PAC-Bayesian Collective Stability
Supplemental Material

Ben London
University of Maryland

Bert Huang
University of Maryland

Ben Taskar
University of Washington

Lise Getoor
University of California, Santa Cruz

A PROBABILITY THEORY

A.1 Coupling

The proof of Theorem 1 requires a theoretical construction known as *coupling*. For random variables $\zeta^{(1)}$ and $\zeta^{(2)}$, with respective distributions $\mathbb{P}$ and $\mathbb{Q}$ over a sample space Ω, a coupling is any joint distribution $\hat{\mathbb{P}}$ for $(\zeta^{(1)}, \zeta^{(2)})$ such that the marginals $\hat{\mathbb{P}}(\zeta^{(1)})$ and $\hat{\mathbb{P}}(\zeta^{(2)})$ are equal to $\mathbb{P}$ and $\mathbb{Q}$ respectively.

Using a construction due to Fiebig (1993), one can create a coupling of two sequences of random variables, such that the probability that any two corresponding variables are different is upper-bounded by the ϑ-mixing coefficients in Definition 7. The following is an adaptation of this result (due to Samson, 2000) for continuous domains.

Lemma 1. *Let $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on a sample space Ω, with strictly positive densities with respect to a reference measure on Ω. Let $\boldsymbol{\zeta}^{(1)} \triangleq (\zeta_j^{(1)})_{j=1}^N$ and $\boldsymbol{\zeta}^{(2)} \triangleq (\zeta_j^{(2)})_{j=1}^N$ be random variables with respective distributions $\mathbb{P}$ and $\mathbb{Q}$. Then there exists a coupling $\hat{\mathbb{P}}$ of $(\boldsymbol{\zeta}^{(1)}, \boldsymbol{\zeta}^{(2)})$, with marginal distributions $\hat{\mathbb{P}}(\boldsymbol{\zeta}^{(1)}) = \mathbb{P}(\boldsymbol{\zeta}^{(1)})$ and $\hat{\mathbb{P}}(\boldsymbol{\zeta}^{(2)}) = \mathbb{Q}(\boldsymbol{\zeta}^{(2)})$, such that, for any $j \in [N]$,*

$$\hat{\mathbb{P}}\left\{\zeta_j^{(1)} \neq \zeta_j^{(2)}\right\} \leq \left\|\mathbb{P}(\boldsymbol{\zeta}_{j:N}^{(1)}) - \mathbb{Q}(\boldsymbol{\zeta}_{j:N}^{(2)})\right\|_{\text{TV}},$$

where $\hat{\mathbb{P}}\left\{\zeta_j^{(1)} \neq \zeta_j^{(2)}\right\}$ is the marginal *probability that $\zeta_j^{(1)} \neq \zeta_j^{(2)}$, under $\hat{\mathbb{P}}$.*

Note that the requirement of strictly positive densities is not restrictive, since one can always construct a positive density from a simply nonnegative one. We defer to Samson (2000) for details.

A.2 Proof of Theorem 1

Recall that $\mathcal{B} \subseteq \mathcal{Z}^n$ is the subset of "bad" inputs. For every $i \in [n]$, there exists a set of "bad starts" for which the probability that $\mathbf{Z}$ is bad is higher than some threshold $\lambda \in [0, 1]$. More formally, for $\mathbf{z} \in \mathcal{Z}^i$, let

$$\nu_i^{\boldsymbol{\pi}}(\mathbf{z}) \triangleq \mathbb{P}\left\{\mathbf{Z} \in \mathcal{B} \,|\, \mathbf{Z}_{\pi_i(1:i)} = \mathbf{z}\right\}. \tag{6}$$

and let

$$\mathcal{C}_i \triangleq \left\{\mathbf{z} \in \mathcal{Z}^i : \nu_i^{\boldsymbol{\pi}}(\mathbf{z}) > \lambda\right\} \tag{7}$$

denote the set of bad starts. Let $\mathcal{B}_i \triangleq \mathcal{C}_i \times \mathcal{Z}^{n-i}$ denote the set of inputs that have bad starts, and note that $\mathbf{Z} \in \mathcal{B}_i$ if and only if $\mathbf{Z}_{\pi_i(1:i)} \in \mathcal{C}_i$. Using the chain rule, we have that

$$\begin{aligned}
\nu &\geq \mathbb{P}\{\mathbf{Z} \in \mathcal{B}\} \\
&\geq \mathbb{P}\{\{\mathbf{Z} \in \mathcal{B}\} \cap \{\mathbf{Z} \in \mathcal{B}_i\}\} \\
&= \mathbb{P}\{\mathbf{Z} \in \mathcal{B} \,|\, \mathbf{Z} \in \mathcal{B}_i\}\, \mathbb{P}\{\mathbf{Z} \in \mathcal{B}_i\} \\
&\geq \inf_{\mathbf{z} \in \mathcal{C}_i} \nu_i^{\boldsymbol{\pi}}(\mathbf{z})\, \mathbb{P}\{\mathbf{Z} \in \mathcal{B}_i\} \\
&\geq \lambda\, \mathbb{P}\{\mathbf{Z} \in \mathcal{B}_i\};
\end{aligned} \tag{8}$$

therefore, $\mathbb{P}\{\mathbf{Z} \in \mathcal{B}_i\} \leq \nu/\lambda$. We then define a *new* set of "bad" inputs, $\mathcal{B}_\lambda \triangleq \bigcup_{i=1}^n \mathcal{B}_i$. Note that $\mathcal{B}_n = \mathcal{B}$, so $\mathcal{B} \subseteq \mathcal{B}_\lambda$. Via the union bound and Equation 8, we obtain

$$\mathbb{P}\{\mathbf{Z} \in \mathcal{B}_\lambda\} \leq \sum_{i=1}^n \mathbb{P}\{\mathbf{Z} \in \mathcal{B}_i\} \leq \frac{n\nu}{\lambda}.$$

What remains is to upper-bound

$$\mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \,|\, \mathbf{Z} \notin \mathcal{B}_\lambda\right]. \tag{9}$$

To do so, we use McDiarmid's *method of bounded differences* (McDiarmid, 1989). We define a Doob martingale difference sequence

$$V_i^{\boldsymbol{\pi}} \triangleq \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{\pi_i(1:i)}] - \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{\pi_i(1:i-1)}],$$

where $V_1^{\boldsymbol{\pi}} \triangleq \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{\pi_i(1)}] - \mathbb{E}[\varphi(\mathbf{Z})]$. Observe that $\mathbb{E}[V_i^{\boldsymbol{\pi}}] = 0$ and

$$\sum_{i=1}^n V_i^{\boldsymbol{\pi}} = \varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})].$$

If $V_i^{\boldsymbol{\pi}}$ is bounded, then the moment-generating function, $\mathbb{E}\left[e^{\tau V_i^{\boldsymbol{\pi}}} \,|\, \mathbf{Z} \notin \mathcal{B}_\lambda\right]$, can be upper-bounded using Hoeffding's lemma (Hoeffding, 1963).

Lemma 2. *If ξ is a random variable, such that $\mathbb{E}[\xi] = 0$ and $a \le \xi \le b$ almost surely, then for any $\tau \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{\tau\xi}\right] \le \exp\left(\frac{\tau^2(b-a)^2}{8}\right).$$

Thus, if we can show that

$$\begin{aligned}
&\sup V_i^{\boldsymbol{\pi}} - \inf V_i^{\boldsymbol{\pi}} \\
&\quad= \sup_{\substack{\mathbf{z}\in\mathcal{Z}^{i-1} \\ z,z'\in\mathcal{Z}}} \begin{pmatrix} \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)] \\ -\mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')] \end{pmatrix} \\
&\quad\le c_i, \qquad (10)
\end{aligned}$$

then we have that

$$\begin{aligned}
&\mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z})-\mathbb{E}[\varphi(\mathbf{Z})])} \,|\, \mathbf{Z}\notin\mathcal{B}_\lambda\right] \\
&\quad= \mathbb{E}\left[e^{\tau\sum_{i=1}^n V_i^{\boldsymbol{\pi}}} \,|\, \mathbf{Z}\notin\mathcal{B}_\lambda\right] \\
&\quad= \mathbb{E}\left[e^{\tau\sum_{i=1}^{n-1} V_i^{\boldsymbol{\pi}}}\, \mathbb{E}\left[e^{V_n^{\boldsymbol{\pi}}} \,|\, \mathbf{Z}_{\pi_i(1:n-1)}\right] \,|\, \mathbf{Z}\notin\mathcal{B}_\lambda\right] \\
&\quad\le \mathbb{E}\left[e^{\tau\sum_{i=1}^{n-1} V_i^{\boldsymbol{\pi}}} \,|\, \mathbf{Z}\notin\mathcal{B}_\lambda\right] e^{\frac{\tau^2 c_n^2}{8}} \\
&\quad\le \mathbb{E}\left[e^{\tau\sum_{i=1}^{n-2} V_i^{\boldsymbol{\pi}}} \,|\, \mathbf{Z}\notin\mathcal{B}_\lambda\right] e^{\frac{\tau^2\left(c_n^2+c_{n-1}^2\right)}{8}} \\
&\quad\le \dots \\
&\quad\le \exp\left(\frac{\tau^2\sum_{i=1}^n c_i^2}{8}\right), \qquad (11)
\end{aligned}$$

via the law of total expectation and Lemma 2. When $Z_1, \dots, Z_n$ are mutually independent, this is straightforward; it becomes complicated when we relax the independence assumption.

To bound each $V_i^{\boldsymbol{\pi}}$, we use Lemma 1 to construct a coupling that bounds the c_i coefficient in Equation 10, using the mixing coefficients and the smoothness of φ. Fix any $\mathbf{z} \in \mathcal{Z}^{i-1}$ and $z, z' \in \mathcal{Z}$, and let $N \triangleq n - i$. (Recall that, by Equation 9, $(\mathbf{z}, z) \notin \mathcal{C}_i$ and $(\mathbf{z}, z') \notin \mathcal{C}_i$; this will be important later on.) Define random variables $\boldsymbol{\zeta}^{(1)} \triangleq (\zeta_j^{(1)})_{j=1}^N$ and $\boldsymbol{\zeta}^{(2)} \triangleq (\zeta_j^{(2)})_{j=1}^N$, with coupling distribution $\hat{\mathbb{P}}$ such that

$$\begin{aligned}
\hat{\mathbb{P}}(\boldsymbol{\zeta}^{(1)}) &\triangleq \mathbb{P}\left(\mathbf{Z}_{\pi_i(i+1:n)} \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right) \\
\text{and}\quad \hat{\mathbb{P}}(\boldsymbol{\zeta}^{(2)}) &\triangleq \mathbb{P}\left(\mathbf{Z}_{\pi_i(i+1:n)} \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right). \qquad (12)
\end{aligned}$$

In other words, the marginal distributions of $\boldsymbol{\zeta}^{(1)}$ and $\boldsymbol{\zeta}^{(2)}$ are equal to the conditional distributions of $\mathbf{Z}_{\pi_i(i+1:n)}$ given $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)$ and $\mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')$ respectively. Note that we have renumbered the coupled variables according to π_i, but this does not affect the distribution.

Denote by π_i^{-1} the inverse of π_i (i.e., $\pi_i^{-1}(\pi_i(1:n)) = [n]$), and let

$$\psi(\mathbf{z}) = \varphi(\mathbf{z}_{\pi_i^{-1}(1:n)}).$$

In other words, ψ inverts the permutation applied to its input, so as to ensure $\psi(\mathbf{z}_{\pi_i(1:n)}) = \varphi(\mathbf{z})$. For convenience, let

$$\Delta\psi \triangleq \psi(\mathbf{z}, z, \boldsymbol{\zeta}^{(1)}) - \psi(\mathbf{z}, z', \boldsymbol{\zeta}^{(2)})$$

denote the difference, and define events

$$\begin{aligned}
B^{(1)} &\triangleq \mathbb{1}\left\{(\mathbf{z}, z, \boldsymbol{\zeta}^{(1)}) \in \mathcal{B}\right\} \\
\text{and}\quad B^{(2)} &\triangleq \mathbb{1}\left\{(\mathbf{z}, z', \boldsymbol{\zeta}^{(2)}) \in \mathcal{B}\right\}.
\end{aligned}$$

Using these definitions, we have the following equivalence:

$$\begin{aligned}
&\mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)] - \mathbb{E}[\varphi(\mathbf{Z}) \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')] \\
&\quad= \hat{\mathbb{E}}\left[\psi(\mathbf{z}, z, \boldsymbol{\zeta}^{(1)}) - \psi(\mathbf{z}, z', \boldsymbol{\zeta}^{(2)})\right] \\
&\quad= \hat{\mathbb{E}}\left[\Delta\psi\right] \\
&\quad= \hat{\mathbb{P}}\left\{\neg B^{(1)} \cap \neg B^{(2)}\right\} \hat{\mathbb{E}}\left[\Delta\psi \,|\, \neg B^{(1)} \cap \neg B^{(2)}\right] \\
&\qquad+ \hat{\mathbb{P}}\left\{B^{(1)} \cup B^{(2)}\right\} \hat{\mathbb{E}}\left[\Delta\psi \,|\, B^{(1)} \cup B^{(2)}\right]. \qquad (13)
\end{aligned}$$

Thus, to bound $V_i^{\boldsymbol{\pi}}$, we must upper-bound the right-hand terms.

By Equation 9, $\mathbf{Z} \notin \mathcal{B}_\lambda$, which implies $\mathbf{Z}_{\pi_i(1:i)} \notin \mathcal{C}_i$; this means that the values we condition on in Equation 12, $(\mathbf{z}, z)$ and $(\mathbf{z}, z')$, are not in the set of bad starts, $\mathcal{C}_i$. It therefore follows, via the union bound and Equations 6 and 7, that

$$\begin{aligned}
\hat{\mathbb{P}}\left\{B^{(1)} \cup B^{(2)}\right\} &\le \hat{\mathbb{P}}\{B^{(1)}\} + \hat{\mathbb{P}}\{B^{(2)}\} \\
&= \nu_i^{\boldsymbol{\pi}}(\mathbf{z}, z) + \nu_i^{\boldsymbol{\pi}}(\mathbf{z}, z') \\
&\le \lambda + \lambda = 2\lambda.
\end{aligned}$$

Further, since φ is α-uniformly range-bounded, $\Delta\psi \le \alpha$. Combining these inequalities, we have that

$$\hat{\mathbb{P}}\left\{B^{(1)} \cup B^{(2)}\right\} \hat{\mathbb{E}}\left[\Delta\psi \,|\, B^{(1)} \cup B^{(2)}\right] \le 2\lambda\alpha. \qquad (14)$$

Now, conditioned on $\neg B^{(1)} \cap \neg B^{(2)}$, we have that $(\mathbf{z}, z, \boldsymbol{\zeta}^{(1)}) \notin \mathcal{B}$ and $(\mathbf{z}, z', \boldsymbol{\zeta}^{(2)}) \notin \mathcal{B}$; in other words, both assignments are "good." We can therefore use the difference-boundedness condition to show that

$$\begin{aligned}
&\hat{\mathbb{E}}\left[\Delta\psi \,|\, \neg B^{(1)} \cap \neg B^{(2)}\right] \\
&\quad\le \hat{\mathbb{E}}\left[\beta\, D_{\mathrm{H}}\left((z, \boldsymbol{\zeta}^{(1)}), (z', \boldsymbol{\zeta}^{(2)})\right) \,|\, \neg B^{(1)} \cap \neg B^{(2)}\right] \\
&\quad\le \beta\, \hat{\mathbb{E}}\left[1 + \sum_{j=1}^N \mathbb{1}\left\{\zeta_j^{(1)} \ne \zeta_j^{(2)}\right\} \,|\, \neg B^{(1)} \cap \neg B^{(2)}\right] \\
&\quad= \beta\left(1 + \sum_{j=1}^N \hat{\mathbb{P}}\left\{\zeta_j^{(1)} \ne \zeta_j^{(2)} \,|\, \neg B^{(1)} \cap \neg B^{(2)}\right\}\right).
\end{aligned}$$

Recall from Lemma 1 and Definition 7 that

$$\begin{aligned}
&1 + \sum_{j=1}^{N} \hat{\mathbb{P}}\left\{\zeta_j^{(1)} \neq \zeta_j^{(2)}\right\} \\
&\quad \leq 1 + \sum_{j=i+1}^{n} \left\| \begin{array}{c} \mathbb{P}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z)\right) \\ -\mathbb{P}\left(\mathbf{Z}_{\pi_i(j:n)} \,|\, \mathbf{Z}_{\pi_i(1:i)} = (\mathbf{z}, z')\right) \end{array} \right\|_{\text{TV}} \\
&\quad \leq \sum_{j=i}^{n} \theta_{i,j}^{\boldsymbol{\pi}}.
\end{aligned}$$

This holds uniformly for all $\mathbf{z} \in \mathcal{Z}^{i-1}$ and $z, z' \in \mathcal{Z}$. Thus,

$$\begin{aligned}
&\hat{\mathbb{P}}\left\{\neg B^{(1)} \cap \neg B^{(2)}\right\} \hat{\mathbb{E}}\left[\Delta\psi \,|\, \neg B^{(1)} \cap \neg B^{(2)}\right] \\
&\quad \leq \beta \left(\hat{\mathbb{P}}\left\{\neg B^{(1)} \cap \neg B^{(2)}\right\} + \sum_{j=1}^{N} \hat{\mathbb{P}}\left\{\zeta_j^{(1)} \neq \zeta_j^{(2)}\right\} \right) \\
&\quad \leq \beta \left(1 + \sum_{j=1}^{N} \hat{\mathbb{P}}\left\{\zeta_j^{(1)} \neq \zeta_j^{(2)}\right\} \right) \\
&\quad \leq \beta \sum_{j=i}^{n} \theta_{i,j}^{\boldsymbol{\pi}}, \qquad (15)
\end{aligned}$$

In the second inequality, we used the fact that $\hat{\mathbb{P}}\left\{\neg B^{(1)} \cap \neg B^{(2)}\right\} \leq 1$.

Substituting the upper-bounds from Equations 14 and 15 into Equation 13, we then have that

$$\begin{aligned}
&\sup V_i^{\boldsymbol{\pi}} - \inf V_i^{\boldsymbol{\pi}} \\
&\quad = \sup_{\substack{\mathbf{z} \in \mathcal{Z}^{i-1} \\ z, z' \in \mathcal{Z}}} \hat{\mathbb{E}}\left[\psi(\mathbf{z}, z, \boldsymbol{\zeta}^{(1)}) - \psi(\mathbf{z}, z', \boldsymbol{\zeta}^{(2)})\right] \\
&\quad \leq 2\lambda\alpha + \beta \sum_{j=i}^{n} \theta_{i,j}^{\boldsymbol{\pi}} \\
&\quad \leq (2\lambda\alpha + \beta) \sum_{j=i}^{n} \theta_{i,j}^{\boldsymbol{\pi}}. \qquad (16)
\end{aligned}$$

The last inequality will help simplify the expression. Then, since we have shown that each $V_i^{\boldsymbol{\pi}}$ is uniformly bounded, we can use Equation 16 to upper-bound c_i in Equation 11, and thus obtain

$$\begin{aligned}
&\mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \,|\, \mathbf{Z} \notin \mathcal{B}_\lambda\right] \\
&\quad \leq \exp\left(\frac{\tau^2 \sum_{i=1}^{n} \left((2\lambda\alpha + \beta) \sum_{j=i}^{n} \theta_{i,j}^{\boldsymbol{\pi}} \right)^2}{8} \right) \\
&\quad \leq \exp\left(\frac{\tau^2 (2\lambda\alpha + \beta)^2 \, n \max_i \left(\sum_{j=i}^{n} \theta_{i,j}^{\boldsymbol{\pi}} \right)^2}{8} \right) \\
&\quad = \exp\left(\frac{\tau^2 (2\lambda\alpha + \beta)^2 \, n \left\| \boldsymbol{\Theta}_n^{\boldsymbol{\pi}} \right\|_\infty^2}{8} \right),
\end{aligned}$$

which completes the proof.

A.3 Implications of Theorem 1

In this section, we discuss some consequences of Theorem 1. For the following, let $\mathbf{Z} \triangleq (Z_i)_{i=1}^n$ be random variables with joint distribution $\mathbb{P}$, and let $\varphi : \mathcal{Z}^n \to \mathbb{R}$ be a measurable function.

We first show that Theorem 1 trivially yields a moment-generating function inequality for uniformly difference-bounded functions.

Corollary 2. *If φ is β-uniformly difference-bounded, then, for any $\tau \in \mathbb{R}$ and $\boldsymbol{\pi} \in \Pi(n)$,*

$$\mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])}\right] \leq \exp\left(\frac{n\tau^2\beta^2 \left\| \boldsymbol{\Theta}_n^{\boldsymbol{\pi}} \right\|_\infty^2}{8} \right).$$

Proof Since φ is β-uniformly difference-bounded, it is also $(\mathbb{P}, 0, \beta)$ difference-bounded; meaning, the measure of the "bad" set is 0. We therefore take $\lambda = 0$ and apply Theorem 1. (We interpret $\nu/\lambda = 0/0$ as 0, so as to directly use Theorem 1 to upper-bound $\mathbb{P}\{\mathbf{Z} \in \mathcal{B}_\lambda\}$; though, one could trivially show that $\mathbb{P}\{\mathbf{Z} \in \mathcal{B}_\lambda\} = 0$.) Since $\mathcal{B}_\lambda = \emptyset$, there is no need to condition on $\mathbf{Z} \notin \mathcal{B}_\lambda$. ■

We can also use Theorem 1 to derive some novel concentration inequalities for functions of interdependent random variables. While not used in this paper, these results may be of use in other contexts.

Corollary 3. *If φ is β-uniformly difference-bounded, then, for any $\epsilon > 0$ and $\boldsymbol{\pi} \in \Pi(n)$,*

$$\mathbb{P}\left\{\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})] \geq \epsilon\right\} \leq \exp\left(\frac{-2\epsilon^2}{n\beta^2 \left\| \boldsymbol{\Theta}_n^{\boldsymbol{\pi}} \right\|_\infty^2} \right).$$

Proof First, note that, for any $\tau \in \mathbb{R}$,

$$\mathbb{P}\left\{\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})] \geq \epsilon\right\} = \mathbb{P}\left\{e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \geq e^{\tau\epsilon}\right\},$$

due to the monotonicity of the exponent. Using Markov's inequality and Corollary 2, we then have that

$$\begin{aligned}
&\mathbb{P}\left\{e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \geq e^{\tau\epsilon}\right\} \\
&\quad \leq \frac{1}{e^{\tau\epsilon}} \mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])}\right] \\
&\quad \leq \frac{1}{e^{\tau\epsilon}} \exp\left(\frac{n\tau^2\beta^2 \left\| \boldsymbol{\Theta}_n^{\boldsymbol{\pi}} \right\|_\infty^2}{8} \right).
\end{aligned}$$

Optimizing with respect to τ, we take $\tau \triangleq \frac{4\epsilon}{n\beta^2 \| \boldsymbol{\Theta}_n^{\boldsymbol{\pi}} \|_\infty^2}$ to complete the proof. ■

Corollary 4. *If φ is $(\mathbb{P}, \beta, \nu)$ difference-bounded, and α-uniformly range-bounded, then, for any $\epsilon > 0$ and $\boldsymbol{\pi} \in \Pi(n)$,*

$$\mathbb{P}\left\{\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})] \geq \epsilon\right\} \leq \exp\left(\frac{-\epsilon^2}{2n\beta^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}\right) + \frac{2n\nu\alpha}{\beta}.$$

Proof Define the event

$$E \triangleq \mathbb{1}\left\{e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \geq e^{\tau\epsilon}\right\},$$

and let $\mathcal{B}_\lambda$ be as defined in the proof of Theorem 1. Using the law of total probability, we have that

$$\begin{aligned}\mathbb{P}\{E\} &= \mathbb{P}\{E \cap \{\mathbf{Z} \notin \mathcal{B}_\lambda\}\} + \mathbb{P}\{E \cap \{\mathbf{Z} \in \mathcal{B}_\lambda\}\} \\ &\leq \mathbb{P}\{E \,|\, \mathbf{Z} \notin \mathcal{B}_\lambda\} + \mathbb{P}\{E \cap \{\mathbf{Z} \in \mathcal{B}_\lambda\}\} \\ &\leq \mathbb{P}\{E \,|\, \mathbf{Z} \notin \mathcal{B}_\lambda\} + \mathbb{P}\{\mathbf{Z} \in \mathcal{B}_\lambda\}.\end{aligned}$$

Via Markov's inequality and Theorem 1,

$$\begin{aligned}\mathbb{P}\{E \,|\, \mathbf{Z} \notin \mathcal{B}_\lambda\} &\leq \frac{1}{e^{\tau\epsilon}} \mathbb{E}\left[e^{\tau(\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})])} \,|\, \mathbf{Z} \notin \mathcal{B}_\lambda\right] \\ &\leq \frac{1}{e^{\tau\epsilon}} \exp\left(\frac{n\tau^2(2\lambda\alpha + \beta)^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8}\right),\end{aligned}$$

and

$$\mathbb{P}\{\mathbf{Z} \in \mathcal{B}_\lambda\} \leq \frac{n\nu}{\lambda}.$$

Combining these inequalities, we have that

$$\begin{aligned}&\mathbb{P}\left\{\varphi(\mathbf{Z}) - \mathbb{E}[\varphi(\mathbf{Z})] \geq \epsilon\right\} \\ &\quad= \mathbb{P}\{E\} \\ &\quad\leq \frac{1}{e^{\tau\epsilon}} \exp\left(\frac{n\tau^2(2\lambda\alpha + \beta)^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8}\right) + \frac{n\nu}{\lambda}.\end{aligned}$$

Taking $\lambda \triangleq \beta/(2\alpha)$ (which is approximately optimal) and $\tau \triangleq \frac{\epsilon}{n\beta^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}$ completes the proof. ∎

These tail bounds extend some current state-of-the-art results. In particular, Corollary 3 extends Kontorovich and Ramanan (2008, Theorem 1.1) by supporting filtrations of the mixing coefficients. Further, when $Z_1, \ldots, Z_n$ are mutually independent (i.e., $\left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty = 1$), we recover McDiarmid's inequality. Corollary 4 extends (Kutin, 2002, Theorem 3.6) to interdependent random variables.

A.4 Bounded $\left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty$ Conditions for Markov Random Fields

In this section, we describe some general settings under which the dependency matrix $\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}$ has bounded infinity norm. Fix a graph $G = (\mathcal{V}, \mathcal{E})$. For any node $i \in \mathcal{V}$, and subsets $A, B \subseteq \mathcal{V}$, define the distance function $\delta_i(A \,|\, B)$ as the length of the shortest path from i to any node in A, in the induced subgraph over $\mathcal{V} \setminus B$. Let $\Sigma_i(k)$ denote the set of all subset pairs $(A, B) : A, B \subseteq \mathcal{V}$ such that $\delta_i(A \,|\, B) \geq k$.

Definition 13. For an MRF $\mathbf{Z}$ on a graph $G = (\mathcal{V}, \mathcal{E})$, with distribution $\mathbb{P}$, define the *distance-based ϑ-mixing coefficients* as

$$\vartheta(k) \triangleq \sup_{\substack{i \in \mathcal{V} \\ (A,B) \in \Sigma_i(k) \\ z, z' \in \mathcal{Z} \\ \mathbf{z} \in \mathcal{Z}^{|B|}}} \left\| \begin{array}{c} \mathbb{P}\left(\mathbf{Z}_A \,|\, \mathbf{Z}_B = \mathbf{z}, Z_i = z\right) \\ -\mathbb{P}\left(\mathbf{Z}_A \,|\, \mathbf{Z}_B = \mathbf{z}, Z_i = z'\right) \end{array} \right\|_{\mathrm{TV}}.$$

The sequence $\vartheta(1), \vartheta(2), \ldots$ roughly measures how dependence decays with graph distance.

Proposition 1. *Let $\mathbf{Z}$ be an MRF on a graph G, with maximum degree Δ_G. For any positive constant $\epsilon > 0$, if $\mathbf{Z}$ has a distance-based ϑ-mixing sequence such that, for all $k \geq 1$, $\vartheta(k) < (\Delta_G + \epsilon)^{-k}$, then there exists a filtration $\boldsymbol{\pi}$ such that*

$$\left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty \leq 1 + \Delta_G/\epsilon.$$

Proof Since $\vartheta(k)$ uniformly upper-bounds the ϑ-mixing coefficients, each upper-triangular entry of $\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}$ is upper-bounded by

$$\theta_{i,j}^{\boldsymbol{\pi}} \leq \vartheta\left(\delta_{\pi_i(i)}\left(\pi_i(j:n) \,|\, \pi_i(1:i-1)\right)\right).$$

We construct the filtration $\boldsymbol{\pi}$ recursively, starting from any initial permutation π_1. Then, for $i = 2, \ldots, n$, we determine each successive permutation using a breadth-first search over the variables not conditioned on in the previous permutation. More precisely, we set $\pi_i(1:i-1) = \pi_{i-1}(1:i-1)$; then, we set $\pi_i(i:n)$ using the trace of a breadth-first search over the induced subgraph of nodes $\pi_{i-1}(i:n)$, starting at any node.

The degree of any node in this induced subgraph is at most the maximum degree of the full graph, Δ_G, so the number of nodes at distance k from node $\pi_i(i)$ is at most Δ_G^k. Therefore,

$$\sum_{j=i}^{n} \theta_{i,j}^{\boldsymbol{\pi}} \leq \sum_{k=0}^{\infty} \Delta_G^k \, \vartheta(k) \leq \sum_{k=0}^{\infty} \left(\frac{\Delta_G}{\Delta_G + \epsilon}\right)^k.$$

This geometric series converges to

$$\frac{1}{1 - \Delta_G/(\Delta_G + \epsilon)} = 1 + \Delta_G/\epsilon,$$

which completes the proof. ∎

Uniformly geometric distance-based ϑ-mixing may seem like a restrictive condition. However, our analysis is overly pessimistic, in that it ignores the structure

of the MRF beyond simply the maximum degree of the graph. Further, it does not take advantage of the actual conditional independencies present in the distribution. Nevertheless, there is a natural interpretation to the above preconditions that follows from considering the mixing coefficients at distance 1. For the immediate neighbors of a node—i.e., its Markov blanket—its ϑ-mixing coefficient must be less than $1/\Delta_G$. This loosely means that the combination of all incoming influence must be less than 1, implying that there is sufficiently strong influence from local features.

A.5 Gaussian Tail Bounds

The proof of Theorem 6 will require tail bounds for certain operations on Gaussian random vectors. To prove these, we begin with some basic properties of the normal distribution.

Lemma 3. *If X is a Gaussian random variable, with mean μ and variance σ^2, then, for any $\tau \in \mathbb{R}$,*

$$\mathbb{E}\left[e^{\tau Y}\right] = \exp\left(\frac{\tau^2\sigma^2}{2}\right); \tag{17}$$

and for any $\epsilon > 0$,

$$\Pr\left\{|X-\mu| \geq \epsilon\right\} \leq 2\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{18}$$

Equation 18 follows from Equation 17. Lemma 3 can now be used to derive the following tail bounds.

Lemma 4. *Let $\mathbf{X} \triangleq (X_i)_{i=1}^d$ be independent Gaussian random variables, with mean vector $\boldsymbol{\mu} \triangleq (\mu_1, \ldots, \mu_d)$ and variance σ^2. Then, for any $a \geq 1$ and $\epsilon > 0$,*

$$\Pr\left\{\|\mathbf{X}-\boldsymbol{\mu}\|_a \geq \epsilon\right\} \leq 2d\exp\left(-\frac{\epsilon^2}{2\sigma^2 d^{2/a}}\right).$$

Proof Observe that, if $\|\mathbf{X}-\boldsymbol{\mu}\|_a \geq \epsilon$, then there must exist at least one coordinate $i \in [d]$ such that $|X_i - \mu_i| \geq \epsilon/d^{1/a}$; otherwise, we would have

$$\begin{aligned}\|\mathbf{X}-\boldsymbol{\mu}\|_a &= \left(\sum_{i=1}^d |X_i-\mu_i|^a\right)^{1/a} \\ &< \left(d\left(\frac{\epsilon}{d^{1/a}}\right)^a\right)^{1/a} = \epsilon.\end{aligned}$$

Accordingly, we apply the union bound and obtain

$$\begin{aligned}\Pr\left\{\|\mathbf{X}-\boldsymbol{\mu}\|_a \geq \epsilon\right\} &\leq \Pr\left\{\exists i : |X_i-\mu_i| \geq \frac{\epsilon}{d^{1/a}}\right\} \\ &\leq \sum_{i=1}^d \Pr\left\{|X_i-\mu_i| \geq \frac{\epsilon}{d^{1/a}}\right\} \\ &\leq \sum_{i=1}^d 2\exp\left(-\frac{\epsilon^2}{2\sigma^2 d^{2/a}}\right).\end{aligned}$$

The last inequality follows from Equation 18. Summing over $i = 1, \ldots, d$ completes the proof. ■

Lemma 5. *Let $\mathbf{X} \triangleq (X_i)_{i=1}^d$ be independent Gaussian random variables, with mean vector $\boldsymbol{\mu} \triangleq (\mu_1, \ldots, \mu_d)$ and variance σ^2. Let $\mathbf{z} \in \mathbb{R}^d$ be a vector with $\|\mathbf{z}\|_2 \leq 1$. Then, for any $\epsilon > 0$,*

$$\Pr\left\{\langle \mathbf{X}-\boldsymbol{\mu}, \mathbf{z}\rangle \geq \epsilon\right\} \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

Proof Let $\mathbf{Y} \triangleq \mathbf{X} - \boldsymbol{\mu}$. For $i = 1, \ldots, d$, let $\tau_i \triangleq \tau|z_i|$. Since each Y_i is independent and zero-mean, by symmetry and Equation 17,

$$\begin{aligned}\mathbb{E}\left[e^{\tau\langle \mathbf{Y}, \mathbf{z}\rangle}\right] &= \mathbb{E}\left[\prod_{i=1}^d e^{\tau z_i Y_i}\right] \\ &= \prod_{i=1}^d \mathbb{E}\left[e^{\tau z_i Y_i}\right] \\ &= \prod_{i=1}^d \mathbb{E}\left[e^{\tau_i \operatorname{sgn}(z_i) Y_i}\right] \\ &= \prod_{i=1}^d \mathbb{E}\left[e^{\tau_i Y_i}\right] \\ &= \prod_{i=1}^d \exp\left(\frac{\tau_i^2\sigma^2}{2}\right) \\ &= \exp\left(\frac{\tau^2\sigma^2}{2}\sum_{i=1}^d |z_i|^2\right).\end{aligned}$$

Observe that $\sum_{i=1}^d |z_i|^2 = \|\mathbf{z}\|_2^2 \leq 1$, since $\|\mathbf{z}\|_2 \leq 1$. Therefore, using Markov's inequality, we have that

$$\begin{aligned}\Pr\left\{\langle \mathbf{X}-\boldsymbol{\mu}, \mathbf{z}\rangle \geq \epsilon\right\} &= \Pr\left\{e^{\tau\langle \mathbf{X}-\boldsymbol{\mu}, \mathbf{z}\rangle} \geq e^{\tau\epsilon}\right\} \\ &\leq \frac{1}{e^{\tau\epsilon}}\mathbb{E}\left[e^{\tau\langle \mathbf{Y}, \mathbf{z}\rangle}\right] \\ &\leq \exp\left(\frac{\tau^2\sigma^2\|\mathbf{z}\|_2^2}{2} - \tau\epsilon\right) \\ &\leq \exp\left(\frac{\tau^2\sigma^2}{2} - \tau\epsilon\right).\end{aligned}$$

Taking $\tau \triangleq \epsilon/\sigma^2$ completes the proof. ■

B PAC-BAYES PROOFS

B.1 Change of Measure

The following lemma, often called the *change of measure* inequality, is due to Donsker and Varadhan (1975).

Lemma 6. *For any measurable function $\varphi : \mathcal{H} \to \mathbb{R}$, and any two distributions $\mathbb{H}, \mathbb{Q}$ on $\mathcal{H}$,*

$$\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}[\varphi(h)] \leq D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[e^{\varphi(h)}\right].$$

A straightforward proof appears in McAllester (2003).

B.2 Stability of the Loss

The following technical lemmas are used in the proofs of Theorems 2 and 3.

Lemma 7. *Let ℓ be (M, Λ)-admissible.*

1. *If h has β-uniform collective stability, then $\ell \circ h$ has $(M + \Lambda\beta)$-uniform collective stability.*
2. *If h has $(\mathbb{P}, \nu, \beta)$ collective stability, then $\ell \circ h$ has $(\mathbb{P}, \nu, M + \Lambda\beta)$ collective stability.*

Proof For any assignments $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$, let $\mathcal{I} \triangleq \{i \in [n] : z_i \neq z'_i\}$ denote the set of coordinates at which their values differ. By definition,

$$\begin{aligned}
&\sum_{j=1}^{n} \left|\ell(y_j, h_j(\mathbf{x})) - \ell(y'_j, h_j(\mathbf{x}'))\right| \\
&\qquad = \sum_{i\in\mathcal{I}} \left|\ell(y_i, h_i(\mathbf{x})) - \ell(y'_i, h_i(\mathbf{x}'))\right| \\
&\qquad\quad + \sum_{j\notin\mathcal{I}} \left|\ell(y_j, h_j(\mathbf{x})) - \ell(y_j, h_j(\mathbf{x}'))\right|.
\end{aligned}$$

Focusing on the first sum, for any $i \in \mathcal{I}$, we have via the first admissibility condition that

$$\begin{aligned}
&\left|\ell(y_i, h_i(\mathbf{x})) - \ell(y'_i, h_i(\mathbf{x}'))\right| \\
&\qquad \leq \left|\ell(y_i, h_i(\mathbf{x})) - \ell(y_i, h_i(\mathbf{x}'))\right| \\
&\qquad\quad + \left|\ell(y_i, h_i(\mathbf{x}')) - \ell(y'_i, h_i(\mathbf{x}'))\right| \\
&\qquad \leq \left|\ell(y_i, h_i(\mathbf{x})) - \ell(y_i, h_i(\mathbf{x}'))\right| + M.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{i\in\mathcal{I}} \left|\ell(y_i, h_i(\mathbf{x})) - \ell(y'_i, h_i(\mathbf{x}'))\right| \\
&\qquad \leq M\left|\mathcal{I}\right| + \sum_{i\in\mathcal{I}} \left|\ell(y_i, h_i(\mathbf{x})) - \ell(y_i, h_i(\mathbf{x}'))\right|.
\end{aligned}$$

Combining this with the second sum, we have that

$$\begin{aligned}
&\sum_{j=1}^{n} \left|\ell(y_j, h_j(\mathbf{x})) - \ell(y'_j, h_j(\mathbf{x}'))\right| \\
&\qquad \leq M\left|\mathcal{I}\right| + \sum_{j=1}^{n} \left|\ell(y_j, h_j(\mathbf{x})) - \ell(y_j, h_j(\mathbf{x}'))\right| \\
&\qquad \leq M\left|\mathcal{I}\right| + \Lambda \sum_{j=1}^{n} \left\|h_j(\mathbf{x}) - h_j(\mathbf{x}')\right\|_1 \\
&\qquad = M\left|\mathcal{I}\right| + \Lambda \left\|h(\mathbf{x}) - h(\mathbf{x}')\right\|_1,
\end{aligned}$$

where we have used the second admissibility condition. Observe that $|\mathcal{I}| = D_{\mathrm{H}}(\mathbf{z}, \mathbf{z}')$. Upper-bounding $\|h(\mathbf{x}) - h(\mathbf{x}')\|_1$ by the uniform or probabilistic collective stability conditions completes the proof. ■

Note that Lemma 7 still holds when the cardinality of $\mathbf{X}$ does not equal that of $\mathbf{Y}$.

The following lemmas follow trivially from the definition of L (Equation 1).

Lemma 8. *For any loss ℓ and hypothesis h:*

1. *If $\ell \circ h$ has β-uniform collective stability, then $L(h, \cdot)$ is $(\beta/(mn))$-uniformly difference-bounded.*
2. *If $\ell \circ h$ has $(\mathbb{P}, \nu, \beta)$ collective stability, then $L(h, \cdot)$ is $(\mathbb{P}, \nu, \beta/(mn))$ difference-bounded.*

Lemma 9. *If ℓ is (M, Λ)-admissible, then L is M-uniformly range-bounded.*

B.3 Proof of Theorem 2

Let

$$\Phi(h, \mathbf{Z}) \triangleq \overline{L}(h) - L(h, \hat{\mathbf{Z}}),$$

and note that

$$\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\Phi(h, \hat{\mathbf{Z}})\right] = \overline{L}(\mathbb{Q}) - L(\mathbb{Q}, \hat{\mathbf{Z}}).$$

Recall that the "bad" set $\mathcal{B}_{\mathcal{H}}$ is the set of hypotheses that do not have β-uniform collective stability. By assumption, this set has measure $\mathbb{Q}\{h \in \mathcal{B}_{\mathcal{H}}\} \leq \eta$ for all applicable posteriors. Further, by Lemma 9, L is M-uniformly range-bounded, so

$$\begin{aligned}
\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\Phi(h, \hat{\mathbf{Z}}) \,|\, h \in \mathcal{B}_{\mathcal{H}}\right] &\leq \sup_{h\in\mathcal{B}_{\mathcal{H}}} \overline{L}(h) - L(h, \hat{\mathbf{Z}}) \\
&\leq \sup_{\substack{h\in\mathcal{B}_{\mathcal{H}} \\ \mathbf{z}\in\mathcal{Z}^n}} L(h, \mathbf{z}) - L(h, \hat{\mathbf{Z}}) \\
&\leq M.
\end{aligned}$$

For Φ' as defined in Equation 4, and a free parameter $u \in \mathbb{R}$, we have that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\Phi(h, \hat{\mathbf{Z}})\right] \\
&= \mathbb{Q}\{h \in \mathcal{B}_{\mathcal{H}}\} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\Phi(h, \hat{\mathbf{Z}}) \,|\, h \in \mathcal{B}_{\mathcal{H}}\right] + \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[\Phi'(h, \hat{\mathbf{Z}})\right] \\
&\leq \eta M + \frac{1}{u} \mathop{\mathbb{E}}_{h\sim\mathbb{Q}}\left[u\,\Phi'(h, \hat{\mathbf{Z}})\right] \\
&\leq \eta M + \frac{1}{u}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[e^{u\Phi'(h, \hat{\mathbf{Z}})}\right]\right), \quad (19)
\end{aligned}$$

where the last inequality follows from Lemma 6.

What remains is to bound $\mathbb{E}_{h\sim\mathbb{H}}\left[e^{u\Phi'(h, \hat{\mathbf{Z}})}\right]$ and optimize u. Since the KL divergence term is a function of the (learned) posterior, we cannot optimize u for

all posteriors simultaneously. We therefore use discretization to cover the range of optimal parameter values for all possible posteriors, thereby ensuring that Equation 19 is bounded with high probability for all discrete values.

Let $\beta_{\ell\circ\mathcal{H}} \triangleq M + \Lambda\beta$. By Lemma 7, $\beta_{\ell\circ\mathcal{H}}$ is a uniform upper bound on the uniform collective stability of $\ell \circ h$ for any "good" hypothesis $h \notin \mathcal{B}_{\mathcal{H}}$. Further, by Lemma 8, $L(h,\cdot)$ is $(\beta_{\ell\circ h}/(mn))$-uniformly difference-bounded. Therefore, since Φ' outputs 0 for any $h \in \mathcal{B}_{\mathcal{H}}$, we have that $\Phi'(h,\cdot)$ is $(\beta_{\ell\circ\mathcal{H}}/(mn))$-uniformly difference-bounded for all $h \in \mathcal{H}$.

For $j = 0, 1, 2, \ldots$, define a parameter

$$u_j \triangleq 2^j \sqrt{\frac{8mn \ln \frac{2}{\delta}}{\beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}},$$

let $\delta_j \triangleq \delta 2^{-(j+1)}$, and define an event

$$E_j \triangleq \mathbb{1}\left\{ \mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[e^{u_j \Phi'(h,\hat{\mathbf{Z}})}\right] \geq \frac{1}{\delta_j} \exp\left(\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right)\right\}.$$

By the union bound, the probability that any E_j occurs is

$$\mathbb{P}\left\{\bigcup_{j=0}^\infty E_j\right\} \leq \sum_{j=0}^\infty \mathbb{P}\{E_j\}.$$

Further, by Markov's inequality and the law of total expectation,

$$\begin{aligned}
&\mathbb{P}\{E_j\} \\
&\leq \delta_j \exp\left(-\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right) \mathop{\mathbb{E}}_{\hat{\mathbf{Z}}\sim\mathbb{P}^m} \mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[e^{u_j \Phi'(h,\hat{\mathbf{Z}})}\right] \\
&= \delta_j \exp\left(-\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right) \mathop{\mathbb{E}}_{h\sim\mathbb{H}} \mathop{\mathbb{E}}_{\hat{\mathbf{Z}}\sim\mathbb{P}^m}\left[e^{u_j \Phi'(h,\hat{\mathbf{Z}})}\right].
\end{aligned}$$

Since $\Phi'(h,\cdot)$ is $(\beta_{\ell\circ\mathcal{H}}/(mn))$-uniformly difference-bounded, and $\mathbb{E}_{\hat{\mathbf{Z}}\sim\mathbb{P}^m}[\Phi'(h,\hat{\mathbf{Z}})] = 0$, we apply Corollary 2 to the above inequality and obtain

$$\begin{aligned}
\mathbb{P}\left\{\bigcup_{j=0}^\infty E_j\right\} &\leq \sum_{j=0}^\infty \delta_j \exp\left(-\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right) \\
&\qquad \times \mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[\exp\left(\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right)\right] \\
&= \sum_{j=0}^\infty \delta_j = \delta.
\end{aligned}$$

Therefore, with probability at least $1-\delta$ over realizations of $\hat{\mathbf{Z}}$, *every* u_j satisfies

$$\mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[e^{u_j \Phi'(h,\hat{\mathbf{Z}})}\right] \leq \frac{1}{\delta_j} \exp\left(\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right). \quad (20)$$

Note that we have applied Corollary 2 to a function of mn interdependent variables. However, since there is independence between examples, the dependency matrix of $\hat{\mathbf{Z}}$ (denoted $\boldsymbol{\Theta}_{mn}^{\boldsymbol{\pi}}$) is in fact block diagonal, with each sub-matrix equal to $\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}$. Therefore, $\left\|\boldsymbol{\Theta}_{mn}^{\boldsymbol{\pi}}\right\|_\infty = \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty$.

For any particular posterior $\mathbb{Q}$, there exists an approximately-optimal $u_{j^\star}$ by taking

$$j^\star \triangleq \left\lfloor \frac{1}{2\ln 2} \ln\left(\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H})}{\ln(2/\delta)} + 1\right)\right\rfloor. \quad (21)$$

Since, for all $v \in \mathbb{R}$, $v - 1 \leq \lfloor v \rfloor \leq v$, we can use Equation 21 to show that

$$\frac{1}{2}\sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H})}{\ln(2/\delta)} + 1} \leq 2^{j^\star} \leq \sqrt{\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H})}{\ln(2/\delta)} + 1};$$

therefore,

$$\begin{aligned}
u_{j^\star} &\geq \sqrt{\frac{2mn\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta}\right)}{\beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}} \\
\text{and} \quad u_{j^\star} &\leq \sqrt{\frac{8mn\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta}\right)}{\beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}}. \quad (22)
\end{aligned}$$

Further, by definition of $\delta_{j^\star}$,

$$\begin{aligned}
\ln\frac{1}{\delta_{j^\star}} &= \ln\frac{2}{\delta} + j\ln 2 \\
&\leq \ln\frac{2}{\delta} + \frac{\ln 2}{2\ln 2}\ln\left(\frac{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H})}{\ln(2/\delta)} + 1\right) \\
&= \ln\frac{2}{\delta} + \frac{1}{2}\ln\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta}\right) - \frac{1}{2}\ln\ln\frac{2}{\delta} \\
&\leq \ln\frac{2}{\delta} + \frac{1}{2}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta}\right)
\end{aligned}$$

for all $\delta \in (0,1)$; therefore,

$$D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{1}{\delta_{j^\star}} \leq \frac{3}{2}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta}\right). \quad (23)$$

Putting it all together, we now have that, with probability at least $1-\delta$, the approximately-optimal $u_{j^\star}$ for any posterior $\mathbb{Q}$ satisfies

$$\begin{aligned}
&\frac{1}{u_{j^\star}}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[e^{u_{j^\star}\Phi'(h,\hat{\mathbf{Z}})}\right]\right) \\
&\quad \leq \frac{1}{u_{j^\star}}\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{1}{\delta_{j^\star}} + \frac{u_{j^\star}^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn}\right) \\
&\quad \leq \frac{3\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta}\right)}{2u_{j^\star}} + \frac{u_{j^\star}\beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{8mn} \\
&\quad \leq \frac{2\beta_{\ell\circ\mathcal{H}}\left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty}{\sqrt{2mn}}\sqrt{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta}}.
\end{aligned}$$

The first inequality is due to Equation 20; the second inequality is from Equation 23; the last inequality uses the lower and upper bounds from Equation 22. Combining this with Equation 19, and replacing $\beta_{\ell\circ\mathcal{H}}$ with its definition, completes the proof.

B.4 Proof of Theorem 3

The proof of Theorem 3 proceeds similarly to that of Theorem 2. Let $\mathcal{B}_{\mathcal{H}}$ denote the set of "bad" hypotheses. In Equation 19, we isolate $\mathcal{B}_{\mathcal{H}}$ in the first term, then focus on the concentration of Φ' for the "good" set. In Definition 5, the measure of the "bad" assignments $\mathcal{B}$ depends on whether h is in $\mathcal{B}_{\mathcal{H}}$; conditioned on $h \notin \mathcal{B}_{\mathcal{H}}$, $\mathcal{B}$ has measure at most ν, under $\mathbb{P}$. Further, every $h \notin \mathcal{B}_{\mathcal{H}}$ satisfies Equation 3 with stability β for every pair of "good" inputs $\mathbf{z}, \mathbf{z}' \notin \mathcal{B}$. Therefore, every good hypothesis $h \notin \mathcal{B}_{\mathcal{H}}$ has $(\mathbb{P}, \nu, \beta)$ collective stability.

We again let $\beta_{\ell\circ\mathcal{H}} \triangleq M + \Lambda\beta$. By Lemma 7, $\ell \circ h$ has $(\mathbb{P}, \nu, \beta_{\ell\circ\mathcal{H}})$ collective stability for any $h \notin \mathcal{B}_{\mathcal{H}}$. It therefore follows from Lemma 8 and Equation 4 that $\Phi'(h, \cdot)$ is $(\mathbb{P}, \nu, \beta_{\ell\circ\mathcal{H}}/(mn))$ difference-bounded for all $h \in \mathcal{H}$. Finally, by Lemma 9, L—hence, Φ'—is M-uniformly range-bounded.

Let

$$\delta' \triangleq \delta - \frac{2\nu M(mn)^2}{\beta_{\ell\circ\mathcal{H}}}.$$

By assumption in the theorem statement,

$$\delta > 2\nu(mn)^2 \geq \frac{2\nu M(mn)^2}{M + \Lambda\beta} = \frac{2\nu M(mn)^2}{\beta_{\ell\circ\mathcal{H}}},$$

so $\delta' > 0$. For $j = 0, 1, 2, \ldots$, define a parameter

$$u_j \triangleq 2^j \sqrt{\frac{2mn \ln \frac{2}{\delta'}}{\beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}},$$

let $\delta_j \triangleq \delta' 2^{-(j+1)}$, and define an event

$$E_j \triangleq \mathbb{1}\left\{ \mathop{\mathbb{E}}_{h\sim\mathbb{H}} \left[e^{u_j \Phi'(h, \hat{\mathbf{Z}})} \right] \geq \frac{1}{\delta_j} \exp\left(\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{2mn} \right) \right\}.$$

Let $\mathcal{B}_\lambda$ be as defined in Theorem 1. Using the law of total probability and the union bound, we have that the probability that any E_j occurs is

$$\begin{aligned}
\mathbb{P}\left\{ \bigcup_{j=0}^\infty E_j \right\} &\leq \mathbb{P}\{\hat{\mathbf{Z}} \in \mathcal{B}_\lambda\} + \mathbb{P}\left\{ \bigcup_{j=0}^\infty E_j \,|\, \hat{\mathbf{Z}} \notin \mathcal{B}_\lambda \right\} \\
&\leq \mathbb{P}\{\hat{\mathbf{Z}} \in \mathcal{B}_\lambda\} + \sum_{j=0}^\infty \mathbb{P}\{E_j \,|\, \hat{\mathbf{Z}} \notin \mathcal{B}_\lambda\}.
\end{aligned}$$

By applying Markov's inequality and rearranging the expectations, we obtain

$$\begin{aligned}
&\mathbb{P}\{E_j \,|\, \mathbf{Z} \notin \mathcal{B}_\lambda\} \\
&\quad\leq \delta_j \exp\left(-\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{2mn} \right) \\
&\qquad\times \mathop{\mathbb{E}}_{h\sim\mathbb{H}} \mathop{\mathbb{E}}_{\hat{\mathbf{Z}}\sim\mathbb{P}^m} \left[e^{u_j \Phi'(h, \hat{\mathbf{Z}})} \,|\, \hat{\mathbf{Z}} \notin \mathcal{B}_\lambda \right].
\end{aligned}$$

We then apply Theorem 1, with $\lambda \triangleq \beta_{\ell\circ\mathcal{H}}/(2Mmn)$. Since $\Phi'(h, \cdot)$ is $(\mathbb{P}, \nu, \beta_{\ell\circ\mathcal{H}}/(mn))$ difference-bounded and M-uniformly range-bounded, we have that

$$\mathop{\mathbb{E}}_{\hat{\mathbf{Z}}\sim\mathbb{P}^m} \left[e^{u_j \Phi'(h, \hat{\mathbf{Z}})} \,|\, \hat{\mathbf{Z}} \notin \mathcal{B}_\lambda \right] \leq \exp\left(\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{2mn} \right),$$

and

$$\mathbb{P}\{\hat{\mathbf{Z}} \in \mathcal{B}_\lambda\} \leq \frac{2\nu M(mn)^2}{\beta_{\ell\circ\mathcal{H}}}.$$

Combining these inequalities, we then have that

$$\begin{aligned}
\mathbb{P}\left\{ \bigcup_{j=0}^\infty E_j \right\} &\leq \frac{2\nu M(mn)^2}{\beta_{\ell\circ\mathcal{H}}} + \sum_{j=0}^\infty \delta_j \\
&= \frac{2\nu M(mn)^2}{\beta_{\ell\circ\mathcal{H}}} + \delta' \\
&= \delta.
\end{aligned}$$

Therefore, with probability at least $1 - \delta$ over realizations of $\hat{\mathbf{Z}}$, every u_j satisfies

$$\mathop{\mathbb{E}}_{h\sim\mathbb{H}} \left[e^{u_j \Phi'(h, \hat{\mathbf{Z}})} \right] \leq \frac{1}{\delta_j} \exp\left(\frac{u_j^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{2mn} \right). \quad (24)$$

Now, using Equation 21 (with δ' instead of δ) to select $j^\star$ for a given posterior $\mathbb{Q}$, we have that

$$\begin{aligned}
u_{j^\star} &\geq \sqrt{\frac{mn\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta'}\right)}{2\beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}} \\
\text{and} \quad u_{j^\star} &\leq \sqrt{\frac{2mn\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta'}\right)}{\beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}}. \quad (25)
\end{aligned}$$

Therefore, with probability at least $1 - \delta$, the approximately-optimal $u_{j^\star}$ satisfies

$$\begin{aligned}
&\frac{1}{u_{j^\star}} \left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{H}} \left[e^{u_{j^\star} \Phi'(h, \hat{\mathbf{Z}})} \right] \right) \\
&\quad\leq \frac{1}{u_{j^\star}} \left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{1}{\delta_{j^\star}} + \frac{u_{j^\star}^2 \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{2mn} \right) \\
&\quad\leq \frac{3\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta'}\right)}{2u_{j^\star}} + \frac{u_{j^\star} \beta_{\ell\circ\mathcal{H}}^2 \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty^2}{2mn} \\
&\quad\leq \frac{4\beta_{\ell\circ\mathcal{H}} \left\|\boldsymbol{\Theta}_n^{\boldsymbol{\pi}}\right\|_\infty}{\sqrt{2mn}} \sqrt{D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{H}) + \ln\frac{2}{\delta'}}.
\end{aligned}$$

The first inequality is due to Equation 24; the second inequality is from Equation 23 (with δ' instead of δ); the last inequality uses the lower and upper bounds from Equation 25. To complete the proof, we combine this with Equation 19 and substitute $\beta_{\ell\circ\mathcal{H}}$ and δ' with their respective definitions, noting that

$$\ln \frac{2}{\delta'} = \ln \frac{2}{\delta - \frac{2\nu M (mn)^2}{M+\Lambda\beta}} \leq \ln \frac{2}{\delta - 2\nu (mn)^2}.$$

C STRONG CONVEXITY AND COLLECTIVE STABILITY

C.1 Strong Convexity

The following definition is a specialization of a more general definition involving an arbitrary norm.

Definition 14. A function $\varphi : \mathcal{S} \to \mathbb{R}$ is *κ-strongly convex* (with respect to the 1-norm) if $\mathcal{S}$ is a convex set and, for any $s, s' \in \mathcal{S}$ and $\tau \in [0,1]$,

$$\frac{\kappa}{2}\tau(1-\tau)\left\|s - s'\right\|_1^2 + \varphi(\tau s + (1-\tau)s') \leq \tau\varphi(s) + (1-\tau)\varphi(s').$$

Strongly convex functions have the following useful properties.

Lemma 10. *Let $\varphi : \mathcal{S} \to \mathbb{R}$ be κ-strongly convex, and let $\dot{s} \triangleq \arg\min_{s\in\mathcal{S}} \varphi(s)$. Then, for any $s \in \mathcal{S}$*

$$\left\|s - \dot{s}\right\|_1^2 \leq \frac{2}{\kappa}\left(\varphi(s) - \varphi(\dot{s})\right).$$

Proof Let $\Delta s \triangleq s - \dot{s}$. By Definition 14, for any $\tau \in [0,1]$,

$$\frac{\kappa}{2}\tau(1-\tau)\left\|\Delta s\right\|_1^2 + \varphi(\dot{s}+\tau\Delta s) - \varphi(\dot{s}) \leq \tau\left(\varphi(s) - \varphi(\dot{s})\right).$$

Since $\dot{s}$ is, by definition, the unique minimizer of φ, it follows that $\varphi(\dot{s} + \tau\Delta s) - \varphi(\dot{s}) \geq 0$; so the above inequality is preserved when this term is dropped. Thus, dividing both sides by $\tau\kappa/2$, we have that

$$\left\|\Delta s\right\|_1^2 \leq (1-\tau)\left\|\Delta s\right\|_1^2 \leq \frac{2}{\kappa}\left(\varphi(s) - \varphi(\dot{s})\right),$$

where the left inequality follows from the fact that $(1-\tau)$ is maximized at $\tau = 0$. ■

Lemma 11. *Let $\varphi : \Omega \times \mathcal{S} \to \mathbb{R}$ be κ-strongly convex in $\mathcal{S}$. If, for all $s \in \mathcal{S}$ and $\omega, \omega' \in \Omega : D_{\text{H}}(\omega, \omega') = 1$, $|\varphi(\omega, s) - \varphi(\omega', s)| \leq \Lambda$, then*

$$\left\|\arg\min_{s\in\mathcal{S}} \varphi(\omega, s) - \arg\min_{s'\in\mathcal{S}} \varphi(\omega', s')\right\|_1 \leq \sqrt{\frac{2\Lambda}{\kappa}}.$$

Proof Let $\dot{s} \triangleq \arg\min_{s\in\mathcal{S}} \varphi(\omega, s)$ and $\dot{s}' \triangleq \arg\min_{s'\in\mathcal{S}} \varphi(\omega', s')$. Without loss of generality, assume that $\varphi(\omega, \dot{s}) \geq \varphi(\omega', \dot{s}')$. (If $\varphi(\omega', \dot{s}') \geq \varphi(\omega, \dot{s})$, we could state this in terms of ω'.) Using Lemma 10, we have that

$$\begin{aligned}\left\|\dot{s}' - \dot{s}\right\|_1^2 &\leq \frac{2}{\kappa}\left(\varphi(\omega, \dot{s}') - \varphi(\omega, \dot{s})\right) \\ &\leq \frac{2}{\kappa}\left(\varphi(\omega, \dot{s}') - \varphi(\omega', \dot{s}')\right) \leq \frac{2}{\kappa}\Lambda.\end{aligned}$$

Taking the square root completes the proof. ■

C.2 Proof of Theorem 4

Lemma 11 implies that the maximum of the energy function has uniform collective stability if the negative energy is strongly convex and uniformly difference-bounded. We prove the latter property in the following lemma.

Lemma 12. *For any $h \in \mathcal{H}_{\mathcal{T}}^{\text{SC}}$, with weights $\mathbf{w}$, and any $\mathbf{s} \in \mathcal{S}$, the energy function $E_{\mathbf{w}}(\cdot, \mathbf{s})$ is $(2\left\|\mathbf{w}\right\|_a C_G)$-uniformly difference-bounded.*

Proof Without loss of generality, assume that assignments $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ differ at a single coordinate i. Using Hölder's inequality, we have that

$$\begin{aligned}&|E_{\mathbf{w}}(\mathbf{x}, \mathbf{s}) - E_{\mathbf{w}}(\mathbf{x}', \mathbf{s})| \\ &\quad= |\langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \mathbf{s})\rangle - \Psi(\mathbf{s}) - \langle \mathbf{w}, \mathbf{f}(\mathbf{x}', \mathbf{s})\rangle + \Psi(\mathbf{s})| \\ &\quad= |\langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \mathbf{s}) - \mathbf{f}(\mathbf{x}', \mathbf{s})\rangle| \\ &\quad\leq \left\|\mathbf{w}\right\|_a \left\|\mathbf{f}(\mathbf{x}, \mathbf{s}) - \mathbf{f}(\mathbf{x}', \mathbf{s})\right\|_b.\end{aligned}$$

Note that the features of $(\mathbf{x}, \mathbf{s})$ and $(\mathbf{x}', \mathbf{s})$ only differ at any grounding involving node i. The number of such groundings is uniformly upper-bounded by C_G, so at most C_G features will change. Further, the b-norm of any feature function is, by Definition 9, upper-bounded by 1. Therefore, using the triangle inequality, we have that

$$\begin{aligned}&\left\|\mathbf{f}(\mathbf{x}, \mathbf{s}) - \mathbf{f}(\mathbf{x}', \mathbf{s})\right\|_b \\ &= \left(\sum_{t\in\mathcal{T}} \left\| \sum_{c\in t(G)} \mathbb{1}\{i \in c\}\left(f_t(\mathbf{x}_c, \mathbf{s}_c) - f_t(\mathbf{x}'_c, \mathbf{s}_c)\right)\right\|_b^b\right)^{1/b} \\ &\leq \sum_{t\in\mathcal{T}} \sum_{c\in t(G)} \mathbb{1}\{i \in c\}\left\|f_t(\mathbf{x}_c, \mathbf{s}_c) - f_t(\mathbf{x}'_c, \mathbf{s}_c)\right\|_b \\ &\leq 2C_G.\end{aligned}$$

Since this holds for any single coordinate perturbation, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$ with $D_{\text{H}}(\mathbf{x}, \mathbf{x}') \geq 1$, we have that

$$|E_{\mathbf{w}}(\mathbf{x}, \mathbf{s}) - E_{\mathbf{w}}(\mathbf{x}', \mathbf{s})| \leq (2\left\|\mathbf{w}\right\|_a C_G)\, D_{\text{H}}(\mathbf{x}, \mathbf{x}'),$$

which completes the proof. ∎

We are now equipped to prove Theorem 4. Fix any $h \in \mathcal{H}_{\mathcal{T}}^{\text{SC}}$, with weights $\mathbf{w}$. By Definition 9, ϕ is convex and Ψ is κ-strongly convex, for some $\kappa > 0$. This implies that $-E$ is at least κ-strongly convex in $\mathcal{S}$. Now, fix any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}^n$, and let $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$ denote their respective maximizers of $E_{\mathbf{w}}$. Using the additive property of linear transformations, and the property $\|\Gamma\|_1 \leq 1$, we have that

$$\begin{aligned}\|h(\mathbf{x}) - h(\mathbf{x}')\|_1 &= \|\Gamma(\mathbf{s}) - \Gamma(\mathbf{s}')\|_1 \\ &= \|\Gamma(\mathbf{s} - \mathbf{s}')\|_1 \\ &\leq \|\Gamma\|_1 \|\mathbf{s} - \mathbf{s}'\|_1 \\ &\leq \|\mathbf{s} - \mathbf{s}'\|_1 .\end{aligned}$$

Thus, telescoping $(\mathbf{s} - \mathbf{s}')$ into a sum of single-site differences, and applying Lemmas 11 and 12, we have that

$$\|\mathbf{s} - \mathbf{s}'\|_1 \leq 2\sqrt{\frac{\|\mathbf{w}\|_a C_G}{\kappa}}\, D_{\text{H}}(\mathbf{x}, \mathbf{x}').$$

Corollary 1 follows directly by using R as a uniform upper bound for $\|\mathbf{w}\|_a$.

C.3 Strong Convexity and p-Norms

The requirement of strong convexity with respect to the 1-norm may at first seem restrictive. However, observe that strong convexity with respect to *any* p-norm suffices for collective stability.

Claim 1. *Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a convex set, for $n \in [1, \infty)$, and suppose $\varphi : \mathcal{S} \to \mathbb{R}$ is a differentiable function that is κ-strongly convex with respect to the p-norm, for $p \geq 1$. Then, for $K \geq n^{2-2/p}$, $K\varphi$ is κ-strongly convex with respect to the 1-norm.*

Proof Since φ is differentiable and strongly convex, using an alternate definition of strong convexity, one can shown that

$$\kappa \|s - s'\|_p^2 \leq \langle \nabla\varphi(s) - \nabla\varphi(s'), s - s' \rangle ,$$

for any $s, s \in \mathcal{S}$. To lower-bound the left-hand side, we use the following p-norm identity: for any $\mathbf{v} \in \mathbb{R}^n$ and $p \geq 1$,

$$\|\mathbf{v}\|_1 \leq n^{1-1/p} \|\mathbf{v}\|_p .$$

Since all p-norms are nonnegative, this inequality holds when we square both sides. Now, let $\tilde{\varphi}(s) \triangleq K\varphi(s)$, and note that $\nabla\tilde{\varphi}(s) = K\nabla\varphi(s)$. We therefore have that

$$\begin{aligned}\kappa \|s - s'\|_1^2 &\leq \kappa\, n^{2-2/p} \|s - s'\|_p^2 \\ &\leq K\kappa \|s - s'\|_p^2 \\ &\leq K \langle \nabla\varphi(s) - \nabla\varphi(s'), s - s' \rangle \\ &= \langle \nabla\tilde{\varphi}(s) - \nabla\tilde{\varphi}(s'), s - s' \rangle\end{aligned}$$

which completes the proof. ∎

It is common that $p = 2$, in which case $K \geq n$ suffices. One should also note that Theorem 4 does not depend on the magnitude of Ψ; thus, we can replace Ψ with any suitably scaled, strongly convex surrogate, without penalty. That said, scaling Ψ may affect the inference algorithm, and therefore also affect the empirical risk.

D PROOFS OF EXAMPLES

D.1 Modified Margin Loss

In collective classification, one often wishes to bound the expected 0-1 loss, ℓ_0. Unfortunately, this loss is not admissible, so one cannot directly apply our generalization bounds. A common workaround is to upper-bound ℓ_0 using a surrogate loss that satisfies admissibility. For this, we use a *modified* margin loss,

$$\ell_{\gamma,\rho}(y, \hat{y}) \triangleq r_{\gamma,\rho}\left(\langle y, \hat{y}\rangle - \max_{y' \in \mathcal{Y} : y \neq y'} \langle y', \hat{y}\rangle\right),$$

where $\gamma \geq 0$, $\rho \geq 0$ and $r_{\gamma,\rho}$ is the *thresholded ramp function*,

$$r_{\gamma,\rho}(\alpha) \triangleq \begin{cases} 1 & \text{for } \alpha \leq \gamma, \\ 1 - (\alpha - \gamma)/\rho & \text{for } \gamma < \alpha < \gamma + \rho, \\ 0 & \text{for } \alpha \geq \gamma + \rho. \end{cases}$$

Note that $\ell_{0,0} \equiv \ell_0$ and $\ell_{\gamma,0} \equiv \ell_\gamma$.

Lemma 13. *The modified margin loss, $\ell_{\gamma,\rho}$, is $(1, 1/\rho)$-admissible for any $\gamma \geq 0$ and $\rho > 0$, and 1-uniformly range-bounded over all $\gamma \geq 0$ and $\rho \geq 0$.*

Proof By definition, $\ell_{\gamma,\rho}$ is bounded in the interval $[0, 1]$, independent of γ and ρ. Thus, it is 1-uniformly range-bounded over all values of γ and ρ, which also establishes the first admissibility condition for a given γ and ρ.

For any $\hat{y}, \hat{y}' \in \hat{\mathcal{Y}}$, let $u \triangleq \arg\max_{y' \in \mathcal{Y} : y \neq y'} \langle y', \hat{y}\rangle$ and $u' \triangleq \arg\max_{y' \in \mathcal{Y} : y \neq y'} \langle y', \hat{y}'\rangle$. Without loss of generality, assume that $\langle y, \hat{y}\rangle - \langle u, \hat{y}\rangle \geq \langle y, \hat{y}'\rangle - \langle u', \hat{y}'\rangle$. For any $y \in \mathcal{Y}$, we have that

$$\begin{aligned}&|(\langle y, \hat{y}\rangle - \langle u, \hat{y}\rangle) - (\langle y, \hat{y}'\rangle - \langle u', \hat{y}'\rangle)| \\ &\quad = |\langle y, \hat{y} - \hat{y}'\rangle + \langle u', \hat{y}'\rangle - \langle u, \hat{y}\rangle| \\ &\quad \leq |\langle y, \hat{y} - \hat{y}'\rangle + \langle u', \hat{y}'\rangle - \langle u', \hat{y}\rangle| \\ &\quad = |\langle y - u', \hat{y} - \hat{y}'\rangle| \\ &\quad \leq \|y - u'\|_\infty \|\hat{y} - \hat{y}'\|_1 \\ &\quad \leq \|\hat{y} - \hat{y}'\|_1 .\end{aligned}$$

Further, for any $a, a' \in \mathbb{R}$,

$$|r_{\gamma,\rho}(a) - r_{\gamma,\rho}(a')| \leq \left|\frac{a - \gamma}{\rho} - \frac{a' - \gamma}{\rho}\right| = \frac{1}{\rho}|a - a'| .$$

Combining these inequalities, we have that $|\ell_{\gamma,\rho}(y,\hat{y}) - \ell_{\gamma,\rho}(y,\hat{y}')| \leq (1/\rho)\left\|\hat{y} - \hat{y}'\right\|_1$, which establishes the second admissibility condition. ■

D.2 Properties of Pairwise TSMs

The class of pairwise TSMs have some useful structural properties.

Lemma 14. *Let G be a graph on n nodes, with maximum degree Δ_G. Let $\mathcal{T}$ contain only the unary and pairwise clique templates. Then, the following hold:*

1. *The maximum number of groundings involving any single variable is at most $\Delta_G + 1$.*
2. *The total number of groundings is at most*

$$\frac{n(\Delta_G + 2)}{2}.$$

Proof Any single node may only participate in one unary grounding and up to Δ_G pairwise groundings. By the handshaking lemma, the number of edges in the graph is equal to the sum of the degrees, divided by two; this is at most $n\Delta_G/2$. Since there are n nodes, this makes the total number of groundings at most $n + n\Delta_G/2$. ■

D.3 Proof of Theorem 5

For the proof, we will use a modified margin loss described in Appendix D.1. The benefit of this loss is that it is admissible, per Lemma 13.

Define the prior $\mathbb{H}$ as the uniform distribution on $(\pm R)^d$. Given a (learned) hypothesis $h \in \mathcal{H}^{\text{PAM}}_{R,\kappa}$, with parameters $\mathbf{w} \in (\pm R)^d$, define the posterior $\mathbb{Q}_h$ as the uniform distribution on the hypercube $(\mathbf{w} \pm \varepsilon)^d \cap (\pm R)^d$, where

$$\varepsilon \triangleq \frac{\kappa\gamma^2}{9n(\Delta_G + 2)}.$$

The proof requires two intermediate lemmas: first, we show that the loss of h is always "close" to that of any hypothesis drawn from the posterior; then, we bound the KL divergence of the constructed prior and posterior.

Lemma 15. *For any $h \in \mathcal{H}^{\text{PAM}}_{R,\kappa}$ and $\mathbf{z} \in \mathcal{Z}^n$,*

$$L^0(h,\mathbf{z}) \leq L^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h,\mathbf{z}) \leq L^{\gamma}(h,\mathbf{z}).$$

Proof Fix any $h' \sim \mathbb{Q}_h$, and let $\mathbf{w}$ and $\mathbf{w}'$ denote the respective weights of h and h'. Recall that $\mathbf{z} = (\mathbf{x},\mathbf{y})$. Let $\mathbf{s}$ and $\mathbf{s}'$ be the respective maximizers of $E_{\mathbf{w}}(\mathbf{x},\cdot)$ and $E_{\mathbf{w}'}(\mathbf{x},\cdot)$. Since Γ is a projection with $\|\Gamma\|_1 \leq 1$,

$$\left\|h(\mathbf{x}) - h'(\mathbf{x})\right\|_1 = \left\|\Gamma(\mathbf{s} - \mathbf{s}')\right\|_1 \leq \left\|\mathbf{s} - \mathbf{s}'\right\|_1.$$

Further, since $-E$ is κ-strongly convex in $\mathcal{S}$, using Lemma 10, we have that

$$\begin{aligned}
\left\|\mathbf{s} - \mathbf{s}'\right\|_1^2 &= \frac{1}{2}\left(\left\|\mathbf{s}' - \mathbf{s}\right\|_1^2 + \left\|\mathbf{s} - \mathbf{s}'\right\|_1^2\right) \\
&\leq \frac{1}{\kappa}\left(E_{\mathbf{w}}(\mathbf{x},\mathbf{s}) - E_{\mathbf{w}}(\mathbf{x},\mathbf{s}')\right. \\
&\quad \left. + E_{\mathbf{w}'}(\mathbf{x},\mathbf{s}') - E_{\mathbf{w}'}(\mathbf{x},\mathbf{s})\right) \\
&= \frac{1}{\kappa}\left(\langle \mathbf{w}, \mathbf{f}(\mathbf{x},\mathbf{s}) - \mathbf{f}(\mathbf{x},\mathbf{s}')\rangle - \Psi(\mathbf{s}) + \Psi(\mathbf{s}')\right. \\
&\quad \left. + \langle \mathbf{w}', \mathbf{f}(\mathbf{x},\mathbf{s}') - \mathbf{f}(\mathbf{x},\mathbf{s})\rangle - \Psi(\mathbf{s}') + \Psi(\mathbf{s})\right) \\
&= \frac{1}{\kappa}\left\langle \mathbf{w} - \mathbf{w}', \mathbf{f}(\mathbf{x},\mathbf{s}) - \mathbf{f}(\mathbf{x},\mathbf{s}')\right\rangle. \qquad (26)
\end{aligned}$$

Now, using Hölder's inequality,

$$\begin{aligned}
&\frac{1}{\kappa}\left\langle \mathbf{w} - \mathbf{w}', \mathbf{f}(\mathbf{x},\mathbf{s}) - \mathbf{f}(\mathbf{x},\mathbf{s}')\right\rangle \\
&\qquad \leq \frac{1}{\kappa}\left\|\mathbf{w} - \mathbf{w}'\right\|_\infty \left\|\mathbf{f}(\mathbf{x},\mathbf{s}) - \mathbf{f}(\mathbf{x},\mathbf{s}')\right\|_1.
\end{aligned}$$

Due to the construction of $\mathbb{Q}_h$,

$$\left\|\mathbf{w} - \mathbf{w}'\right\|_\infty \leq \varepsilon = \frac{\kappa\gamma^2}{9n(\Delta_G + 2)}.$$

Moreover, since the features of $\mathcal{H}^{\text{PAM}}_{R,\kappa}$ obey the simplex constraint,

$$\begin{aligned}
&\left\|\mathbf{f}(\mathbf{x},\mathbf{s}) - \mathbf{f}(\mathbf{x},\mathbf{s}')\right\|_1 \\
&\qquad = \sum_{t\in\mathcal{T}}\left\|\sum_{c\in t(G)}\left(f_t(\mathbf{x}_c,\mathbf{s}_c) - f_t(\mathbf{x}_c,\mathbf{s}'_c)\right)\right\|_1 \\
&\qquad \leq \sum_{t\in\mathcal{T}}\sum_{c\in t(G)}\left\|f_t(\mathbf{x}_c,\mathbf{s}_c) - f_t(\mathbf{x}_c,\mathbf{s}'_c)\right\|_1 \\
&\qquad \leq \sum_{t\in\mathcal{T}}\sum_{c\in t(G)} 2 \\
&\qquad \leq n(\Delta_G + 2), \qquad (27)
\end{aligned}$$

where the last inequality follows from Lemma 14. Combining these inequalities, we have that

$$\begin{aligned}
&\left\|h(\mathbf{x}) - h'(\mathbf{x})\right\|_\infty \\
&\qquad \leq \left\|h(\mathbf{x}) - h'(\mathbf{x})\right\|_1 \\
&\qquad \leq \left\|\mathbf{s} - \mathbf{s}'\right\|_1 \\
&\qquad \leq \sqrt{\frac{1}{\kappa}\left\|\mathbf{w} - \mathbf{w}'\right\|_\infty \left\|\mathbf{f}(\mathbf{x},\mathbf{s}) - \mathbf{f}(\mathbf{x},\mathbf{s}')\right\|_1} \\
&\qquad \leq \sqrt{\frac{1}{\kappa}\cdot\frac{\kappa\gamma^2}{9n(\Delta_G + 2)}\cdot n(\Delta_G + 2)} \\
&\qquad = \frac{\gamma}{3}.
\end{aligned}$$

This means that each coordinate of the output vectors differs by at most $\gamma/3$. As a result,

$$\begin{aligned}
\ell_0(y_i, h_i(\mathbf{x})) = 1 &\implies \ell_{\frac{\gamma}{3},\frac{\gamma}{3}}(y_i, h'_i(\mathbf{x})) = 1; \\
\ell_{\frac{\gamma}{3},\frac{\gamma}{3}}(y_i, h'_i(\mathbf{x})) < 1 &\implies \ell_0(y_i, h_i(\mathbf{x})) = 0.
\end{aligned}$$

Similarly,

$$0 < \ell_{\frac{\gamma}{3},\frac{\gamma}{3}}(y_i, h_i'(\mathbf{x})) \le 1 \implies \ell_\gamma(y_i, h_i(\mathbf{x})) = 1;$$
$$\ell_\gamma(y_i, h_i(\mathbf{x})) = 0 \implies \ell_{\frac{\gamma}{3},\frac{\gamma}{3}}(y_i, h_i'(\mathbf{x})) = 0.$$

Therefore, for any $h \in \mathcal{H}^{\text{PAM}}_{R,\kappa}$, $h' \sim \mathbb{Q}_h$, $\mathbf{z} \in \mathcal{Z}^n$ and $i \in [n]$,

$$\ell_0(y_i, h_i(\mathbf{x})) \le \ell_{\frac{\gamma}{3},\frac{\gamma}{3}}(y_i, h_i'(\mathbf{x})) \le \ell_\gamma(y_i, h_i(\mathbf{x})),$$

which means that

$$L^0(h, \mathbf{z}) \le L^{\frac{\gamma}{3},\frac{\gamma}{3}}(h', \mathbf{z}) \le L^\gamma(h, \mathbf{z}). \tag{28}$$

Taking the expectation over $h' \sim \mathbb{Q}_h$ completes the proof. ∎

Lemma 16. *For any $h \in \mathcal{H}^{\text{PAM}}_{R,\kappa}$,*

$$D_{\text{KL}}(\mathbb{Q}_h \| \mathbb{H}) \le d \ln\left(\frac{18Rn(\Delta_G + 2)}{\kappa\gamma^2}\right).$$

Proof For a uniform distribution $\mathbb{U}$, denote by $\text{dom}(\mathbb{U})$ its domain, and define its *volume* as

$$\text{vol}(\mathbb{U}) \triangleq \int \mathbb{1}\{x \in \text{dom}(\mathbb{U})\}\, \mathrm{d}x.$$

Recall that $\mathcal{H}^{\text{PAM}}_{R,\kappa}$ is essentially just the hypercube $(\pm R)^d$; therefore, $\text{vol}(\mathbb{H}) = (2R)^d$. Similarly,

$$\text{vol}(\mathbb{Q}_h) \ge \varepsilon^d = \left(\frac{\kappa\gamma^2}{9n(\Delta_G + 2)}\right)^d.$$

(The lower-bound, ε^d, is because $\mathbb{Q}_h$ is truncated if $\mathbf{w}$ is at a corner of the hypercube $(\pm R)^d$.) Denote by p and q_h the respective density functions of $\mathbb{H}$ and $\mathbb{Q}_h$. By definition,

$$\begin{aligned} D_{\text{KL}}(\mathbb{Q}_h \| \mathbb{H}) &= \int q_h(h') \ln \frac{q_h(h')}{p(h')}\, \mathrm{d}h' \\ &= \int \frac{\mathbb{1}\{h' \in \text{dom}(\mathbb{Q}_h)\}}{\text{vol}(\mathbb{Q}_h)} \ln \frac{\text{vol}(\mathbb{H})}{\text{vol}(\mathbb{Q}_h)}\, \mathrm{d}h' \\ &= \ln \frac{\text{vol}(\mathbb{H})}{\text{vol}(\mathbb{Q}_h)} \\ &\le \ln\left(2R \cdot \frac{9n(\Delta_G + 2)}{\kappa\gamma^2}\right)^d, \end{aligned}$$

which completes the proof. ∎

We are now ready to prove Theorem 5. Via Corollary 1 and Lemma 14, the class $\mathcal{H}^{\text{PAM}}_{R,\kappa}$ has β-uniform collective stability with

$$\beta \triangleq 2\sqrt{\frac{RC_G}{\kappa}} \le 2\sqrt{\frac{R(\Delta_G + 1)}{\kappa}}.$$

It therefore has $(\mathbb{Q}, 0, \beta)$ collective stability with respect to any posterior $\mathbb{Q}$. We apply Theorem 2 to the prior $\mathbb{H}$ and posterior $\mathbb{Q}_h$, using Lemma 13 for the admissibility of $\ell_{\frac{\gamma}{3},\frac{\gamma}{3}}$ and Lemma 16 to upper-bound $D_{\text{KL}}(\mathbb{Q}_h \| \mathbb{H})$, which yields

$$\begin{aligned} &\overline{L}^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h) - L^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) \\ &\quad\le 0 \times M + \frac{2\left\|\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n\right\|_\infty}{\sqrt{2mn}}\left(1 + \frac{6}{\gamma}\sqrt{\frac{R(\Delta_G + 1)}{\kappa}}\right) \\ &\qquad\times \sqrt{d \ln\left(\frac{18Rn(\Delta_G + 2)}{\kappa\gamma^2}\right) + \ln\frac{2}{\delta}}. \end{aligned}$$

To complete the proof, we use Lemma 15 to lower-bound $\overline{L}^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h)$ and upper-bound $L^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h, \hat{\mathbf{Z}})$. Since $L^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h, \mathbf{Z})$ dominates $L^0(h, \mathbf{Z})$, taking the expectation over $\mathbf{Z}$ yields $\overline{L}^0(h) \le \overline{L}^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h)$. Similarly, since $L^\gamma(h, \mathbf{Z})$ dominates $L^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h, \mathbf{Z})$, it follows for m examples that $L^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) \le L^\gamma(h, \hat{\mathbf{Z}})$.

D.4 Proof of Theorem 6

As in the proof of Theorem 5, we will use the modified margin loss from Appendix D.1. Throughout this section, it will be convenient to use the shorthand

$$\omega \triangleq \frac{\mathbf{w}}{\kappa} \tag{29}$$

for parameters $(\mathbf{w}, \kappa)$ of a hypothesis $h \in \mathcal{H}^{\text{PVC}}$. We define the prior $\mathbb{H}$ as an isotropic, unit-variance Gaussian, with density

$$p(h) \triangleq \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|\omega\|_2^2\right).$$

Let

$$\varsigma \triangleq \left(\frac{9n(\Delta_G + 2)}{\gamma^2}\right)^2 \ln(mn). \tag{30}$$

(Note that $\varsigma \ge 1$, due to our assumptions that $n \ge 2$ and $\gamma \le \sqrt{n}$.) Given a (learned) hypothesis $h \in \mathcal{H}^{\text{PVC}}$, we define the posterior $\mathbb{Q}_h$ as an isotropic Gaussian, with mean $\omega = \mathbf{w}/\kappa$ and variance $1/\varsigma$, whose density is

$$q_h(h') \triangleq \left(\frac{\varsigma}{2\pi}\right)^{d/2} \exp\left(-\frac{\varsigma}{2}\|\omega' - \omega\|_2^2\right).$$

The proof proceeds similarly to that of Theorem 5, via a sequence of intermediate lemmas. We first show that the loss of the deterministic predictor, h, is almost always "close" to the loss of a hypothesis drawn from the posterior (i.e., the *Gibbs classifier*). We then bound the KL divergence of the constructed prior and posterior. Finally, we show that, with high probability, the collective stability of a random predictor from $\mathcal{H}^{\text{PVC}}$ is within a constant additive factor of the stability of h.

Lemma 17. *For any $h \in \mathcal{H}^{\text{PVC}}$ and $\mathbf{z} \in \mathcal{Z}^n$,*

$$L^0(h, \mathbf{z}) \leq L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(\mathbb{Q}_h, \mathbf{z}) + \frac{1}{\sqrt{mn}}$$
$$\text{and} \quad L^{\gamma}(h, \mathbf{z}) \geq L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(\mathbb{Q}_h, \mathbf{z}) - \frac{1}{\sqrt{mn}}.$$

Proof The proof is similar to that of Lemma 15. Fix any $h' \sim \mathbb{Q}_h$, and let $(\mathbf{w}, \kappa)$ and $(\mathbf{w}', \kappa')$ denote the respective parameters of h and h'. Let $\mathbf{s}$ and $\mathbf{s}'$ be the respective maximizers of $E_{\mathbf{w},\kappa}(\mathbf{x}, \cdot)$ and $E_{\mathbf{w}',\kappa'}(\mathbf{x}, \cdot)$. Adapting Equation 26, we have that

$$\begin{aligned}
&\|\mathbf{s} - \mathbf{s}'\|_1^2 \\
&\quad = \frac{1}{2}\left(\|\mathbf{s} - \mathbf{s}'\|_1^2 + \|\mathbf{s}' - \mathbf{s}\|_1^2\right) \\
&\quad \leq \frac{1}{\kappa'}\left(E_{\mathbf{w}',\kappa'}(\mathbf{x}, \mathbf{s}') - E_{\mathbf{w}',\kappa'}(\mathbf{x}, \mathbf{s})\right) \\
&\qquad + \frac{1}{\kappa}\left(E_{\mathbf{w},\kappa}(\mathbf{x}, \mathbf{s}) - E_{\mathbf{w},\kappa}(\mathbf{x}, \mathbf{s}')\right) \\
&\quad = \frac{1}{\kappa'}\left(\langle \mathbf{w}', \mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})\rangle - \kappa'\Psi(\mathbf{s}') + \kappa'\Psi(\mathbf{s})\right) \\
&\qquad + \frac{1}{\kappa}\left(\langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \mathbf{s}) - \mathbf{f}(\mathbf{x}, \mathbf{s}')\rangle - \kappa\Psi(\mathbf{s}) + \kappa\Psi(\mathbf{s}')\right) \\
&\quad = \frac{1}{\kappa'}\langle \mathbf{w}', \mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})\rangle \\
&\qquad + \frac{1}{\kappa}\langle \mathbf{w}, \mathbf{f}(\mathbf{x}, \mathbf{s}) - \mathbf{f}(\mathbf{x}, \mathbf{s}')\rangle \\
&\quad = \langle \omega', \mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})\rangle + \langle \omega, \mathbf{f}(\mathbf{x}, \mathbf{s}) - \mathbf{f}(\mathbf{x}, \mathbf{s}')\rangle \\
&\quad = \langle \omega' - \omega, \mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})\rangle .
\end{aligned}$$

We also have that

$$\begin{aligned}
&\mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s}) \\
&\qquad = \frac{\mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})}{\|\mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})\|_1} \cdot \|\mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})\|_1 \\
&\qquad \leq \frac{\mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})}{\|\mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})\|_1} \cdot n(\Delta_G + 2),
\end{aligned}$$

via Equation 27, since the features of $\mathcal{H}^{\text{PVC}}$ obey the simplex constraint and the templates are unary and pairwise. For notational convenience, let

$$\Delta\mathbf{f} \triangleq \frac{\mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})}{\|\mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})\|_1}.$$

Note that $\Delta\mathbf{f}$ has $\|\Delta\mathbf{f}\|_2 \leq 1$.

Define the event

$$E \triangleq \mathbb{1}\left\{\langle \omega' - \omega, \Delta\mathbf{f}\rangle \geq \frac{\gamma^2}{9n(\Delta_G + 2)}\right\}.$$

Since $\mathbb{Q}_h$ is Gaussian, with mean ω and variance $1/\varsigma$, using Lemma 5 and Equation 30, we have that

$$\begin{aligned}
\mathbb{Q}_h\{E\} &\leq \exp\left(-\frac{\varsigma}{2}\left(\frac{\gamma^2}{9n(\Delta_G + 2)}\right)^2\right) \\
&= \exp\left(-\frac{\ln(mn)}{2}\right) \\
&= \frac{1}{\sqrt{mn}}. \qquad (31)
\end{aligned}$$

This means that, with probability at least $1 - (mn)^{-1/2}$ over draws of $h' \sim \mathbb{Q}_h$, $\langle \omega' - \omega, \Delta\mathbf{f}\rangle \leq \frac{\gamma^2}{9n(\Delta_G+2)}$, and

$$\begin{aligned}
\|h(\mathbf{x}) - h'(\mathbf{x})\|_1 &= \|\Gamma(\mathbf{s} - \mathbf{s}')\|_1 \\
&\leq \|\mathbf{s} - \mathbf{s}'\|_1 \\
&\leq \sqrt{\langle \omega' - \omega, \mathbf{f}(\mathbf{x}, \mathbf{s}') - \mathbf{f}(\mathbf{x}, \mathbf{s})\rangle} \\
&\leq \sqrt{\langle \omega' - \omega, \Delta\mathbf{f}\rangle\, n(\Delta_G + 2)} \\
&\leq \sqrt{\frac{\gamma^2}{9n(\Delta_G + 2)} \cdot n(\Delta_G + 2)} \\
&= \frac{\gamma}{3}.
\end{aligned}$$

Now, for any $\mathbf{z} \in \mathcal{Z}^n$,

$$\begin{aligned}
&L^0(h, \mathbf{z}) - L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(\mathbb{Q}_h, \mathbf{z}) \\
&\quad = L^0(h, \mathbf{z}) - \mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(h', \mathbf{z})\right] \\
&\quad = \mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[L^0(h, \mathbf{z}) - L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(h', \mathbf{z})\right] \\
&\quad \leq \mathbb{Q}_h\{E\} \mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[L^0(h, \mathbf{z}) - L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(h', \mathbf{z}) \,|\, E\right] \\
&\qquad + \mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[L^0(h, \mathbf{z}) - L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(h', \mathbf{z}) \,|\, \neg E\right].
\end{aligned}$$

Recall from Lemma 13 that $\ell_{\gamma,\rho}$—hence, $L^{\gamma,\rho}$—is 1-uniformly range-bounded over all inputs and values of γ and ρ. Therefore, using Equation 31 to upper-bound the measure of E, we have that

$$\mathbb{Q}_h\{E\} \mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[L^0(h, \mathbf{z}) - L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(h', \mathbf{z}) \,|\, E\right] \leq \frac{1}{\sqrt{mn}}.$$

Further, conditioned on $\neg E$, we have that each coordinate of the output vectors differs by at most $\gamma/3$. Using the same reasoning as in the proof of Lemma 15, we would then have that Equation 28 holds, so

$$\mathop{\mathbb{E}}_{h' \sim \mathbb{Q}_h}\left[L^0(h, \mathbf{z}) - L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(h', \mathbf{z}) \,|\, \neg E\right] \leq 0.$$

Therefore, combining the inequalities, we have that

$$L^0(h, \mathbf{z}) - L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(\mathbb{Q}_h, \mathbf{z}) \leq \frac{1}{\sqrt{mn}}.$$

By the same reasoning, it is straightforward to show that

$$L^{\frac{\gamma}{3}, \frac{\gamma}{3}}(\mathbb{Q}_h, \mathbf{z}) - L^{\gamma}(h, \mathbf{z}) \leq \frac{1}{\sqrt{mn}}.$$

The lemma follows directly from these inequalities. ■

Lemma 18. *For any $h \in \mathcal{H}^{\text{PVC}}$,*

$$D_{\text{KL}}(\mathbb{Q}_h \| \mathbb{H}) \leq d \ln\left(\frac{9n(\Delta_G+2)}{\gamma^2}\sqrt{\ln(mn)}\right) + \frac{\|\mathbf{w}\|_2^2}{2\kappa^2}.$$

Proof By definition,

$$\begin{aligned}
&D_{\text{KL}}(\mathbb{Q}_h \| \mathbb{H}) \\
&= \int_{h'} q_h(h') \ln \frac{q_h(h')}{p(h')} \mathrm{d}h' \\
&= \int_{h'} q_h(h') \ln \frac{\left(\frac{\varsigma}{2\pi}\right)^{d/2} e^{-\frac{\varsigma}{2}\|\omega'-\omega\|_2^2}}{\left(\frac{1}{2\pi}\right)^{d/2} e^{-\frac{1}{2}\|\omega'\|_2^2}} \mathrm{d}h' \\
&= \int_{h'} q_h(h') \left(\frac{d}{2}\ln\varsigma + \frac{1}{2}\|\omega'\|_2^2 - \frac{\varsigma}{2}\|\omega'-\omega\|_2^2\right) \mathrm{d}h' \\
&\leq \int_{h'} q_h(h') \left(\frac{d}{2}\ln\varsigma + \frac{1}{2}\|\omega'\|_2^2 - \frac{1}{2}\|\omega'-\omega\|_2^2\right) \mathrm{d}h' \\
&\leq \int_{h'} q_h(h') \left(\frac{d}{2}\ln\varsigma + \frac{1}{2}\|\omega'-\omega'+\omega\|_2^2\right) \mathrm{d}h' \\
&= \int_{h'} q_h(h') \left(\frac{d}{2}\ln\varsigma + \frac{1}{2}\|\omega\|_2^2\right) \mathrm{d}h' \\
&= \frac{d}{2}\ln\varsigma + \frac{1}{2}\|\omega\|_2^2 .
\end{aligned}$$

The first inequality follows from $\varsigma \geq 1$, by assumptions $n \geq 2$ and $\gamma \leq \sqrt{n}$; the second inequality follows from the triangle inequality. Substituting Equation 29 for ω, and Equation 30 for ς, completes the proof. ■

Lemma 19. *For any $h \in \mathcal{H}^{\text{PVC}}$, the class $\mathcal{H}^{\text{PVC}}$ has*

$$\left(\mathbb{Q}_h\,,\, \frac{2d}{mn}\,,\, 2\sqrt{\left(\frac{\|\mathbf{w}\|_\infty}{\kappa}+1\right)(\Delta_G+1)}\right)$$

collective stability.

Proof Define the "bad" set as

$$\mathcal{B}^h_{\mathcal{H}^{\text{PVC}}} \triangleq \{h' \in \mathcal{H}^{\text{PVC}} : \|\omega'-\omega\|_\infty \geq 1\}.$$

Since $\mathbb{Q}_h$ is Gaussian, with mean ω and variance $1/\varsigma$, using Lemma 4, with $a \triangleq \infty$ and $\epsilon \triangleq 1$, we have that

$$\begin{aligned}
&\mathbb{Q}_h\left\{h' \in \mathcal{B}^h_{\mathcal{H}^{\text{PVC}}}\right\} \\
&\quad= \mathbb{Q}_h\left\{\|\omega'-\omega\|_\infty \geq 1\right\} \\
&\quad\leq 2d\exp\left(-\frac{\varsigma}{2}\right) \\
&\quad= 2d\exp\left(-\frac{1}{2}\left(\frac{9n(\Delta_G+2)}{\gamma^2}\right)^2 \ln(mn)\right) \\
&\quad\leq 2d\exp\left(-\ln(mn)\right) \\
&\quad= \frac{2d}{mn}.
\end{aligned}$$

The second inequality uses the fact that $\gamma \leq \sqrt{n}$ and $9(\Delta_G+2) \geq \sqrt{2}$.

Since $\mathcal{H}^{\text{PVC}}$ is a subset of $\mathcal{H}^{\text{SC}}_{\mathcal{T}}$, Theorem 4 holds for every hypothesis in $\mathcal{H}^{\text{PVC}}$. Thus, using Equation 29, any $h' \in \mathcal{H}^{\text{PVC}}$ has β-uniform collective stability for

$$\beta \triangleq 2\sqrt{\frac{\|\mathbf{w}'\|_\infty}{\kappa'} C_G} = 2\sqrt{\|\omega'\|_\infty C_G}.$$

For any $h' \notin \mathcal{B}^h_{\mathcal{H}^{\text{PVC}}}$, using the triangle inequality, we have that

$$\begin{aligned}
\|\omega'\|_\infty &= \|\omega\|_\infty + \|\omega'\|_\infty - \|\omega\|_\infty \\
&\leq \|\omega\|_\infty + \|\omega'-\omega\|_\infty \\
&\leq \|\omega\|_\infty + 1.
\end{aligned}$$

Therefore, using Lemma 14, every $h' \notin \mathcal{B}^h_{\mathcal{H}^{\text{PVC}}}$ must have

$$\begin{aligned}
\beta &\leq 2\sqrt{(\|\omega\|_\infty+1)\, C_G} \\
&\leq 2\sqrt{(\|\omega\|_\infty+1)\,(\Delta_G+1)}.
\end{aligned}$$

Replacing ω with Equation 29 completes the proof. ■

The proof of Theorem 6 now follows directly from Theorem 2, using Lemma 13 for the admissibility constants, Lemma 19 for collective stability, and Lemma 18 for the KL divergence. With probability at least $1-\delta$ over realizations of $\hat{\mathbf{Z}}$,

$$\begin{aligned}
&\overline{L}^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h) - L^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) \leq \frac{2d}{mn} \\
&\quad+ \frac{2\,\|\boldsymbol{\Theta}^{\boldsymbol{\pi}}_n\|_\infty}{\sqrt{2mn}}\left(1+\frac{6}{\gamma}\sqrt{\left(\frac{\|\mathbf{w}\|_\infty}{\kappa}+1\right)(\Delta_G+1)}\right) \\
&\quad\times\sqrt{d\ln\left(\frac{9n(\Delta_G+2)}{\gamma^2}\sqrt{\ln(mn)}\right) + \frac{\|\mathbf{w}\|_2^2}{2\kappa^2} + \ln\frac{2}{\delta}}.
\end{aligned}$$

Further, using Lemma 17, we have that

$$\begin{aligned}
&\overline{L}^0(h) - L^\gamma(h, \hat{\mathbf{Z}}) \\
&\qquad\leq \overline{L}^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h) - L^{\frac{\gamma}{3},\frac{\gamma}{3}}(\mathbb{Q}_h, \hat{\mathbf{Z}}) + \frac{2}{\sqrt{mn}}.
\end{aligned}$$

Combining these inequalities completes the proof.

References

M. Donsker and S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 28(1):1–47, 1975.

D. Fiebig. Mixing properties of a class of Bernoulli processes. *Transactions of the American Mathematical Society*, 338:479–492, 1993.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

A. Kontorovich and K. Ramanan. Concentration inequalities for dependent random variables via the martingale method. *Annals of Probability*, 36(6): 2126–2158, 2008.

S. Kutin. Extensions to McDiarmid's inequality when differences are bounded with high probability. Technical report, University of Chicago, 2002.

D. McAllester. Simplified PAC-Bayesian margin bounds. In *Conference on Computational Learning Theory*, 2003.

C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, volume 141 of *London Mathematical Society Lecture Note Series*, pages 148–188. Cambridge University Press, 1989.

P.-M. Samson. Concentration of measure inequalities for Markov chains and ϕ-mixing processes. *Annals of Probability*, 28(1):416–461, 2000.