Gibbs-Based Information Criteria and the Over-Parameterized Regime

Haobo Chen
University of Florida
haobo.chen@ufl.edu

Yuheng Bu
University of Florida
buyuheng@ufl.edu

Gregory W. Wornell
Massachusetts Institute of Technology
gww@mit.edu

Abstract

Double-descent refers to the unexpected drop in test loss of a learning algorithm beyond an interpolating threshold with over-parameterization, which is not predicted by information criteria in their classical forms due to the limitations in the standard asymptotic approach. We update these analyses using the information risk minimization framework and provide Bayesian Information Criterion (BIC) for models trained by the Gibbs algorithm. Notably, the BIC penalty term for the Gibbs algorithm corresponds to a specific information measure, i.e., KL divergence. We extend this information-theoretic analysis to over-parameterized models by characterizing the Gibbs-based BIC for the random feature model in the regime where the number of parameters p and the number of samples n tend to infinity, with p/n fixed. Our experiments demonstrate that the Gibbs-based BIC can select the high-dimensional model and reveal the mismatch between marginal likelihood and population risk in the over-parameterized regime, providing new insights for understanding the double-descent phenomenon.

1 Introduction

The classical understanding of model selection is that more complex models can capture more complex patterns but tend to overfit and have large generalization error [1]. This tradeoff results in a $\cup$-shaped curve characterized by the classical model selection criterion when test loss is plotted against model complexity. As a result, the models that minimize test loss tend to have moderate complexity. Recently, the success of deep learning challenges this classical picture since neural networks are often extremely complex (e.g., able to fit random labels [2]) while *also* generalizing well to yield low test error on unseen samples.

An emerging explanation of this behavior is *double-descent* [3], which posits that: 1) The classical $\cup$-shaped curve is only valid when the number of model parameters p is smaller than the number of samples n. 2) In the over-parameterized regime where p is significantly larger than n, and models are complex enough to fit training data perfectly, test loss can decrease with increased model complexity.

To better understand the double-descent phenomenon, we revisit the classical information criteria and discern that the penalty term in Akaike Information Criterion (AIC) can be interpreted as the generalization error, while Bayesian Information Criterion (BIC) approximates the marginal likelihood using the empirical risk minimization solution. We update the classical analyses of BIC using the information risk minimization framework proposed in [4] with the *Gibbs algorithm*. We make the following contributions in this paper:

1. We provide an information-theoretic analysis for the marginal likelihood of the model learned by the Gibbs algorithm, resulting in Gibbs-based BIC (equation (5)) with KL divergence as the penalty term.

Mathematics of Modern Machine Learning Workshop at NeurIPS 2023.

2. We generalize our information-theoretic analysis to over-parameterized random feature (RF) models, which results in an over-parameterized Gibbs-based BIC (equation (9)) that favors over-parameterized RF model, while classical information criteria cannot.
3. We empirically demonstrate the mismatch between marginal likelihood (BIC) and generalization error (AIC) in the over-parameterized setting, where AIC exhibits double-descent but BIC does not. Such a phenomenon is highly affected by the choice of prior distributions.

2 Preliminaries

Let $S = \{Z_i\}_{i=1}^n$ be the training set, where each random sample $Z_i = \{(X_i, Y_i)\} \in \mathcal{Z}$ are independent and identically distributed (i.i.d.) from the same data-generating distribution P_Z. We denote the parameter of a machine learning model by $w \in \mathcal{W}$, where $\mathcal{W}$ is the parameter space. The performance of the model is measured by a loss function $\ell \colon \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$, and the log loss $\ell_{\log}(w, \boldsymbol{z}) \triangleq -\log P(y|\boldsymbol{x}; w)$ associated with a parametric probabilistic model $P(y|\boldsymbol{x}; w)$ is of particular interest to us. We can define the empirical risk and the population risk for a given w as $L_E(w, \boldsymbol{z}^n) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(w, \boldsymbol{z}_i)$, and $L_P(w, P_Z) \triangleq \mathbb{E}_{P_Z}\big[\ell(w, Z)\big]$, respectively. A learning algorithm can be modeled as a randomized mapping from the training set S onto a model parameter $\hat{W} \in \mathcal{W}$ according to the conditional distribution $P_{\hat{W}|S}$. The expected generalization error quantifying the degree of over-fitting can be expressed in the form

$$\overline{gen}(P_{\hat{W}|S}, P_S) \triangleq \mathbb{E}_{P_{\hat{W},S}}\big[L_P(\hat{W}, P_Z) - L_E(\hat{W}, S)\big], \tag{1}$$

where the expectation is taken over the joint distribution $P_{\hat{W},S} = P_{\hat{W}|S} \otimes P_S$.

3 Gibbs-based Information Criteria

3.1 Information Risk Minimization and Gibbs Algorithm

As detailed in Appendix B, classical AIC and BIC depend on maximum likelihood estimate (MLE), which can be viewed as empirical risk minimization. Instead, we consider the Gibbs algorithm, which minimizes both empirical risk and an information-theoretic generalization error bound.

Lemma 1 ([5])**.** *Suppose the loss function $\ell(w, z) \in [0, 1]$ is bounded, and $S = \{Z_i\}_{i=1}^n$ contains n i.i.d. training samples, then $|\overline{gen}(P_{\hat{W}|S}, P_S)| \leq \sqrt{I(\hat{W}; S)/(2n)}$.*

This upper bound motivates the following information risk minimization (IRM) problem [4, 5, 6]

$$P^*_{\hat{W}|S} = \underset{P_{\hat{W}|S}}{\operatorname{argmin}}\ \mathbb{E}_{P_{\hat{W},S}}\big[L_E(\hat{W}, S)\big] + \frac{1}{\beta} D(P_{\hat{W}|S} \| \pi | P_S), \tag{2}$$

where $\beta > 0$ controls the regularization term and balances empirical risk and generalization error. Note that instead of regularizing $I(\hat{W}; S)$ which requires the knowledge of $P_{\hat{W}}$, IRM replaces it with an upper bound $D(P_{\hat{W}|S} \| \pi | P_S) \geq I(\hat{W}; S)$, where π is an arbitrary prior distribution over $\mathcal{W}$. The following lemma characterizes the solution to the IRM problem.

Lemma 2 ([4, 6])**.** *The minimum value of the following information risk minimization problem is*

$$\min_{P_{\hat{W}|S}} \mathbb{E}_{P_{\hat{W},S}}\big[L_E(\hat{W}, S)\big] + \frac{1}{\beta} D(P_{\hat{W}|S} \| \pi | P_S) = -\frac{1}{\beta}\mathbb{E}_{P_S}\big[\log \mathbb{E}_{\pi}[e^{-\beta L_E(W,S)}]\big], \tag{3}$$

which is achieved by the Gibbs algorithm (distribution) $P^*_{\hat{W}|S}(w|s) = \frac{\pi(w) e^{-\beta L_E(w,s)}}{\mathbb{E}_{\pi}\left[e^{-\beta L_E(W,s)}\right]}$, for $\beta > 0$.

3.2 Gibbs-based BIC

The Gibbs-based BIC is constructed by computing the marginal likelihood $m(\boldsymbol{z}^n)$ using the IRM framework. As such, it differs from the standard approach in classical (MLE-based) BIC, as no Laplace approximation is needed.

Proposition 1. *(proved in Appendix C) For the Gibbs algorithm $P^*_{\hat{W}|S}$, if we adopt the log-loss function $\ell(w, \boldsymbol{z}) = -\log P(y|\boldsymbol{x}; w)$, and set $\beta = n$, the marginal likelihood is*

$$-\frac{1}{n}\log m(\boldsymbol{z}^n) = \mathbb{E}_{P^*_{\hat{W}|S=\boldsymbol{z}^n}}\big[L_E(\hat{W}, \boldsymbol{z}^n)\big] + \frac{1}{n} D(P^*_{\hat{W}|S=\boldsymbol{z}^n} \| \pi). \tag{4}$$

Motivated by Proposition 1, we propose the Gibbs-based BIC to approximate the marginal likelihood:

$$\mathrm{BIC}^{+} \triangleq L_E(\hat{W}^*, \boldsymbol{z}^n) + \frac{1}{n} D(P^*_{\hat{W}|S=\boldsymbol{z}^n} \| \pi). \tag{5}$$

It can be verified that the Gibbs-based BIC has the same BIC penalty term as the classical regime. Due to space limitations, the discussion of Gibbs-based AIC is provided in Appendix D.

4 BIC for Over-Parameterized RF Model

The classical AIC and BIC fail to capture double-descent behavior. This is because the generalization error and marginal likelihood have different asymptotic behaviors in the over-parameterized regime. The classical analyses heavily rely on the asymptotic normality of MLE and Laplace approximation under certain regularization assumptions, which ignores the prior distribution as $n \to \infty$. Unfortunately, none of these properties hold in the over-parameterized regime. However, the Gibbs-based BIC^{+} in (5) defined using KL divergence can be generalized to over-parameterized models, as Proposition 1 holds regardless of the values of p and n. So, we can provide analysis of BIC^{+} to approximate the marginal likelihood in the over-parameterized regime. Here, we refine the Gibbs-based BIC^{+} analysis to this regime for the random feature (RF) model.

4.1 Random Feature Model

The RF model [7] takes the form of a two-layer neural network with fixed random weights in the first layer. Specifically, the output of RF model with input data $\boldsymbol{x} \in \mathbb{R}^d$ is

$$g(\boldsymbol{x}) \triangleq \sum_{j=1}^{p} f\Big(\frac{\langle \boldsymbol{x}, \boldsymbol{F}_j \rangle}{\sqrt{d}}\Big) \boldsymbol{w}_j = f\Big(\frac{\boldsymbol{x}^\top \boldsymbol{F}}{\sqrt{d}}\Big) \boldsymbol{w}, \tag{6}$$

where $\boldsymbol{w} \in \mathbb{R}^p$ denotes the weights of the model. Moreover, $\boldsymbol{F}_j \in \mathbb{R}^d$ denotes the jth random feature vector, which is the jth column of the random feature matrix $\boldsymbol{F} \in \mathbb{R}^{d \times p}$ whose entries are drawn i.i.d. from $\mathcal{N}(0,1)$. Finally, $f(\cdot)$ is a point-wise activation function. In our setting, there are n training samples $\boldsymbol{z}^n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and the $\boldsymbol{x}_i$ are drawn i.i.d. from $\mathcal{N}(0, \boldsymbol{I}_d)$.

We adopt a Gaussian prior distribution $\boldsymbol{w} \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} \boldsymbol{I}_p)$, and the weights of the RF model $\boldsymbol{w}$ can be obtained by the Gibbs algorithm using the regularized log-loss

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2\sigma^2} \|\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{w}\|_2^2 + \frac{n}{2} \log(2\pi\sigma^2) + \frac{\lambda n \|\boldsymbol{w}\|_2^2}{2\sigma^2}, \quad \text{where } \boldsymbol{B} \triangleq f(\boldsymbol{X}\boldsymbol{F}/\sqrt{d}) \in \mathbb{R}^{n \times p}, \tag{7}$$

and we collect the training data in a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and a vector $\boldsymbol{Y} \in \mathbb{R}^n$ to simplify the notation. As discussed in [8, 9], a significant benefit of using the random feature model is that the dimensionality of the input data d is not entangled with the number of parameters p.

4.2 Gibbs-based BIC for Over-Parameterized RF Model

To generalize BIC^{+} to the over-parameterized RF model, it suffices to focus on the second KL-divergence term in 5. In the random feature model, it can be shown (see Appendix E for details) that the Gibbs algorithm reduces to the Gaussian posterior distribution $P^*_{\hat{W}|S} \sim \mathcal{N}(\hat{W}_\lambda, \boldsymbol{\Sigma}_w)$, with $\hat{W}_\lambda = (\lambda n \boldsymbol{I}_p + \boldsymbol{B}^\top \boldsymbol{B})^{-1} \boldsymbol{B}^\top \boldsymbol{Y}$, and $\boldsymbol{\Sigma}_w = \sigma^2 (\lambda n \boldsymbol{I}_p + \boldsymbol{B}^\top \boldsymbol{B})^{-1}$.

Thus, the KL-divergence between the Gibbs posterior distribution and prior $\mathcal{N}(0, \frac{\sigma^2}{\lambda n} \boldsymbol{I}_p)$ is given by

$$D(P^*_{\hat{W}|S=\boldsymbol{z}^n} \| \pi) = \frac{1}{2} \Big[\frac{\lambda n}{\sigma^2} \|\hat{W}_\lambda\|_2^2 + \log \frac{\det(\frac{\sigma^2}{\lambda n} \boldsymbol{I}_p)}{\det(\boldsymbol{\Sigma}_w)} + \mathrm{tr}(\frac{\lambda n}{\sigma^2} \boldsymbol{\Sigma}_w) - p \Big]. \tag{8}$$

To obtain a convenient expression for the determinant and trace terms, we first impose restrictions on the activation function $f(\cdot)$. Therefore, these two terms can be characterized using random matrix theory by studying the eigenvalues of $\boldsymbol{\Sigma} \triangleq \boldsymbol{B}^\top \boldsymbol{B}/(\lambda n) + \boldsymbol{I}_p$ in the over-parameterized regime.

Theorem 1 (full statement and proofs can be found Appendix F) motivates us to define the following Gibbs-based BIC for the over-parameterized RF model to approximate the marginal likelihood,

$$\mathrm{BIC}^{+} \triangleq L_E(\hat{W}^*, z^n) + \frac{\lambda}{2\sigma^2} \|\hat{W}_\lambda\|_2^2 - \frac{\lambda}{8} \mathcal{F}(\frac{1}{\lambda}, r) + \frac{1}{2} V(1/\lambda, r). \tag{9}$$

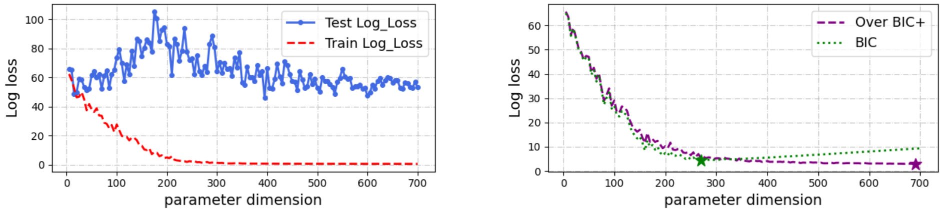

Figure 1: A comparison of the test and training log loss achieved by Gibbs with varying p when $\lambda = 0.005$ (**left**). A comparison of over-parameterized BIC^+ and classical BIC with varying p (**right**). The preferred models selected by different information criteria are marked using stars.

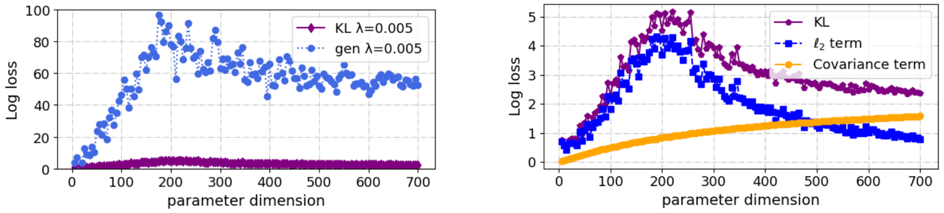

Figure 2: A comparison between the KL-divergence term in BIC^+ and the generalization error term (**left**). A decomposition of the terms in over-parameterized BIC^+ in (9) with $\lambda = 0.005$ (**right**).

The penalty term consists of the ℓ_2 norm of the learned weights, and two other terms, $\mathcal{F}$ and V (see Appendix F for detailed definition), capture the log determinant and trace in the over-parameterized regime, which will be referred to as the covariance term altogether in the next section.

5 Experiments

In the over-parameterized regime, we evaluate the Gibbs-based BIC^+ using $n = 200$ samples generated by the linear model

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{w}^* + \epsilon_i, \quad \boldsymbol{w}^* \in \mathbb{R}^p, \quad \|\boldsymbol{w}^*\|_2^2 = 1, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{10}$$

with true model $p = 400$ and noise $\sigma^2 = 0.1$.

As depicted in Figure 1 (left), the peak of test Log-Loss is located at the interpolation threshold, i.e., when $p = n = 200$, resulting in the highest generalization error. As p continues to increase, the test error begins to decline again, even falling below the levels observed in the under-parameterized regime $p < n$. In the right panel of Figure 1, it is evident that the classic BIC fails to select the true model, whereas our over-parameterized BIC^+ in (9) succeeds. Note that marginal likelihood does not exhibit double-descent behavior, and a similar mismatch between marginal likelihood and generalization for ERM has been observed in [10].

We further investigate the inconsistency between the marginal likelihood and population risk for the Gibbs algorithm by plotting the KL divergence and the generalization error in Figure 2 (left). Unlike the classical BIC, where the penalty term $(p/(2n)) \log n$ is order-wise larger than the p/n term in classical AIC, it can be seen that the KL divergence term in the over-parameterized BIC^+ can be smaller than the generalization error, depending on the value of λ and p. Thus, the mismatch between marginal likelihood (BIC) and population risk (AIC) is even more complicated in the over-parameterized setting due to the influence of prior distribution.

In Figure 2 (right), we decompose the penalty term of the over-parameterized BIC^+ in (9) into ℓ_2 term, covariance term. When $p \leq n$, the model prior can be ignored, and the training loss becomes the dominant factor of BIC^+. In this case, the KL divergence and the covariance term increase with p, corresponding to the classical BIC. When $p \geq n$, the double descent behavior of the ℓ_2 norm term dominates the over-parameterized BIC^+. In this regime, multiple weights exist that can fit the training data perfectly. We observe that the ℓ_2 norm of the Gibbs solution decreases as p increases. Note that similar phenomena are observed for the model learned by SGD, and generalization error bounds using different weights norms are established in [11, 12], which shows the profound connection of the double descent between generalization error (AIC) and the marginal likelihood (BIC).

References

[1] S. Geman, E. Bienenstock, and R. Doursat, “Neural networks and the bias/variance dilemma,” *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[2] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, “Understanding deep learning requires rethinking generalization,” in *Proc. International Conference on Learning Representations (ICLR)*, 2017.

[3] M. Belkin, D. Hsu, S. Ma, and S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off,” *Proceedings of the National Academy of Sciences*, vol. 116, no. 32, pp. 15849–15854, 2019.

[4] T. Zhang, “Information-theoretic upper and lower bounds for statistical estimation,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1307–1321, 2006.

[5] A. Xu and M. Raginsky, “Information-theoretic analysis of generalization capability of learning algorithms,” in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2524–2533, 2017.

[6] S. M. Perlaza, G. Bisson, I. Esnaola, A. Jean-Marie, and S. Rini, “Empirical risk minimization with relative entropy regularization: Optimality and sensitivity analysis,” in *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 684–689, IEEE, 2022.

[7] A. Rahimi and B. Recht, “Random features for large-scale kernel machines,” in *Advances in neural information processing systems*, pp. 1177–1184, 2008.

[8] S. Mei and A. Montanari, “The generalization error of random features regression: Precise asymptotics and the double descent curve,” *Communications on Pure and Applied Mathematics*, vol. 75, no. 4, pp. 667–766, 2022.

[9] S. d’Ascoli, L. Sagun, and G. Biroli, “Triple descent and the two kinds of overfitting: Where & why do they appear?,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 3058–3069, 2020.

[10] S. Lotfi, P. Izmailov, G. Benton, M. Goldblum, and A. G. Wilson, “Bayesian model selection, the marginal likelihood, and generalization,” in *International Conference on Machine Learning*, pp. 14223–14247, PMLR, 2022.

[11] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, “Exploring generalization in deep learning,” *Advances in neural information processing systems*, vol. 30, 2017.

[12] P. L. Bartlett, D. J. Foster, and M. J. Telgarsky, “Spectrally-normalized margin bounds for neural networks,” *Advances in neural information processing systems*, vol. 30, 2017.

[13] J. Chen and Z. Chen, “Extended Bayesian information criteria for model selection with large model spaces,” *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[14] Y. Fan and C. Y. Tang, “Tuning parameter selection in high dimensional penalized likelihood,” *Journal of the Royal Statistical Society: SERIES B: Statistical Methodology*, pp. 531–552, 2013.

[15] M. S. Advani, A. M. Saxe, and H. Sompolinsky, “High-dimensional dynamics of generalization error in neural networks,” *Neural Networks*, 2020.

[16] M. Geiger, S. Spigler, S. d’Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli, and M. Wyart, “Jamming transition as a paradigm to understand the loss landscape of deep neural networks,” *Physical Review E*, vol. 100, no. 1, p. 012115, 2019.

[17] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, “Deep double descent: Where bigger models and more data hurt,” in *International Conference on Learning Representations*, 2019.

[18] M. Belkin, D. Hsu, and J. Xu, “Two models of double descent for weak features,” *SIAM Journal on Mathematics of Data Science*, vol. 2, pp. 1167–1180, jan 2020.

[19] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, “Surprises in high-dimensional ridgeless least squares interpolation,” *The Annals of Statistics*, vol. 50, no. 2, pp. 949–986, 2022.

[20] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, “Benign overfitting in linear regression,” *Proceedings of the National Academy of Sciences*, 2020.

[21] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, “Harmless interpolation of noisy data in regression,” *IEEE Journal on Selected Areas in Information Theory*, 2020.

[22] Z. Deng, A. Kammoun, and C. Thrampoulidis, "A model of double descent for high-dimensional binary linear classification," *Information and Inference: A Journal of the IMA*, vol. 11, no. 2, pp. 435–495, 2022.

[23] G. R. Kini and C. Thrampoulidis, "Analytic study of double descent in binary classification: The impact of loss," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2527–2532, IEEE, 2020.

[24] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, "Generalisation error in learning with random features and the hidden manifold model," in *International Conference on Machine Learning*, pp. 3452–3462, PMLR, 2020.

[25] S. d'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala, "Double trouble in double descent: Bias and variance (s) in the lazy regime," in *International Conference on Machine Learning*, pp. 2280–2290, PMLR, 2020.

[26] E. Weinan, C. Ma, and L. Wu, "A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics," *Science China Mathematics*, pp. 1–24, 2020.

[27] F. Liu, J. Suykens, and V. Cevher, "On the double descent of random features models trained with SGD," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34966–34980, 2022.

[28] Z. Yang, Y. Yu, C. You, J. Steinhardt, and Y. Ma, "Rethinking bias-variance trade-off for generalization of neural networks," in *International Conference on Machine Learning*, pp. 10767–10777, PMLR, 2020.

[29] R. Dwivedi, C. Singh, B. Yu, and M. J. Wainwright, "Revisiting complexity and the bias-variance tradeoff," *arXiv preprint arXiv:2006.10189*, 2020.

[30] B. Adlam and J. Pennington, "Understanding double descent requires a fine-grained bias-variance decomposition," *Advances in neural information processing systems*, vol. 33, pp. 11022–11032, 2020.

[31] L. Lin and E. Dobriban, "What causes the test error? going beyond bias-variance via anova," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 6925–7006, 2021.

[32] P. Nakkiran, "More data can hurt for linear regression: Sample-wise double descent," *arXiv preprint arXiv:1912.07242*, 2019.

[33] R. Heckel and F. F. Yilmaz, "Early stopping in deep networks: Double descent and how to eliminate it," *arXiv preprint arXiv:2007.10099*, 2020.

[34] A. Immer, M. Bauer, V. Fortuin, G. Rätsch, and K. M. Emtiyaz, "Scalable marginal likelihood estimation for model selection in deep learning," in *International Conference on Machine Learning*, pp. 4563–4573, PMLR, 2021.

[35] H. Akaike, "Likelihood of a model and information criteria," *Journal of Econometrics*, vol. 16, no. 1, pp. 3–14, 1981.

[36] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[37] G. Aminian, Y. Bu, L. Toni, M. Rodrigues, and G. Wornell, "An exact characterization of the generalization error for the Gibbs algorithm," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8106–8118, 2021.

[38] K. P. Murphy, "Conjugate bayesian analysis of the gaussian distribution," *def*, vol. 1, no. $2\sigma 2$, p. 16, 2007.

[39] J. Pennington and P. Worah, "Nonlinear random matrix theory for deep learning," in *Advances in Neural Information Processing Systems*, pp. 2637–2646, 2017.

[40] A. M. Tulino, S. Verdú, and S. Verdu, *Random matrix theory and wireless communications*. Now Publishers Inc, 2004.

A Related Works

Previous work has extended the classical BIC to high dimensions, e.g.,[13, 14]. However, these works seek to *replace* maximizing marginal likelihood with a different criterion, substituting a penalty term $pf(n)$ in place of the $p\log n$ in BIC. By contrast, we retain the BIC criterion but analyze it beyond the classical regime for Gibbs with an information-theoretic analysis.

Double-descent of the population risk with increasing model size was introduced in [3]; see also [15, 16]. An empirical demonstration of double-descent in modern deep networks is provided in [17]. A variety of work develops simplified models where the characterization of the double-descent curve can be obtained. For example, double-descent in linear regression models is investigated in [18, 19, 20, 21], and in linear classification models in [22, 23, 24]. The RF model has been adopted to understand double-descent in [8], which provides a generalization analysis of the performance achieved with ridge regression in the over-parameterized regime. In some of our analysis we likewise adopt this RF model [8, 25, 24, 26, 27], but with a different objective and analysis tools.

Double-descent phenomena have been explained from different perspectives. In [25, 28, 29], double-descent curves are explained via a refined version of bias-variance tradeoff, where the bias of the model decreases monotonically with the increase of p, but the variance increases and then decreases with p. In addition, [30] and [31] employ the analysis of variance (ANOVA) to decompose the variance of the test error to identify contributing factors to the double descent phenomenon observed in linear problems with two-layer neural networks. In particular, [31] calculates the limits of the variance components of the Marchenko-Pastur distribution, a powerful analytical tool that we have adopted and further expanded upon in our paper. The connection of gradient descent dynamics and double-descent is discussed in [26, 15]. In [32, 9], sample-wise double-descent is studied under linear regression, and [33] shows that by adjusting the step sizes, sample-wise double-descent can be eliminated by early stopping.

Among recent work, our paper is most related to [34, 10], which also examines the difference between marginal likelihood and generalization error in model selection. However, by focusing on the Gibbs, we are able to interpret the mismatch between AIC and BIC via information measures, which is more insightful in understanding double descent and other complex behaviors.

B The Classical Forms of AIC and BIC

The standard derivation of the AIC and the BIC arises from the classical asymptotic analysis of MLE. Assume we have K candidate models $M_1, M_2, \dots, M_K$, and each model M_k is characterized by a parametric probabilistic model $P_k(y|\boldsymbol{x};\boldsymbol{\theta}_k)$, a prior distribution $\pi_k(\boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k \in \mathcal{W}_k \subset \mathbb{R}^{p_k}$ is the parameter vector. We demonstrate that AIC selects the model with the smallest population risk, and BIC identifies the true data-generating model by maximizing the marginal likelihood.

AIC This criterion [35] ranks statistical models based on the Kullback-Leibler (KL) divergence between the true data distribution P_Z and the learned parametric model. With $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^{(k)}$ denoting the MLE of the kth model, AIC selects the model as the solution to

$$\operatorname*{argmin}_k D(P_Z \| P_k(y|\boldsymbol{x};\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^{(k)})) = \operatorname*{argmin}_k \mathbb{E}_{P_Z}\big[-\log P_k(y|\boldsymbol{x};\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^{(k)})\big]. \tag{11}$$

The term $\mathbb{E}_{P_Z}[-\log P_k(y|\boldsymbol{x};\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^{(k)})]$ can be interpreted as the population risk $L_P(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^{(k)}, P_Z)$ of the MLE under log-loss. From this perspective, AIC measures how well the model fits the unknown data distribution P_Z, with smaller AIC values suggesting a lower population risk.

As the true distribution P_Z is unknown, the AIC is obtained by approximating the population risk as the sum of empirical risk, i.e., the negative log-likelihood of $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^{(k)}$ on training samples and a penalty term corresponding to the generalization error. In the classic regime where p_k is fixed and $n \to \infty$, the asymptotic normality of MLE yields

$$\mathrm{AIC}(M_k) \triangleq -\frac{\hat{L}_k(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^{(k)})}{n} + \frac{p_k}{n}, \text{ whence } \mathrm{AIC} = -\frac{\hat{L}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})}{n} + \frac{p}{n}. \tag{12}$$

Note that our form of AIC differs by a factor of 2 from its classical form to facilitate a direct comparison to population risk and generalization error.

BIC This criterion [36] ranks statistical models by their marginal likelihoods of generating the data, with smaller values of the BIC correspond to larger marginal likelihoods. Approximating the marginal likelihood of observing $\boldsymbol{z}^n$ for M_k, i.e., $m_k(\boldsymbol{z}^n) \triangleq \int P_k(y^n|\boldsymbol{x}^n;\boldsymbol{\theta}_k)\,\pi_k(\boldsymbol{\theta}_k)\,d\boldsymbol{\theta}_k$, using Laplace's method yields

$$\log m_k(\boldsymbol{z}^n) = \hat{L}_k(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^{(k)}) - \frac{p_k}{2}\log n + O(1), \quad n \to \infty. \tag{13}$$

In turn, BIC is obtained from (13) by dropping terms that don't scale with n and scaling by $-1/n$:

$$\mathrm{BIC}(M_k) \triangleq -\frac{\hat{L}_k(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}^{(k)})}{n} + \frac{p_k \log n}{2n}, \ \text{ whence } \ \mathrm{BIC} = -\frac{\hat{L}(\hat{\boldsymbol{\theta}}_{\mathrm{ML}})}{n} + \frac{p\log n}{2n}. \tag{14}$$

When, further, $P(M_k) = 1/K$, we obtain

$$P(M_k|\boldsymbol{z}^n) = \frac{m_k(\boldsymbol{z}^n)P(M_k)}{\sum_{k=1}^{K} m_k(\boldsymbol{z}^n)P(M_k)} \propto m_k(\boldsymbol{z}^n). \tag{15}$$

Thus, when we assume a uniform prior over different models, the BIC ranks models by their posterior probability of generating the training data, and choosing the smallest BIC corresponds to the maximum a posteriori rule for model selection.

Both (12) and (14) share a common first term, representing the average negative log-likelihood of the training data for MLE, which can be interpreted as the empirical risk with log-loss, decreasing as we adopt more complex models. We note that AIC and BIC in the classical $n \to \infty$ regime are independent of the form of the model family $P(y|\boldsymbol{x};\boldsymbol{\theta})$ and the prior distribution $\pi(\boldsymbol{\theta})$, which makes it compatible with general distribution families subject to mild smoothness constraints. But because AIC and BIC differ in the second penalty term, they select different models.

C Proof of Proposition 1

If we adopt the log-loss function $\ell(w,\boldsymbol{z}) = -\log P(y|\boldsymbol{x};w)$, and set $\beta = n$, the Gibbs distribution can be viewed as the Bayesian posterior distribution, i.e.,

$$P^*_{\hat{W}|S}(\boldsymbol{w}|\boldsymbol{z}^n) = \frac{\pi(\boldsymbol{w})\prod_{i=1}^{n} P(y_i|\boldsymbol{x}_i;\boldsymbol{w})}{V(\boldsymbol{z}^n)}, \text{ with } V(\boldsymbol{z}^n) = \int \pi(\boldsymbol{w})\prod_{i=1}^{n} P(y_i|\boldsymbol{x}_i;\boldsymbol{w})d\boldsymbol{w}. \tag{16}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{P^*_{\hat{W}|S=\boldsymbol{z}^n}}\big[L_E(\hat{W},\boldsymbol{z}^n)\big] + \frac{1}{n}D(P^*_{\hat{W}|S=\boldsymbol{z}^n}\|\pi)\\
&= \mathbb{E}_{P^*_{\hat{W}|S=\boldsymbol{z}^n}}\big[L_E(\hat{W},\boldsymbol{z}^n)\big] + \frac{1}{n}\mathbb{E}_{P^*_{\hat{W}|S=\boldsymbol{z}^n}}\big[\log\frac{\exp\big(-nL_E(\hat{W},\boldsymbol{z}^n)\big)}{V(z^n)}\big]\\
&= -\frac{1}{n}\log V(z^n)\\
&= -\frac{1}{n}\log m(\boldsymbol{z}^n),
\end{aligned} \tag{17}$$

which completes the proof.

D Gibbs-based AIC

As we discussed in Section B, the penalty term in AIC can be viewed as the generalization error of MLE with log-loss. Thus, we start with the following result from [37], which provides an exact characterization for the generalization error of the Gibbs algorithm using information measure.

Proposition 2 ([37]). *For the Gibbs algorithm $P^*_{\hat{W}|S}$, the expected generalization error is*

$$\overline{gen}(P^*_{\hat{W}|S}, P_S) = I_{\mathrm{SKL}}(P^*_{\hat{W}|S}, P_S)/\beta, \tag{18}$$

where $I_{\mathrm{SKL}}(P^*_{\hat{W}|S}, P_S)$ is the symmetrized KL information between $\hat{W}$ and S, defined as follows

$$I_{\mathrm{SKL}}(P_{Y|X}, P_X) \triangleq D(P_{X,Y}\|P_X \otimes P_Y) + D(P_X \otimes P_Y\|P_{X,Y}). \tag{19}$$

Notably, information risk minimization in (2) regularizes the mutual information $I(\hat{W};S)$ as a proxy of the generalization error, but the exact generalization error of the Gibbs algorithm is the symmetrized KL information, which is always larger than the mutual information.

Thus, Proposition 2 motivates the following *Gibbs-based* AIC:

$$\mathrm{AIC}^{+} \triangleq L_E(\hat{W}_{\mathrm{Gibbs}}, \boldsymbol{z}^n) + \frac{1}{\beta} I_{\mathrm{SKL}}(P^{*}_{\hat{W}|S}, P_S). \tag{20}$$

Observe that the penalty term in Gibbs-based AIC is an information measure that characterizes the generalization error of the Gibbs algorithm. By investigating the asymptotic behavior of $I_{\mathrm{SKL}}(P^{*}_{\hat{W}|S}, P_S)$, we have the following theorem characterizes the Gibbs-based AIC in the classical asymptotic regime.

Evidently, our information-theoretic analysis has the same AIC penalty term for the AIC^{+} in the classical regime, which suggests that the generalization error of the Gibbs algorithm (Gibbs) has the same order of p/n as that of the MLE (SGD) in this regime.

E Gibbs Distribution of Random Feature Model

For random feature model with the prior distribution $\pi(w) \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n}\boldsymbol{I}_p)$ and $L_E(w,s) = \frac{1}{n}\sum_{i=1}^{n}\log(y_i|x_i,w)$, the log-posterior $\log(P^{*}_{\hat{W}|S}) \propto \log\pi(w) + \log(e^{-nL_E(S)})$, where

$$L_E(S) = \frac{1}{2n\sigma^2}\|\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{w}\|_2^2 + \frac{1}{2}\log(2\pi\sigma^2).$$

Thus, the Gibbs algorithm, in this case, is given by the following Gaussian posterior distribution, as shown in [38],

$$P^{*}_{\hat{W}|S} \sim \mathcal{N}(\hat{W}_\lambda, \boldsymbol{\Sigma}_w), \tag{21}$$

where $\hat{W}_\lambda = (\lambda n\boldsymbol{I}_p + \boldsymbol{B}^\top\boldsymbol{B})^{-1}\boldsymbol{B}^\top\boldsymbol{Y}$, and $\boldsymbol{\Sigma}_w = \sigma^2(\lambda n\boldsymbol{I}_p + \boldsymbol{B}^\top\boldsymbol{B})^{-1}$.

F Gibbs-based BIC for Over-Parameterized RF Model

In particular, for activation functions $f(\cdot)$ that satisfy conditions

$$\mathbb{E}[f(\varepsilon)] = 0, \quad \mathbb{E}[f(\varepsilon)^2] = 1, \quad \mathbb{E}[f'(\varepsilon)] = 0, \quad \left|\mathbb{E}[f(\varepsilon)^k]\right| < \infty, \qquad \text{for } k > 1, \tag{22}$$

where $\varepsilon \sim \mathcal{N}(0,1)$, the following theorem characterizes the KL divergence term in the over-parameterized RF model.

Theorem 1. *For activation functions $f(\cdot)$ satisfying the conditions in* (22)*, as $n, d, p \to \infty$ with $p/d \to r_1$, $n/d \to r_2$, and $r_1/r_2 = r$, where $r_1, r_2 \in (0,\infty)$, we have*

$$\frac{1}{n}D(P^{*}_{\hat{W}|S=\boldsymbol{z}^n}\|\pi) \to \frac{\lambda}{2\sigma^2}\|\hat{W}_\lambda\|_2^2 - \frac{\lambda}{8}\mathcal{F}(\frac{1}{\lambda}, r) + \frac{1}{2}V(1/\lambda, r) \tag{23}$$

almost surely, where

$$V(\gamma,r) \triangleq r\log\Big(1+\gamma-\frac{1}{4}\mathcal{F}(\gamma,r)\Big) - \frac{1}{4\gamma}\mathcal{F}(\gamma,r) + \log\Big(1+\gamma r - \frac{1}{4}\mathcal{F}(\gamma,r)\Big), \tag{24}$$

with

$$\mathcal{F}(\gamma,r) \triangleq \Big(\sqrt{\gamma(1+\sqrt{r})^2+1} - \sqrt{\gamma(1-\sqrt{r})^2+1}\Big)^2. \tag{25}$$

F.1 Proof of Theorem 1

We note that the KL divergence between two Gaussian distributions can be written as

$$D(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\|\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) = \frac{1}{2}\left[(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - p + \log\frac{\det\boldsymbol{\Sigma}_2}{\det\boldsymbol{\Sigma}_1}\right].$$

The KL divergence between the Gibbs posterior of the RF model and the prior can be computed by,

$$\frac{1}{n}D(P^*_{\hat{W}|S=\boldsymbol{z}^n}\|\pi) = \frac{1}{2n}\left[\hat{W}_\lambda^\top(\frac{\sigma^2}{\lambda n}\boldsymbol{I}_p)^{-1}\hat{W}_\lambda + \mathrm{tr}(\frac{\lambda n}{\sigma^2}\boldsymbol{\Sigma}_w) + \log\frac{\det(\frac{\sigma^2}{\lambda n}\boldsymbol{I}_p)}{\det(\boldsymbol{\Sigma}_w)} - p\right] \tag{26}$$

$$= \frac{1}{2n}\left[\frac{\lambda n}{\sigma^2}\|\hat{W}_\lambda\|_2^2 + \mathrm{tr}\big((\boldsymbol{I}_p + \frac{\boldsymbol{B}^\top\boldsymbol{B}}{\lambda n})^{-1}\big) + \log\det(\boldsymbol{I}_p + \frac{\boldsymbol{B}^\top\boldsymbol{B}}{\lambda n}) - p\right].$$

The trace and the log determinant of the random matrix $\boldsymbol{\Sigma} = \frac{\boldsymbol{B}^\top\boldsymbol{B}}{\lambda n} + \boldsymbol{I}_p$ can be computed using the following results from [39], which characterizes the probability density function of the eigenvalues of the random matrix $\boldsymbol{B}^\top\boldsymbol{B}/n$ in the over-parameterized regime.

Lemma 3. *[39] Let the matrix $\boldsymbol{M} = \frac{1}{n}\boldsymbol{B}^\top\boldsymbol{B} \in \mathbb{R}^{p\times p}$, where $\boldsymbol{B} = f\Big(\frac{\boldsymbol{X}\boldsymbol{F}}{\sqrt{d}}\Big) \in \mathbb{R}^{n\times p}$, all the elements in $\boldsymbol{F} \in \mathbb{R}^{d\times p}$ and $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ are generated i.i.d from $\mathcal{N}(0,1)$. Suppose that the activation function has zero mean and finite moments, i.e.,*

$$\mathbb{E}[f(\varepsilon)] = 0, \quad \mathbb{E}[f(\varepsilon)^k] < \infty, \ \textit{for}\ k > 1, \quad \varepsilon \sim \mathcal{N}(0,1). \tag{27}$$

and define constants η and ξ as

$$\eta \triangleq \mathbb{E}[f(\varepsilon)^2], \quad \xi \triangleq \mathbb{E}[f'(\varepsilon)]^2, \quad \varepsilon \sim \mathcal{N}(0,1), \tag{28}$$

as $n, d, p \to \infty$ with $d/p \to \psi$, $d/n \to \phi$, where $\psi, \phi \in (0, \infty)$, then the Stieltjes transform $G(z)$ of the spectral density of random matrix $\boldsymbol{M}$ satisfies

$$dF_{\boldsymbol{M}}(x) = \frac{1}{\pi}\lim_{\epsilon\to 0^+}\mathrm{Im}G(x - i\epsilon), \quad G(z) = \frac{\psi}{z}A\Big(\frac{1}{z\psi}\Big) + \frac{1-\psi}{z}, \tag{29}$$

$$A(t) = 1 + (\eta - \xi)tA_\phi(t)A_\psi(t) + \frac{A_\phi(t)A_\psi(t)t\xi}{1 - A_\phi(t)A_\psi(t)t\xi}, \tag{30}$$

where $A_\phi(t) = 1 + (A(t)-1)\phi$ and $A_\psi(t) = 1 + (A(t)-1)\psi$.

This lemma characterizes the spectral density of random matrix $\boldsymbol{M}$ for any zero-mean activation functions. However, these implicit equations need to be evaluated numerically, and it is hard to obtain a closed-form expression or provide more insights.

If we further assume that the assumptions in (22) are satisfied, i.e., $\mathbb{E}[f(\varepsilon)^2] = 1$, and $\mathbb{E}[f'(\varepsilon)]^2 = 0$, then the result in Lemma 3 can be simplified significantly, as the probability density of the eigenvalues for random matrix $\boldsymbol{M}$ will converge to the well-known Marchenko-Pastur distribution with shape parameter $r = p/n$, i.e.,

$$dF_{\boldsymbol{M}}(x) \to (1 - \frac{1}{r})^+\delta(x) + \frac{\sqrt{(x-a)^+(b-x)^+}}{2\pi r x}, \tag{31}$$

as n, d, p all go to infinity, where $(z)^+ \triangleq \max\{0, z\}$, and $a \triangleq (1-\sqrt{r})^2$, and $b \triangleq (1+\sqrt{r})^2$. Thus, we focus on this case to obtain a convenient, closed-form expression for mutual information.

The following lemma from Sections 2.2.2 and 2.2.3 in [40] characterizes the η-transform and Shannon transform of the Marchenko-Pastur distribution.

Lemma 4. *The η and Shannon transform of a nonnegative random variable X are defined as*

$$\eta_X(\gamma) \triangleq \mathbb{E}[\frac{1}{1+\gamma X}], \quad \mathcal{V}_X(\gamma) \triangleq \mathbb{E}[\log(1+\gamma X)], \tag{32}$$

respectively. If X is distributed according to Marchenko-Pastur distribution with shape parameter $r = p/n$, then

$$\eta_X(\gamma) = 1 - \frac{\mathcal{F}(\gamma, r)}{4r\gamma}, \tag{33}$$

$$\mathcal{V}_X(\gamma) = \log\left(1 + \gamma - \frac{1}{4}\mathcal{F}(\gamma,r)\right) + \frac{1}{r}\log\left(1 + \gamma r - \frac{1}{4}\mathcal{F}(\gamma,r)\right) - \frac{1}{4r\gamma}\mathcal{F}(\gamma,r), \tag{34}$$

where

$$\mathcal{F}(\gamma, r) \triangleq \left(\sqrt{\gamma(1+\sqrt{r})^2 + 1} - \sqrt{\gamma(1-\sqrt{r})^2 + 1}\right)^2. \tag{35}$$

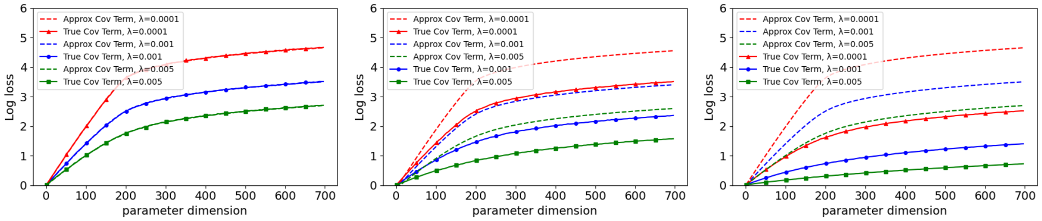

Figure 3: A comparison between $1/n(\log|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma})^{-1} - p)$ and the asymptotic approximation of the covariance term in Theorem 1 for different value of λ and for different activation functions: $f(x) = (x^2-1)/\sqrt{2}$ (**left**), ReLU (**middle**) and Sigmoid (**right**). We adopt the same experiment settings as in Section 5, and we change $r = p/n$ by fixing $n = 200$ and varying p.

Equipped with the aforementioned tools from random matrix theory, we could proceed our analysis,

$$\frac{1}{n}\log\det\big(\boldsymbol{I}_p + \frac{1}{\lambda n}\boldsymbol{B}^\top\boldsymbol{B}\big) = \frac{r}{p}\sum_{i=1}^{p}\log\big(1+\frac{1}{\lambda}\lambda_i(\frac{1}{n}\boldsymbol{B}^\top\boldsymbol{B})\big), \tag{36}$$

where the notation $\lambda_i(\cdot)$ denote the eigenvalues of the matrix for $i = 1, \cdots, p$. As shown in Lemma 4, we have

$$\frac{1}{p}\sum_{i=1}^{p}\log\big(1+\frac{1}{\lambda}\lambda_i(\frac{1}{n}\boldsymbol{B}^\top\boldsymbol{B})\big) \to \int_0^\infty \log\big(1+\frac{x}{\lambda}\big)dF^n_{\boldsymbol{M}}(x) = \mathcal{V}_X(1/\lambda) \tag{37}$$

almost surely, when $n, d, p \to \infty$, $p/n = r$. Thus, in the over-parameterized regime, we have

$$\frac{1}{n}\log\det\big(\boldsymbol{I}_p + \frac{1}{\lambda n}\boldsymbol{B}^\top\boldsymbol{B}\big) \to r\cdot\mathcal{V}_X(1/\lambda) = V(1/\lambda, r). \tag{38}$$

And the trace term can be simplified as,

$$\frac{1}{n}\mathrm{tr}\big(\boldsymbol{I}_p + \frac{1}{\lambda n}\boldsymbol{B}^\top\boldsymbol{B}\big)^{-1} = r\frac{1}{p}\sum_{i=1}^{p}\frac{1}{\big(1+\frac{1}{\lambda}\lambda_i(\frac{1}{n}\boldsymbol{B}^\top\boldsymbol{B})\big)}, \tag{39}$$

which will converge to the following expression by Lemma 4, when $n, d, p \to \infty$, $p/n = r$,

$$r\frac{1}{p}\sum_{i=1}^{p}\frac{1}{\big(1+\frac{1}{\lambda}\lambda_i(\frac{1}{n}\boldsymbol{B}^\top\boldsymbol{B})\big)} \to r\int_0^\infty \frac{1}{1+\frac{x}{\lambda}}dF^n_{\boldsymbol{M}}(x) = r\big(1 - \frac{\mathcal{F}(\frac{1}{\lambda}, r)}{4r\frac{1}{\lambda}}\big). \tag{40}$$

Combine (38) and (40) with (26), we obtain the following result

$$\frac{1}{n}D(P^*_{\hat{W}|S=\boldsymbol{z}^n}\|\pi) \to \frac{\lambda}{2\sigma^2}\|\hat{W}_\lambda\|_2^2 - \frac{\lambda}{8}\mathcal{F}(\frac{1}{\lambda}, r) + \frac{1}{2}V(1/\lambda, r). \tag{41}$$

F.2 Empirical Behavior of Covariance Term

To show that Theorem 1 can provide a good approximation for the asymptotic behavior of the log determinant and trace term in (8), we plot in Figure 3 both the term $1/n(\log|\boldsymbol{\Sigma}| + \mathrm{tr}(\boldsymbol{\Sigma})^{-1} - p)$ with finite data and $\frac{1}{2}V(1/\lambda, r) - \frac{\lambda}{8}\mathcal{F}(\frac{1}{\lambda}, r)$ in the over-parameterized Gibbs-based BIC for different activation functions with varying regularizer parameters λ. As shown from Figure 3 (left), our theoretical results provide a good proxy for the asymptotic behavior of the covariance term, even for activation functions (e.g., ReLU and Sigmoid in Figure 3 (middle and right)) that do not satisfy the assumptions in Theorem 1. This is evidence that the particular choice of activation function does not significantly influence the asymptotic behavior of the covariance term in Gibbs-based BIC.

F.3 Additional Experiments with anisotropic data distribution

The experimental setup remains consistent with the setting in Section 5, except for the data generating distribution of the training samples $\boldsymbol{z}^n = (\boldsymbol{x}_i, y_i)_{i=1}^n$. Here, $\boldsymbol{x}_i$ values are drawn i.i.d. from anistropic distribution $\mathcal{N}(0, \boldsymbol{V})$. The variance in this distribution, denoted by $\boldsymbol{V}$, varies within a range of 0 to 2. As Figure 4 and 5 show, there is no significant difference from the figures we contained in Section 5, and the proposed Gibbs-based BIC works for general covariance matrices.

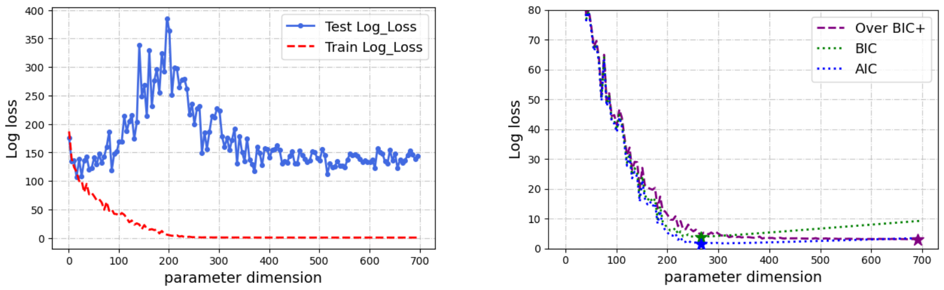

Figure 4: A comparison of the test and training log loss achieved by Gibbs with varying p when $\lambda = 0.005$ (**left**). A comparison of over-parameterized BIC^+ and classical BIC with varying p (**right**). The preferred models selected by different information criteria are marked using stars.

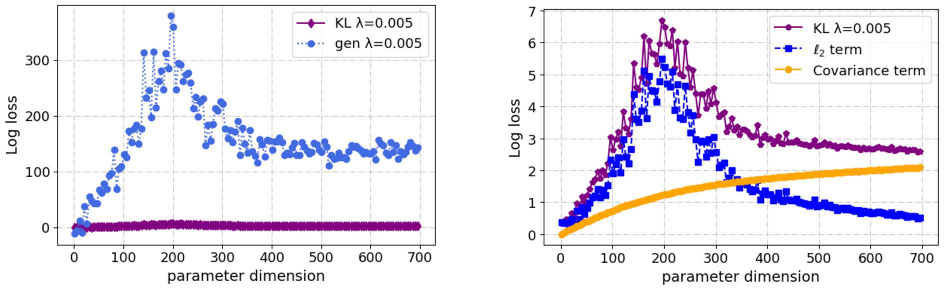

Figure 5: A comparison between the KL-divergence term in BIC^+ and the generalization error term (**left**). A decomposition of the terms in over-parameterized BIC^+ in (9) with $\lambda = 0.005$ (**right**).