PAC-Bayes Generalization Bounds for Randomized Structured Prediction

Ben London
University of Maryland
blondon@cs.umd.edu

Bert Huang
University of Maryland
bert@cs.umd.edu

Ben Taskar
University of Washington
taskar@cs.washington.edu

Lise Getoor
University of Maryland
getoor@cs.umd.edu

Abstract

We present a new PAC-Bayes generalization bound for structured prediction that is applicable to perturbation-based probabilistic models. Our analysis explores the relationship between perturbation-based modeling and the PAC-Bayes framework, and connects to recently introduced generalization bounds for structured prediction. We obtain the first PAC-Bayes bounds that guarantee better generalization as the size of each structured example grows.

1 Introduction

Perturbation-based models represent a powerful new framework for structured probabilistic modeling where sampling is, by construction, efficient and exact. A perturbation-based model uses a distribution over a space of efficiently solvable optimizations. In some cases, they can be designed to explicitly mimic equivalent exponential-family Markov random fields (e.g., [5, 15]). In other cases, the class of distributions is distinct from standard probabilistic models (e.g., [18]). By defining their distributions around tractable optimization problems, perturbation-based models admit efficient, exact sampling procedures. These sampling procedures typically generate optimization parameters from simple distributions, such as Gaussian or Gumbel; the sampled parameters are then used to define an efficient optimization problem, such as Gaussian inference, graph cuts, or matching. These methods have been shown empirically to yield effective learning algorithms for novel structured prediction tasks [5, 18]. In this paper, we introduce new theory to characterize the generalization properties of learning a perturbation-based model, including new *PAC-Bayes* analysis for structured predictors that yields tighter bounds than previous analyses.

PAC-Bayes is a theoretical framework for analyzing the generalization error of Bayesian learning and randomized prediction. Perturbation-based sampling can be viewed as the *Gibbs classifier* in the PAC-Bayes paradigm. In PAC-Bayes, this Gibbs classifier performs a random draw of a predictor from a distribution over the hypothesis space. The PAC-Bayes hypothesis space corresponds to the parameter space in which perturbations are made during perturb-and-MAP. Hypothesis complexity in PAC-Bayes analysis is measured as the Kullback-Leibler (KL) divergence between a fixed prior distribution and a learned posterior. PAC-Bayes analysis was introduced by McAllester [12] and later refined by a number of authors [1, 9, 16], achieving some of the tightest known generalization bounds for both randomized and deterministic predictors.

We connect the PAC-Bayes paradigm to new generalization bounds for structured prediction [11]. The tightest known PAC-Bayes bounds for structured prediction [14] decrease proportionally to the number of training examples. Our new PAC-Bayes bound decreases with both the number of examples and the *size* of each example. Accordingly, provided the data distribution exhibits

suitably weak dependence within each structure, and the hypothesis class has certain properties—in particular, a form of predictive smoothness which we call *collective stability*—our bounds can be much tighter than previous bounds when training on a limited number of very large examples—even just one.

Our bounds suggest a class of parameter distributions that may guarantee generalization. Because the Gibbs classifier in PAC-Bayes analysis relates to the sampling process for perturbation-based models, these generalization bounds can be applied to a restricted class of randomized optimum models.

2 Preliminaries

We analyze generalization by relating it to the boundedness of the loss function, the smoothness of the structured prediction, and weak dependency of the data generating process. We introduce some notation and terminology useful for formalizing these concepts and applying them in our analysis. In the structured prediction framework we consider, each example contains n interdependent random variables, $\mathbf{Z} \triangleq \{Z_i\}_{i=1}^n \triangleq \{(X_i, Y_i)\}_{i=1}^n$, with joint distribution $\mathbb{P}$.[1] Each Z_i takes values in a sample space $\mathcal{Z} \triangleq \mathcal{X} \times \mathcal{Y}$.

We are interested in predicting $\mathbf{Y} \triangleq \{Y_i\}_{i=1}^n$, conditioned on $\mathbf{X} \triangleq \{X_i\}_{i=1}^n$. Let $\mathcal{H} \subseteq \{h : \mathcal{X}^n \to \hat{\mathcal{Y}}^n\}$ denote a class of hypotheses, where $\hat{\mathcal{Y}}$ is not necessarily the same as $\mathcal{Y}$. (For example, h could output a label *score* instead of a label.) Let $\mathbb{H}$ denote a predetermined prior distribution over $\mathcal{H}$, and let $\hat{\mathbb{H}}$ denote a posterior distribution, typically learned from training data. In the PAC-Bayes framework, prediction is stochastic. Given an input $\mathbf{X}$, we first draw a hypothesis $h \in \mathcal{H}$ according to $\hat{\mathbb{H}}$, then compute the prediction $\hat{\mathbf{Y}} = h(\mathbf{X})$.

To make the relationship between this PAC-Bayes setting and perturbation-based models more concrete, consider the following pairwise MRF.

$$p(\mathbf{Y} \mid \mathbf{X}) \triangleq \frac{1}{\Pi(\mathbf{X})} \exp\left(\sum_{i \in \mathcal{V}} \langle w_i, f_i(X_i, Y_i)\rangle + \sum_{\{i,j\} \in \mathcal{E}} \langle w_{i,j}, f_{i,j}(Y_i, Y_j)\rangle\right) \tag{1}$$

Given evidence $\mathbf{X} = \mathbf{x}$, the Gibbs classifier would first draw a random weight vector $\mathbf{w}$ according to a posterior distribution over the parameter space, then solve the desired inference problem (e.g., marginal or MAP). If the inference problem were efficiently solvable (such as a convex optimization), then this would equivalent to a single round of sampling in a randomized optimum model.

For a loss function ℓ and hypothesis h, denote the average loss on a set of m structured examples, $\hat{\mathbf{Z}} \triangleq \{\mathbf{Z}^{(l)}\}_{l=1}^m = \{\{Z_i^{(l)}\}_{i=1}^n\}_{l=1}^m$, by

$$L(h, \hat{\mathbf{Z}}) \triangleq \frac{1}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} \ell\left(Y_i^{(l)}, h_i(\mathbf{X}^{(l)})\right).$$

Let $\overline{L}(h) \triangleq \mathbb{E}[L(h, \mathbf{Z})]$ denote the expected average loss (also known as the *risk*) over realizations of a single example $\mathbf{Z}$, which corresponds to the error h will incur on future predictions. Since prediction is a stochastic process, we are also interested in the expectation of these measures over draws of h, which we denote by $L(\hat{\mathbb{H}}, \mathbf{Z}) \triangleq \mathbb{E}_{h \sim \hat{\mathbb{H}}}[L(h, \mathbf{Z})]$ and $\overline{L}(\hat{\mathbb{H}}) \triangleq \mathbb{E}_{h \sim \hat{\mathbb{H}}}[\overline{L}(h)]$.

We restrict our analysis to loss functions that satisfy certain admissibility conditions.

Definition 1. We say that a loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$ is (M, λ)-*admissible* if, for any $y, y' \in \mathcal{Y}$ and $\hat{y}, \hat{y}' \in \mathcal{Y}$, the following hold:

1. $|\ell(y, \hat{y}) - \ell(y', \hat{y})| \leq M$.
2. $|\ell(y, \hat{y}) - \ell(y, \hat{y}')| \leq \lambda \left\|\hat{y} - \hat{y}'\right\|_1$.

[1] We have assumed a one-to-one correspondence between input and output variables so as to minimize bookkeeping, but this assumption can be relaxed.

2.1 Collective Stability

A key component of our analysis is the *algorithmic stability* of joint inference. Stability ensures that small changes to the input result in bounded variation in the output. In learning theory, it has traditionally been used to quantify the variation in the output of a learning algorithm upon adding or removing training examples [2]. We apply this concept to an arbitrary class of vector-valued functions, $\mathcal{F} \triangleq \{\phi : \mathcal{Z}^n \to \mathbb{R}^N\}$, where N does not necessarily equal n. For the following, let $\text{dist}_{\text{H}}(\mathbf{z}, \mathbf{z}') \triangleq \sum_{i=1}^{n} \mathbb{1}\{z_i \neq z_i'\}$ denote the Hamming distance between two vectors $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$.

Definition 2. We say that a function $\phi \in \mathcal{F}$ has *uniform collective stability* β if, for any two inputs $\mathbf{z}, \mathbf{z}' \in \mathcal{Z}^n$ such that $\text{dist}_{\text{H}}(\mathbf{z}, \mathbf{z}') = 1$, $\|\phi(\mathbf{z}) - \phi(\mathbf{z}')\|_1 \leq \beta$. Similarly, we say that the class $\mathcal{F}$ has uniform collective stability β if every $\phi \in \mathcal{F}$ has uniform collective stability β.

The collective stability of a hypothesis extends to any admissible loss function.

Lemma 1 (London et al. [11])**.** *If a loss function ℓ is (M, λ)-admissible, and a hypothesis h has uniform collective stability β, then $\ell \circ h$ has uniform collective stability $(M + \lambda\beta)$.*

2.2 Statistical Tools

We now review some supporting definitions that are used in our generalization bounds. For a fixed permutation π of the variables, we use a *dependency matrix* $\mathbf{\Theta}_n^\pi$ to measure the dependence between variables. See London et al. [11] for the formal definition. Put simply, each upper-off-diagonal entry, $\theta_{i,j}^\pi : i < j$, measures the amount of influence that variable $Z_{\pi(i)}$ has on variables $Z_{\pi(j)}, \ldots, Z_{\pi(n)}$. Our bound quantifies the overall dependence using the standard matrix infinity norm, $\|\mathbf{\Theta}_n^\pi\|_\infty \triangleq \max_{i \in [n]} \sum_{j=1}^{n} \left|\theta_{i,j}^\pi\right|$. Note that, If the variables are independent, then $\mathbf{\Theta}_n^\pi$ is the identity matrix, and $\|\mathbf{\Theta}_n^\pi\|_\infty = 1$.

We do not assume that $\mathbf{Z}$ corresponds to a temporal process, which is why the ordering π has such a strong impact on $\|\mathbf{\Theta}_n^\pi\|_\infty$. In general, given a graph topology and an ordering of the vertices, $\|\mathbf{\Theta}_n^\pi\|_\infty$ measures the decay of dependence over graph distance. For instance, for Markov a tree process, Kontorovich [7] orders the variables via a breadth-first traversal from the root; for an Ising model on a lattice, Chazottes et al. [3] order the variables with a spiraling traversal from the origin. In both of these instances, under suitable contraction or temperature regimes, the authors show that $\|\mathbf{\Theta}_n^\pi\|_\infty$ is bounded independent of n (i.e., $\|\mathbf{\Theta}_n^\pi\|_\infty = \mathrm{O}(1)$). We posit that the same holds for any graph with bounded degree when the mixing coefficients exhibit geometric decay.

3 PAC-Bayes Bounds

Our PAC-Bayes proofs are based on a martingale technique due to Lever et al. [10] and Seldin et al. [17]. The so-called "one-sided" bounds we present, while not as tight as some "two-sided" bounds, are arguably more interpretable, and are easily obtained using martingale-based concentration inequalities. The proof is provided in the appendix.

Theorem 1. *Fix any $m \geq 1$, $n \geq 1$, $\delta \in (0, 1)$ and π. Let $\mathcal{H}$ be a hypothesis class with uniform collective stability β, and let ℓ be a (M, λ)-admissible loss function. Then, for any prior distribution $\mathbb{H}$ on $\mathcal{H}$, with probability at least $1 - \delta$ over realizations of m examples, $\hat{\mathbf{Z}} \triangleq \{\{Z_i^{(l)}\}_{i=1}^n\}_{l=1}^m$, the following holds simultaneously for all posteriors $\hat{\mathbb{H}}$ on $\mathcal{H}$:*

$$\overline{L}(\hat{\mathbb{H}}) \leq L(\hat{\mathbb{H}}, \hat{\mathbf{Z}}) + \|\mathbf{\Theta}_n^\pi\|_\infty (M + \lambda\beta) \sqrt{\frac{2\,\mathrm{KL}(\hat{\mathbb{H}} \| \mathbb{H}) + 2 \ln \frac{2}{\delta}}{mn}}. \tag{2}$$

Note that, if $\|\mathbf{\Theta}_n^\pi\|_\infty$ and β do not grow with n, and $\mathrm{KL}(\hat{\mathbb{H}} \| \mathbb{H})$ is sublogarithmic in m and n, then Equation 2 decreases with both m and n. This makes it potentially tighter than existing bounds when each structured example is large and the number of examples is small. Even for $m = 1$, Equation 2 goes to zero as n increases, meaning one can generalize from a single, large example.

It is also worth noting that, unlike some previous PAC-Bayes bounds for structured prediction [6, 14], our bounds do not have a $\ln m$ or $\ln n$ term in the numerator, though this may be added when bounding the KL divergence term.

4 Application to Perturbation Methods

We can directly apply Theorem 1 to a specific class of randomized optimum models. To obtain nontrivial risk bounds, one needs to show that (1) $\mathrm{KL}(\hat{\mathbb{H}}\|\mathbb{H}) = \mathrm{O}(\ln(mn))$, and (2) $\beta = \mathrm{O}(1)$. The first precondition is satisfied by certain constructions of the prior and posterior; we refer the reader to Langford and Shawe-Taylor [9] and McAllester [13] for examples, and posit that these techniques are easily extended to perturbation-based methods. London et al. [11] showed that the second condition, uniform collective stability, is satisfied by a broad class of structured predictors; in particular, if the model is *templated* (that is, uses *parameter tying*), the feature and weight norms are uniformly bounded, and the (log-linear) inference objective (sometimes called the *energy function*) is *strongly* convex.

To make this concrete, recall the pairwise MRF given in Equation 1. Suppose we replace the per-clique weights, w_i and $w_{i,j}$, with weights w_s for singletons and w_p for pairs. We could then express the conditional distribution as

$$p(\mathbf{Y} \mid \mathbf{X}) \triangleq \frac{1}{\Pi(\mathbf{X})} \exp\left(\langle w_s, \mathbf{f}_s(\mathbf{X}, \mathbf{Y})\rangle + \langle w_p, \mathbf{f}_p(\mathbf{Y})\rangle\right),$$

where $f_s(\mathbf{X}, \mathbf{Y}) \triangleq \sum_{i \in \mathcal{V}} f_i(X_i, Y_i)$ and $\mathbf{f}_p(\mathbf{Y}) \triangleq \sum_{\{i,j\} \in \mathcal{E}} f_{i,j}(Y_i, Y_j)$. To perform approximate marginal inference in this graphical model, one solves the optimization

$$\arg\max_{\mu \in \mathcal{M}} \langle w_s, \mu_s\rangle + \langle w_p, \mu_p\rangle + \Psi(\mu),$$

where $\mathcal{M}$ is the relaxed marginal polytope and Ψ is a κ-strongly convex surrogate for the negative entropy (e.g., the convexified Bethe approximation [19]). London et al. [11] showed that this model has uniform collective stability $(2\sqrt{R(\Delta+1)/\kappa})$, where R is a uniform upper bound on $\|(w_s, w_p)\|_\infty$, and Δ is the maximum degree of the graph (which is assumed to be independent of n).

In the perturbation framework, a randomized optimum model would first draw a random (w_s, w_p), according to a posterior $\hat{\mathbb{H}}$, then solve the resulting convex optimization of maximizing $E_{\mathbf{w}}$. Learning in this setting involves learning the parameters of $\hat{\mathbb{H}}$ (e.g., mean and covariance for the weights).

The parameter tying differs from the traditional perturbation framework, in which one typically samples the weights for each clique independently. Templating may affect the statistical properties of some perturbation-based methods, which might result in larger sampling requirements. This tradeoff between computational complexity and model complexity (i.e., generalization) in the perturbation framework is an interesting direction for future research.

Requiring the weight norms to be bounded is a simple modification to standard perturbation-based models, but it precludes the possibility of using the full Gaussian or Gumbel distributions that yield exact mapping between MRFs and perturb-and-MAP models. Yet the short tail of these distributions means that the weight norms—thus, the collective stability—are almost-surely bounded. In recently submitted work, we provide extensions to our theory, and applications thereof, that accommodate distributions over unbounded parameter spaces.

The strong convexity condition on inference also differs from current instantiations of perturbation-based models. We are also extending our theory to accommodate linear optimization objectives, which encompass many of the optimization problems of interest for perturbation models.

Acknowledgments

This work was partially supported by NSF CAREER grants 0746930 and 1054215, NSF grant IIS1218488, and IARPA via DoI/NBC contract number D12PC00337. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, IARPA, DoI/NBC, or the U.S. Government.

References

[1] A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. In *Advances in Neural Information Processing Systems 19*, 2006.

[2] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

[3] J. Chazottes, P. Collet, C. Külske, and F. Redig. Concentration inequalities for random fields via coupling. *Probability Theory and Related Fields*, 137:201–225, 2007.

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

[5] T. Hazan and T. Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[6] J. Keshet, D. McAllester, and T. Hazan. PAC-Bayesian approach for minimization of phoneme error rate. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2224–2227, 2011.

[7] L. Kontorovich. *Measure Concentration of Strongly Mixing Processes with Applications*. PhD thesis, Carnegie Mellon University, 2007.

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

[9] J. Langford and J. Shawe-Taylor. PAC-Bayes and margins. In *Advances in Neural Information Processing Systems 15*, 2002.

[10] G. Lever, F. Laviolette, and J. Shawe-Taylor. Distribution-dependent PAC-Bayes priors. In *Proceedings of the 21st International Conference on Algorithmic Learning Theory*, 2010.

[11] Ben London, Bert Huang, Benjamin Taskar, and Lise Getoor. Collective stability in structured prediction: Generalization from one example. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[12] D. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory*, pages 164–170, 1999.

[13] D. McAllester. Simplified PAC-Bayesian margin bounds. In *COLT*, pages 203–215, 2003.

[14] D. McAllester. Generalization bounds and consistency for structured labeling. In G. Bakir, T. Hofmann, B. Schölkopf, A. Smola, B. Taskar, and S. Vishwanathan, editors, *Predicting Structured Data*. MIT Press, 2007.

[15] G. Papandreou and A. Yuille. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *Proceedings of IEEE International Conference on Computer Vision*, pages 193–200, 2011.

[16] M. Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

[17] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58(12):7086–7093, 2012.

[18] D. Tarlow, R. P. Adams, and R. S. Zemel. Randomized optimum models for structured prediction. In *Proceedings of the 15th Conference on Artificial Intelligence and Statistics*, pages 21–23, 2012.

[19] M. Wainwright. Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research*, 7:1829–1859, 2006.

A Proof of Theorem 1

Our proof requires the following concentration inequality, which is an adaptation of Kontorovich and Ramanan [8, Theorem 1.1].

Theorem 2. *Let $\mathbf{Z} \triangleq \{Z_i\}_{i=1}^n$ be a set of random variables with joint distribution $\mathbb{P}$. Let $\phi : \mathcal{Z}^n \to \mathbb{R}$ be a measurable function that is (β/n)-Lipschitz w.r.t. the Hamming distance. Then, for any $\tau \in \mathbb{R}$, $\epsilon > 0$ and permutation π, with $\mathbf{\Theta}_n^\pi$ as defined in London et al. [11],*

$$\mathbb{P}\left\{e^{\tau(\phi(\mathbf{Z})-\mathbb{E}[\phi(\mathbf{Z})])} \geq \epsilon\right\} \leq \frac{1}{\epsilon}\exp\left(\frac{\tau^2\beta^2\left\|\mathbf{\Theta}_n^\pi\right\|_\infty^2}{8n}\right). \tag{3}$$

We now prove Theorem 1. Start by defining a function $\phi(h, \hat{\mathbf{Z}}) \triangleq \overline{L}(h) - L(h, \hat{\mathbf{Z}})$, and a free parameter $u \in \mathbb{R}$. Using the *change of measure* inequality, due to Donsker and Varadhan [4], we have for any prior and posterior distributions $\mathbb{H}, \hat{\mathbb{H}}$ over $\mathcal{H}$,

$$\overline{L}(\hat{\mathbb{H}}) - L(\hat{\mathbb{H}}, \hat{\mathbf{Z}}) = \frac{1}{u}\mathop{\mathbb{E}}_{h\sim\hat{\mathbb{H}}}[u\,\phi(h, \hat{\mathbf{Z}})] \leq \frac{1}{u}\left(\mathrm{KL}(\hat{\mathbb{H}}\|\mathbb{H}) + \ln \mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[e^{u\,\phi(h,\hat{\mathbf{Z}})}\right]\right). \tag{4}$$

The remainder of the proof focuses on upper-bounding $e^{u\,\phi(h,\hat{\mathbf{Z}})}$ and optimizing u. Since u may be a function of the posterior, we can't optimize u for *all* posteriors simultaneously. We therefore adopt a technique due to Seldin et al. [17] in which we discretize the space of u, then apply the union bound. This *approximately* optimizes the bound for all posteriors simultaneously.

Let $\beta_{\ell\circ\mathcal{H}} \triangleq M + \lambda\beta$. By Lemma 1, $\beta_{\ell\circ\mathcal{H}}$ is a uniform upper bound on the uniform collective stability of $\ell \circ h$, for any $h \in \mathcal{H}$, since $\mathcal{H}$ has uniform collective stability β. It is then easy to show that $\phi(h, \cdot)$ is $(\beta_{\ell\circ\mathcal{H}}/(mn))$-Lipschitz w.r.t. the Hamming distance, for all $h \in \mathcal{H}$.

Define an infinite sequence of parameters, $u_0, u_1, \ldots$, where

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

Let $\delta_j \triangleq \delta 2^{-(j+1)}$ and define the event

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

By Theorem 2, we have that E_j happens with probability less than δ_j; therefore, by the union bound, with probability at least $1 - \sum_{j=0}^\infty \delta_j = 1 - \delta$, *every* u_j satisfies

$$\mathop{\mathbb{E}}_{h\sim\mathbb{H}}\left[e^{u_j\phi(h,\hat{\mathbf{Z}})}\right] \leq \frac{1}{\delta_j}\exp\left(\frac{u_j^2\beta_{\ell\circ\mathcal{H}}^2\left\|\mathbf{\Theta}_n^\pi\right\|_\infty^2}{8mn}\right). \tag{5}$$

Here we have used the fact that $\left\|\mathbf{\Theta}_{mn}^\pi\right\|_\infty = \left\|\mathbf{\Theta}_n^\pi\right\|_\infty$ because $\mathbf{\Theta}_{mn}^\pi$ is block diagonal, with each sub-matrix equal to $\mathbf{\Theta}_n^\pi$. Combining Equations 4 and 5, we now have, with probability at least $1-\delta$, every u_j satisfies

$$\overline{L}(\hat{\mathbb{H}}) - L(\hat{\mathbb{H}}, \hat{\mathbf{Z}}) \leq \frac{\mathrm{KL}(\hat{\mathbb{H}}\|\mathbb{H}) + \ln\frac{1}{\delta_j}}{u_j} + \frac{u_j\beta_{\ell\circ\mathcal{H}}^2\left\|\mathbf{\Theta}_n^\pi\right\|_\infty^2}{8mn}. \tag{6}$$

Now, for any particular posterior $\hat{\mathbb{H}}$, there exists an approximately-optimal value $u_{j^\star}$ by taking

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

Since, for all $v \in \mathbb{R}$, $v - 1 \leq \lfloor v \rfloor \leq v$, one can easily show that

$$\frac{1}{2}\sqrt{\frac{8mn\left(\mathrm{KL}(\hat{\mathbb{H}}\|\mathbb{H}) + \ln\frac{2}{\delta}\right)}{\beta_{\ell\circ\mathcal{H}}^2\left\|\mathbf{\Theta}_n^\pi\right\|_\infty^2}} \leq u_{j^\star} \leq \sqrt{\frac{8mn\left(\mathrm{KL}(\hat{\mathbb{H}}\|\mathbb{H}) + \ln\frac{2}{\delta}\right)}{\beta_{\ell\circ\mathcal{H}}^2\left\|\mathbf{\Theta}_n^\pi\right\|_\infty^2}}. \tag{7}$$

One can further show that

$$\ln \frac{1}{\delta_{j^\star}} \leq \ln \frac{2}{\delta} + \frac{\ln 2}{2 \ln 2} \ln \left(\frac{\mathrm{KL}(\hat{\mathbb{H}} \| \mathbb{H})}{\ln(2/\delta)} + 1 \right) \leq \ln \frac{2}{\delta} + \frac{1}{2} \left(\mathrm{KL}(\hat{\mathbb{H}} \| \mathbb{H}) + \ln \frac{2}{\delta} \right) \tag{8}$$

for all $\delta \in (0, 1)$. Combining Equation 6 with the lower and upper bounds from Equations 7 and 8, we then have that, with probability at least $1 - \delta$, $u_{j^\star}$ satisfies

$$\overline{L}(\hat{\mathbb{H}}) - L(\hat{\mathbb{H}}, \hat{\mathbf{Z}}) \leq \beta_{\ell \circ \mathcal{H}} \left\| \boldsymbol{\Theta}_n^\pi \right\|_\infty \sqrt{\frac{2 \, \mathrm{KL}(\hat{\mathbb{H}} \| \mathbb{H}) + 2 \ln \frac{2}{\delta}}{mn}}.$$

Substituting the definition of $\beta_{\ell \circ \mathcal{H}}$ completes the proof.