Supplementary Material: Additional Proofs

To prove Theorem 2 we will use the following lemma.

**Lemma 1.** Suppose that for some functions \( q \) and \( \phi \), the loss function is of the form:

\[
\ell(z, \theta) = q(\theta) \cdot \phi(z) .
\]

Furthermore, suppose there exist constants \( n_0 \) and \( \phi_0 \) such that, for any training set \( Z = \{z_i \mid 1 \leq i \leq n\} \), where \( Z \sim D^n \) and \( D \sim D_1 \),

\[
E_{D \sim D_1}[E_{z \sim D}[\phi(z) \mid Z]] = \frac{\phi_0 + \sum_{i=1}^{n} \phi(z_i)}{n + n_0} .
\]

Then, there exists a perfect, Bayes-optimal regularizer of the form:

\[
R^*(\theta) = \frac{1}{n} q(\theta) \cdot \phi(\theta) .
\]

**Proof.** Let \( \tilde{L}(\theta, Z) \equiv E_{D \sim D_1}[E_{z \sim D}[\ell(z, \theta) \mid Z]] \) be the conditional expected test loss. By linearity of expectation,

\[
\tilde{L}(\theta, Z) = q(\theta) \cdot E_{D \sim D_1}[E_{z \sim D}[\phi(z) \mid Z]] = q(\theta) \cdot \frac{\phi_0 + \sum_{i=1}^{n} \phi(z_i)}{n + n_0} .
\]

Meanwhile, average training loss is \( \bar{L}(\theta) = \frac{1}{n} q(\theta) \cdot \sum_{i=1}^{n} \phi(z_i) . \) Thus,

\[
(n + n_0) \tilde{L}(\theta, Z) - n \bar{L}(\theta) = n R^*(\theta) .
\]

Rearranging, \( \tilde{L}(\theta) + R^*(\theta) = \frac{n + n_0}{n} \bar{L}(\theta, Z) \), so \( R^* \) is perfect and Bayes-optimal.

**Proof of Theorem 2.** By assumption, \( P(z \mid \theta) \) is an exponential family distribution, meaning that for some functions \( h, g, \eta, \) and \( T \), we have

\[
P(z \mid \theta) = h(z) g(\theta) \exp(\eta(\theta) \cdot T(z)) .
\]

Setting \( q(\theta) = (-\log g(\theta)) \oplus -\eta(\theta) \) and \( \phi(z) = \langle 1 \rangle \oplus T(z) \), we have

\[
-\log(P(z \mid \theta)) = q(\theta) \cdot \phi(z) - \log h(z) .
\]

Because the \( -\log h(z) \) term does not depend on \( \theta \), minimizing \( -\log(P(z \mid \theta)) \) is equivalent to using the loss function \( \ell(z, \theta) = q(\theta) \cdot \phi(z) \).

The conjugate prior for an exponential family has the form

\[
P(\eta(\theta)) = \frac{1}{Z_0} g(\theta)^{n_0} \exp(\eta(\theta) \cdot \tau_0) .
\]

where \( \tau_0 \) and \( n_0 \) are hyperparameters. One of the distinguishing properties of exponential families is that when \( \theta^* \) is drawn from a conjugate prior, the posterior expectation of \( T(z) \) has a linear form (Diaconis & Ylvisaker, 1979):

\[
E_{\theta^* \sim P(\theta)}[E_{z \sim P(z \mid \theta^*)}[T(z) \mid Z]] = \frac{\tau_0 + \sum_{i=1}^{n} T(z_i)}{n_0 + n} .
\]

Thus if we set \( \phi_0 = \langle n_0 \rangle \oplus \tau_0 \),

\[
E_{D \sim D_1}[E_{z \sim D}[\phi(z) \mid Z]] = \frac{\phi_0 + \sum_{i=1}^{n} \phi(z_i)}{n + n_0} .
\]

Lemma 1 then shows that a perfect regularizer is:

\[
R^*_z(\theta) = \frac{1}{n} q(\theta) \cdot \phi(\theta)
\]

\[
= \frac{1}{n} (-n_0 \log(g(\theta)) - \tau_0 \cdot \eta(\theta))
\]

\[
= \frac{1}{n} (-\log P(\eta(\theta)) - \log(Z_0)) .
\]

Because \( R^*_z \) and \( R^* \) differ by a constant, \( R^* \) is also perfect. \( \Box \)

**References**