Supplementary Material: Additional Proofs

To prove Theorem 2 we will use the following lemma.

Lemma 1. *Suppose that for some functions q and ϕ, the loss function is of the form:*

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

Furthermore, suppose there exist constants n_0 and ϕ_0 such that, for any training set $Z = \{z_i \mid 1 \le i \le n\}$, where $Z \sim \mathcal{D}^n$ and $\mathcal{D} \sim \mathcal{D}_1$,

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{D}_1}[\mathbb{E}_{z\sim\mathcal{D}}[\phi(z)] \mid Z] = \frac{\phi_0 + \sum_{i=1}^n \phi(z_i)}{n+n_0}\,.$$

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

$$R^*(\theta) = \frac{1}{n}q(\theta)\cdot\phi_0\,.$$

Proof. Let $\bar{L}(\theta, Z) \equiv \mathbb{E}_{D\sim\mathcal{D}_1}[E_{z\sim D}[\ell(z,\theta)] \mid Z]$ be the conditional expected test loss. By linearity of expectation,

$$\begin{aligned}\bar{L}(\theta, Z) &= q(\theta)\cdot E_{D\sim\mathcal{D}_1}[E_{z\sim D}[\phi(z)] \mid Z] \\ &= q(\theta)\cdot\frac{\phi_0 + \sum_{i=1}^n \phi(z_i)}{n+n_0}\,.\end{aligned}$$

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

$$(n+n_0)\bar{L}(\theta,Z) - n\hat{L}(\theta) = nR^*(\theta)\,.$$

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

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

$$P(z|\theta) = h(z)g(\theta)\exp(\eta(\theta)\cdot T(z))\,.$$

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

$$-\log(P(z|\theta)) = q(\theta)\cdot\phi(z) - \log h(z)\,.$$

Because the $-\log h(z)$ term does not depend on θ, minimizing $-\log(P(z|\theta))$ 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 τ_0 and n_0 are hyperparameters. One of the distinguishing properties of exponential families is that when θ^* is drawn from a conjugate prior, the posterior expectation of $T(z)$ has a linear form (Diaconis & Ylvisaker, 1979):

$$\mathbb{E}_{\theta^*\sim P(\theta)}[\mathbb{E}_{z\sim P(z|\theta^*)}[T(z)] \mid Z] = \frac{\tau_0 + \sum_{i=1}^n T(z_i)}{n_0+n}\,.$$

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

$$\mathbb{E}_{\mathcal{D}\sim\mathcal{D}_1}[\mathbb{E}_{z\sim\mathcal{D}}[\phi(z)] \mid Z] = \frac{\phi_0 + \sum_{i=1}^n \phi(z_i)}{n+n_0}\,.$$

Lemma 1 then shows that a perfect regularizer is:

$$\begin{aligned}R_1^*(\theta) &= \frac{1}{n}q(\theta)\cdot\phi_0 \\ &= \frac{1}{n}\left(-n_0\log(g(\theta)) - \tau_0\cdot\eta(\theta)\right) \\ &= \frac{1}{n}\left(-\log P(\eta(\theta)) - \log(Z_0)\right)\,.\end{aligned}$$

Because R_1^* and R^* differ by a constant, R^* is also perfect. □

References

Diaconis, P. and Ylvisaker, D. Conjugate priors for exponential families. *The Annals of Statistics*, pp. 269–281, 1979.