq \frac{1}{2} \ln \frac{2\beta^2}{\sigma^2} + \frac{1}{\beta} \left[ \mu \left( 1 - \exp \left\{ -\frac{2\mu^2}{\pi\sigma^2} \right\} \right) + \frac{2\sqrt{2}\sigma}{\sqrt{\pi}} \exp \left\{ -\frac{\mu^2}{2\sigma^2} \right\} \right] - \frac{1}{2} \ln(\pi e).
\]

APPENDIX D. PROOF OF THEOREM 4.1

Choose \( Q_{\theta^*; \phi} = N(\theta^*, \phi I) \). With \( P_0 = T_\nu(0, \sigma^2 I) \), we have (Appendix C.2)

\[
\text{KL}(Q_{\theta^*; \phi} | P_0) \leq \ln \Lambda_{\nu, n} + n \frac{\nu^{n/2}}{\Gamma((\nu + 1)/2)^{(n-1)/2}} \sigma^2 - \frac{n}{2} \ln 2\nu + \frac{n(\nu + n)}{2\nu} \phi^2 + \frac{\nu + n}{2} \ln \left( 1 + \frac{1}{\nu \sigma^2} \| \theta^* \|^2 \right),
\]

where

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

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

\[
L_{\text{Bayes}}(Z) \leq \inf_{\theta^*} \text{L}_{\theta^*} + \frac{Tc\phi^2}{2} + \frac{n}{2} \ln \frac{\nu + 1}{\nu} + \frac{n}{2} \ln \frac{\sigma^2}{\phi^2} + \frac{n(\nu + n)}{2\nu} \phi^2 + \frac{\nu + n}{2} \ln \left( 1 + \frac{1}{\nu \sigma^2} \| \theta^* \|^2 \right)
\]

Choosing \( \phi^2 = \frac{\nu \sigma^2 n}{Tc\sigma^2 + (\nu + n)n} \) 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 \( \{ (x_i^{(k)}, y_i^{(k)}) \}_{k=1}^K \) at each time step. Each output is predicted using a separate weight vector \( \theta^{(k)} \), so the \( k \)-th likelihood is \( p(y | \theta^{(k)}, x) \), \( k = 1, \ldots, K \). Write \( Z^{(k)} \triangleq \{ (x_i^{(k)}, y_i^{(k)}) \}_{i=1}^T \). Instead of using independent Gaussian priors on \( \theta^{(1)}, \ldots, \theta^{(K)} \), place a prior over the means of the \( K \) priors. For each dimension \( j = 1, \ldots, n \), let

\[
\mu_j \mid \sigma_0^2 \sim N(0, \sigma_0^2) \quad (E.1)
\]

and

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

and write \( \theta_j^{(1:K)} \triangleq (\theta_j^{(1)}, \ldots, \theta_j^{(K)}) \). Integrating out \( \mu_j \) yields

\[
\theta_j^{(1:K)} \mid \sigma_0^2, \sigma^2 \sim N(0, \Sigma), \quad (E.3)
\]

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

\[
\Sigma \triangleq s^2 \rho 1_K + s^2 (1 - \rho) I \quad s^2 \triangleq \sigma_0^2 + \sigma^2 \quad \rho \triangleq \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}, \quad (E.4)
\]

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

\[
\| x_t^{(k)} \|_2 \leq 1 \quad \text{for all } t, k. \quad (A2')
\]


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

$$
R_{Bayes}^{HG-sm}(Z, \theta^*) \leq \frac{1}{2\gamma^2} \sum_{k=1}^K \|\theta^{(k)}\|^2 + \frac{\sigma_2^2}{\sigma^2} \sum_{k=\ell}^K \|\theta^{(k)} - \theta^{(\ell)}\|^2 + \frac{n}{2} \ln \left(1 + \frac{K\sigma_0^2}{\sigma^2} + \frac{nK}{\gamma^2} + \frac{Tcs^2}{n}\right),
$$

(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 $\sum_k R_{Bayes}^G(Z(k), \theta^{*}(k))$ with prior variance $s^2 = \sigma_0^2 + \sigma^2$. To do so, we find $\Delta(\theta^*) \triangleq \sum_k R_{Bayes}^G(Z(k), \theta^{(k)}) - R_{Bayes}^{HG}(Z, \theta^*)$:

$$
\Delta(\theta^*) = \frac{(K-1)\sigma_0^2}{2\gamma^2s^2} \sum_{k=1}^K \|\theta^{(k)}\|^2 - \frac{\sigma_2^2}{\sigma^2} \sum_{k=\ell}^K \|\theta^{(k)} - \theta^{(\ell)}\|^2
- \frac{nK}{2} \ln \left(\frac{n^2\sigma_0^2(1 - \sigma^2)}{n + Tcs^2}\right) - \frac{n}{2} \ln \left(1 + \frac{K\sigma_0^2}{\sigma^2} + \frac{4n + Tcs^2}{n}\right).
$$

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

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

then the hierarchical model has a smaller regret bound than the non-hierarchical model. As long as $Tcs^2 > 2n$, the condition becomes $4\|\theta^{(1)} - \theta^{(2)}\|^2 \leq \|\theta^{(1)}\|^2 + \|\theta^{(2)}\|^2 + Cs^2n$ 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 $\|\theta^{(1)}\|^2$ and $\|\theta^{(2)}\|^2$ is $s^2n$, as long as $\|\theta^{(1)}\|^2$ and $\|\theta^{(2)}\|^2$ 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 $\|\theta^{(1)} - \theta^{(2)}\|$ must be significantly larger than the expected magnitudes of $\|\theta^{(1)}\|^2$ and $\|\theta^{(2)}\|^2$ 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

$$
\beta \sim N(0, \sigma_0^2I),
$$

(E.6)

$$
\mu^{(s)} \sim N(\beta, \sigma_0^2I), \quad s = 1, \ldots, S
$$

(E.7)

$$
\theta^{(k)} \sim N(\mu^{(sk)}, \sigma_0^2I), \quad k = 1, \ldots, K.
$$

(E.8)

Integrating out $\beta$, we immediately obtain:

$$
\mu_i^{(1:S)} \sim N(0, \Sigma_\mu),
$$

(E.9)

where $\Sigma_\mu \triangleq \sigma_0^2 I_S + \sigma_1^2 I$. Writing $\mu_i = \mu_i^{(1:S)}$ and $\theta_i = \theta_i^{(1,K)}$, we have

$$
\left(\begin{array}{c}
\mu_i \\
\theta_i
\end{array}\right) \sim N(0, \Sigma), \quad \Sigma \triangleq \left(\begin{array}{cc}
\Sigma_\mu & \Sigma_{\mu\theta} \\
\Sigma_{\mu\theta}^\top & \Sigma_\theta
\end{array}\right).
$$

(E.10)

Hence,

$$
\theta_i \mid \mu_i \sim N(\Sigma_{\mu\theta}^\top \Sigma_{\mu\theta}^{-1} \mu_i, \Sigma_{\theta} - \Sigma_{\mu\theta}^\top \Sigma_{\mu\theta}^{-1} \Sigma_{\theta}).
$$

(E.11)

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

Hence, the prior on $\theta_i$ is $P_0 = N(0, \Sigma_\theta)$. Choose $Q_{\theta_i^*} = N(\theta_i^*, \text{diag} \phi)$, yielding

$$
\text{KL}(Q_{\theta_i^*} \mid P_0) = \frac{1}{2} \left\{ \ln \frac{\Sigma_\theta}{\phi_k^2} - k - \text{Tr}(\Sigma_\theta^{-1}) \sum_k \phi_k^2 + (\theta_i^*)^\top \Sigma_\theta^{-1} \theta_i^* \right\}.
$$

(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
\[
\sum_{i=1}^{n} (\theta_i^*)^\top \Sigma_{\theta}^{-1} \theta_i^* + \sum_{k=1}^{K} \frac{n}{2} \ln \left( 2 \text{Tr}(\Sigma_{\theta}^{-1}) + \frac{cT(k)}{n} \right) + \frac{n}{2} \ln |\Sigma_{\theta}|.
\] (E.13)

E.3. **Proof of Theorem [E.1]** First take \( n = 1 \), which will later generalize to arbitrary \( n \). Choose \( Q_{\theta^*(1:K),\phi} = N(\theta^*(1:K), \phi^2 I) \) and note that
\[
|\Sigma| = \sigma^{2K-2}(K\sigma_0^2 + \sigma^2) = \sigma^{2K-2}\gamma^2 \quad \text{and} \quad \Sigma^{-1} = -\frac{\sigma_0^2}{\sigma^{2\gamma^2}} I_K + \frac{1}{\sigma^2} I.
\]
Thus (Appendix [C.1])
\[
\text{KL}(Q_{\theta^*(1:K),\phi} \| P_0) = \frac{1}{2} \left\{ \ln \frac{|\Sigma|}{|\phi^2 I|} - K + \phi^2 \text{Tr}(\Sigma^{-1}) + (\theta^*(1:K))^\top \Sigma^{-1} \theta^*(1:K) \right\}
= \frac{K}{2} \ln \frac{\sigma^{2\gamma^2/2} I_K}{\phi^2\sigma^{2/2} I} - \frac{K}{2} + \frac{K(\gamma^2 - \sigma_0^2)}{2\sigma^{2\gamma^2}} \phi^2
+ \frac{1}{2\gamma^2} \sum_{k=1}^{K} (\theta^*(k))^2 + \frac{\sigma_0^2}{2\gamma^2} \sum_{k<\ell} (\theta^*(k) - \theta^*(\ell))^2.
\]
Moving to the case of general \( n \), since \( \text{Var}_{Q_{\theta^*,\phi}}[\sum_k \theta_j^k] = K\phi^2 \) for all \( j = 1, \ldots, n \), applying Theorem 2.2 gives
\[
L_{\text{Bayes}}(Z) \leq \frac{K}{2} \ln \frac{nK(\gamma^2 - \sigma_0^2)}{2\sigma^{2\gamma^2}} \phi^2 + \frac{1}{2\gamma^2} \sum_{k=1}^{K} (\theta^*(k))^2 + \frac{\sigma_0^2}{2\gamma^2} \sum_{k<\ell} (\theta^*(k) - \theta^*(\ell))^2.
\]
Choosing \( \phi^2 = \frac{\sigma^{2\gamma^2}}{n(\gamma^2 - \sigma_0^2) + T \gamma^2 n^2} \) yields the theorem.

E.4. **Proof of Theorem [E.2]** The proof is similar to that for Theorem [E.1] However, use separate variances for each source:
\[
Q_{\theta^*(1:K),\phi} = \prod_k Q_{\theta^*(1:K),\phi_k} = \prod_k N(\theta^*(k), \phi_k^2).
\]
The error term from the Taylor expansion used in Theorem 2.2 is \( \sum_k \frac{T(k)c\phi_k^2}{2} \), so
\[
L_{\text{Bayes}}(Z) \leq \sum_{k=1}^{K} L_{\theta^*(1:K)}(Z(k)) + \frac{T(k)c\phi_k^2}{2} + \frac{n}{2} \ln \frac{\sigma^{2K\gamma^2}}{2\sigma^{2\gamma^2} I_k} - \frac{nK}{2}
+ \frac{n(\gamma^2 - \sigma_0^2)}{2\sigma^{2\gamma^2}} \sum_k \phi_k^2 + \frac{1}{2\gamma^2} \sum_{k=1}^{K} (\theta^*(k))^2 + \frac{\sigma_0^2}{2\gamma^2} \sum_{k<\ell} (\theta^*(k) - \theta^*(\ell))^2.
\]
Choosing \( \phi_k^2 = \frac{n\sigma^{2\gamma^2}}{(\gamma^2 - \sigma_0^2) + T \gamma^2 n^2} \) 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 \( \theta_i \sim \text{Lap}(\theta_i, \beta) \), \( i = 1, \ldots, n \), then
\[
\mathcal{R}(Z, \theta^*) \leq \frac{1}{2\beta} \sum_i \min \left\{ \sqrt{\frac{2}{\sigma_\theta^2} (\theta_i^*)^2}, |\theta_i^*| \right\}
+ \frac{n}{2} \ln \left( \frac{2T^2 c^2 \beta^4}{\left( \sqrt{2n^2 + Tcn\beta^2} + \sqrt{2n^2} \right)^2} \right).
\] (F.1)
In the regime of $Tc\beta^2 \ll n$, (F.1) becomes (approximately)
\[
R(Z, \theta^*) \leq \frac{1}{2\beta} \sum_i \min \left\{ \sqrt{\frac{2}{\pi \phi^2}} (\theta_i^*)^2, |\theta_i^*| \right\} + Cn
\]
for some constant $C$ independent of $\beta$ and $\epsilon$. Hence, even for sparse $\theta^*$, 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, at least in the Gaussian regression case.

F.2. Proof of Theorem F.1. Apply Theorem 2.2 with $Q_{\theta^*, \phi} = N(\theta^*, \phi^2 I)$. Since $p_0(\theta) = \prod_i \text{Lap}(\theta_i, \beta)$, we have (see Appendix C.3)
\[
\text{KL}(Q_{\theta^*, \phi} || P_0) \leq \frac{n}{2} \ln \frac{2\beta}{\phi^2} - \frac{n}{2} \ln(\pi e) + \frac{1}{2\beta} \sum_i \left[ \theta_i^* \sqrt{1 - \exp \left\{ -\frac{2(\theta_i^*)^2}{\pi \phi^2} \right\}} + \frac{2\sqrt{2} \phi}{\sqrt{\pi}} \exp \left\{ -\frac{(\theta_i^*)^2}{2\phi^2} \right\} \right] \leq \frac{n}{2} \ln \frac{2\beta}{\phi^2} - \frac{n}{2} \ln(\pi e) + \frac{1}{2\beta} \sum_i \min \left\{ \sqrt{\frac{2}{\pi \phi^2}} (\theta_i^*)^2, |\theta_i^*| \right\}.
\]
Since $\text{Var}_{Q_{\theta^*, \phi}}[\theta_i] = \phi^2$,
\[
L_{\text{Bayes}}(Z) \leq \inf_{\theta^*} L_{\theta^*}(Z) + \frac{Tc\phi^2}{2} - \frac{n}{2} \ln(\pi e) + \frac{2\sqrt{2n} \phi}{\sqrt{\pi} \beta} + \frac{n}{2} \ln \frac{2\beta}{\phi^2} + \frac{1}{2\beta} \sum_i \min \left\{ \sqrt{\frac{2}{\pi \phi^2}} (\theta_i^*)^2, |\theta_i^*| \right\}.
\]
Choosing $\phi^2 = \frac{(\sqrt{2n^2 + Tc\beta^2 \phi^2} - \sqrt{2n^2})^2}{Tc^2 \beta^2 \pi}$ gives the desired result.

F.3. Proof of Theorem 4.3. Fix some $\theta^*$. If $\theta_i^* = 0$, then let $Q_{\theta_i^*, \phi^2} = \delta_0$, so $\text{KL}(Q_{\theta_i^*, \phi^2} || P_0) = \frac{1}{p}$. If $\theta_i^* = 0$, then let $Q_{\theta_i^*, \phi^2} = N(\theta_i^*, \phi^2)$, so
\[
\text{KL}(Q_{\theta^*, \phi^2} || P_0) = \text{KL}(Q_{\theta^*, \phi^2} || N(0, \sigma^2)) + \ln \frac{1}{1 - p}.
\]
The rest of the proof of (14) then closely follows earlier ones. To obtain (15), we observe that if $p = q^{1/n}$, then
\[
m \ln \frac{1}{1 - p} = m \ln \frac{1}{1 - q^{1/n}} \leq m \ln \frac{n}{1 - q}
\]
and
\[
(n - m) \ln \frac{1}{p} = \frac{n - m}{n} \ln \frac{1}{q} \leq \ln \frac{1}{q}.
\]