Supplementary Material for "Variational Inference and Model Selection with Generalized Evidence Bounds"

Chenyang Tao, Liqun Chen, Ruiyi Zhang, Ricardo Henao, Lawrence Carin
Electrical & Computer Engineering
Duke University
Durham, NC 27708, USA
chenyang.tao, liqun.chen, rz68, ricardo.henao, lcarin@duke.edu

To simplify our notation, we denote

$$W(Z_{1:K}) \triangleq \frac{1}{K}\sum_{k=1}^{K}\frac{p_\alpha(x,Z_k)}{q_\beta(Z_k|x)}. \tag{1}$$

A Proof for Theorem 2

Proof. We first prove $\text{GLBO}(x;K) \leq \phi(p_\alpha(x))$. This is a direct result of Jensen's inequality

$$\begin{aligned}\text{GLBO}(x;K) &\leq \psi^{-1}\left(h\left(\mathbb{E}_{Z_{1:K}\sim q}\tfrac{1}{K}\textstyle\sum\frac{p_\alpha(x,Z_k)}{q_\beta(Z_k|x)}\right)\right)\\ &= \psi^{-1}(h(p_\alpha(x))) = \phi(p_\alpha(x)).\end{aligned}$$

To prove that $\text{GLBO}(x;K)$ is non-decreasing wrt to K, we apply a similar technique used in [2]. We assume $0 < K_1 < K_2$. Let $I \subset \{1,\cdots,K_2\}$ with $\#(I) = K_1$ be a uniformly distributed subset of distinct indices from $\{1,\cdots,K_2\}$. Note it holds that $\mathbb{E}_{I=\{i_1,\cdots,i_{K_1}\}}\left[\frac{a_{i_1}+\cdots+a_{i_{K_1}}}{K_1}\right] = \frac{1}{K_2}\sum_i a_i$ for any sequence of numbers $\{a_1,\cdots,a_{K_2}\}$. Together with Jensen's inequality, we have

$$\begin{aligned}&\text{GLBO}(x;K_1) = \psi^{-1}\left(\mathbb{E}_{Z_{1:K_1}}\left[h(W(Z_{1:K_1}))\right]\right)\\ &= \psi^{-1}\left(\mathbb{E}_{Z_{1:K_2}}\left[\mathbb{E}_{I=\{i_1,\cdots,i_{K_1}\}}\left[h(W(Z_{i_1:i_{K_1}}))\right]\right]\right)\\ &\leq \psi^{-1}\left(\mathbb{E}_{Z_{1:K_2}}\left[h(\mathbb{E}_{I=\{i_1,\cdots,i_{K_1}\}}\left[W(Z_{i_1:i_{K_1}})\right])\right]\right)\\ &= \psi^{-1}\left(\mathbb{E}_{Z_{1:K_2}}\left[h(W(Z_{1:K_2}))\right]\right)\\ &= \text{GLBO}(x;K_2).\end{aligned}$$

$\text{GLBO}(x;K)$'s convergence to $\phi(p_\alpha(x))$ as K goes to infinity can be proved by applying the law of large numbers.

B Proof for Theorem 3

Proof. We know that $\psi(u)$ is convex and non-decreasing, this implies $\psi^{-1}(u)$ is concave. Using Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}_{Z_{1:K}}[\phi(W(Z_{1:K}))] &= \mathbb{E}_{Z_{1:K}\sim q}\left[\psi^{-1}\left(h\left(W(Z_{1:K})\right)\right)\right]\\ &\leq \psi^{-1}\left(\mathbb{E}_{Z_{1:K}\sim q}\left[h\left(W(Z_{1:K})\right)\right]\right).\end{aligned}$$

This concludes our proof.

35^{th} International Conference on Machine Learning (ICML 2018), Stockholm, Sweden.

C Proof for Theorem 5

Proposition 9. (Lyapunov inequality) For a random variable X and numbers $0 < r < s < +\infty$, it holds that

$$(\mathbb{E}[|X|^r])^{1/r} \leq (\mathbb{E}[|X|^s])^{1/s}.$$

Proof for Theorem 5. We have

$$\text{CLBO}(x; K, T) = \log\left(\left(\mathbb{E}_{Z_{1:K}\sim q}\left[\left(W(Z_{1:K})\right)^{1/T}\right]\right)^T\right),$$

and the result follows by applying the Lyapunov inequality from Proposition 9 while noting that $\log(u)$ is monotonically increasing.

D Proof for Theorem 6

Lemma 10. $h(u; T) = \exp(\frac{1}{T}\log(u))$ is concave when $T > 1$.

Proof. We only need to prove $h''(u; T) < 0$. It is easy to show

$$h'(u; T) = \frac{1}{Tu}h(u; T),$$
$$h''(u; T) = \frac{1-T}{T^2u^2}h(u; T).$$

Since $u > 0, T > 1$ and $h(u; T) > 0$, therefore we have $h''(u; T) < 0$.

Proof for Theorem 6.
1. By Theorem 2 and Lemma 10, we have

$$\text{CLBO}(x; 1, T) \leq \text{CLBO}(x; K, T) \leq \log p_\alpha(x).$$

Since we know $\lim_{T\to 1}\text{CLBO}(x; 1, T) \to \log p_\alpha(x)$ from D. Blei's χ-VI paper, the result follows.
2. Use Taylor expansion.

E Proof for Theorem 8.

Proof. By the use of Jensen's inequality, we have

$$\begin{aligned}\text{RVB}(x; K, T) &= \mathbb{E}_{Z_{1:K}}\left[T\log\left(\tfrac{1}{K}\textstyle\sum_{k=1}^K\left(W(Z_k)\right)^{1/T}\right)\right]\\ &\leq T\log\left(\mathbb{E}_{Z_{1:K}}\left[\tfrac{1}{K}\textstyle\sum_{k=1}^K\left(W(Z_k)\right)^{1/T}\right]\right)\\ &\leq T\log\left(\mathbb{E}_{Z_{1:K}}\left[\left(W(Z_{1:K})\right)^{1/T}\right]\right)\\ &= \text{CLBO}(x; K, T),\end{aligned}$$

where $W(Z_k) = \frac{p_\alpha(x, Z_k)}{q_\beta(Z_k|x)}$.

F Upper bounds

To establish the ϕ-evidence upper bounds, we exchange the concavity and convexity in the theories discussed in the main text. Now we assume that: (iv) $\phi(u)$ is convex, (v) $\psi(u)$ is concave, and (vi) $h(u) \triangleq \psi(\phi(u))$ is convex. Reverting the inequalities in Theorem 2, Theorem 3 and Theorem 5 gives the respective upper bound counterpart. We omit the the proofs as they are similar to those for the lower bounds.

As a concrete example, consider the empirical χ^2 evidence upper bound estimator used in [3]

$$\chi^2(x; K) \triangleq \mathbb{E}_{Z_{1:K}\sim q}\left[\frac{1}{K}\sum_{k=1}^K\left(\frac{p_\alpha(x, Z_k)}{q_\beta(Z_k|x)}\right)^2\right],$$

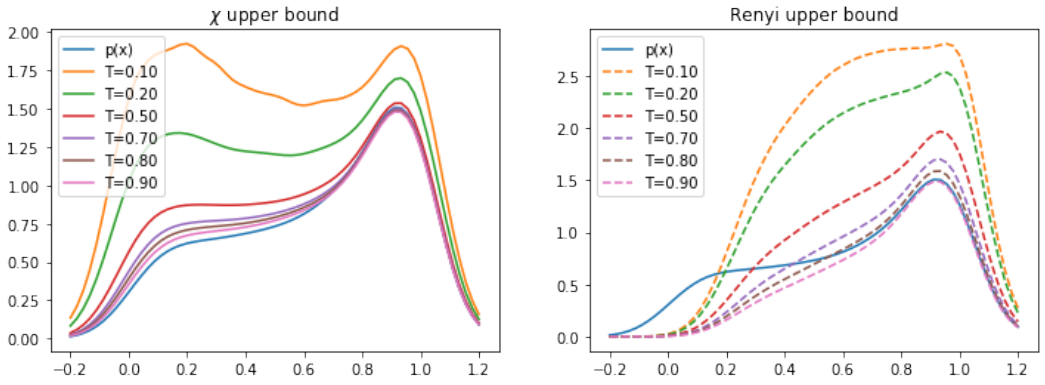


Figure SM 1: Comparison of theoretical upper bounds on the toy distribution. $K = 2$ in this experiment.

this is our ϕ-evidence upper bound with $K = 1, \phi(u) = u^2$ and $\psi(u) = u$. Our framework lends $\chi^2(x; K)$ more theoretical justifications: unlike RVB(x, K), $\mathbb{E}_{Z_{1:K}}\left[\chi^2(x, K)\right]$ is guaranteed to be an upper bound on χ^2-evidence score. Our theory also provides ways to improve $\chi^2(x; K)$'s performance, as leveraging importance sampling and a concave ψ is guaranteed to sharpen the bound. We compare our K-sample generalized upper bound with Rényi upper bound on the toy model distribution in Figure SM 1.

We remark that optimizing the upper bound is numerically more difficult than the lower bound. The challenge comes from sample estimate of the term $\frac{1}{q_\beta(z|x)}$ in importance-weighted estimator (1). Large values of $\frac{1}{q_\beta(z|x)}$ will be sampled with vanishingly small probability. Large values of $\frac{1}{q_\beta(z|x)}$, which will be sampled predominately, usually does not contribute much to the actual integral. Additionally, since $W(Z), Z \sim q_\beta(z|x)$ can be vary across a large numerical range, very unstable gradient estimates (high variance) can be expected for complex problems. In our experiments, we are unable to use a reasonable number of posterior samples to successfully optimize the upper bound. We hypothesize that introducing an auxiliary proposal distribution that is more informative on the geometry of the IW-estimate $W(z)$ can help. We leave this idea for future exploration.

G Maximal entropy argument for model selection with a saturating $\phi(u)$

Now we provide an alternative justification for using a saturating evidence function $\phi(u)$. Let $\tilde{\mathcal{D}}_m = \{\tilde{x}_i\}_{i=1}^m$ be m iid samples from data distribution $p_d(x)$, and $\hat{p}_m = \frac{1}{m}\sum_{i=1}^m \delta_{\tilde{x}_i}$ be the corresponding empirical distribution. Consider a discrete approximation $q(x)$ to $p_d(x)$, such that samples are only allowed to take values from $\boldsymbol{X}_m$. A natural choice $\hat{q}_m$ yielding a good (discrete) approximation would be $\hat{q}_m(x_i) \propto q_{\beta^*}(x_i)$, where $q_{\beta^*}(x)$ is a model that best explains the training samples[1]. We also know that as $m \to \infty$, $\hat{p}_m(x)$ converges to $p_d(x)$. Therefore, we want the difference between $\hat{q}_m(x)$ and $\hat{p}_m(x)$, for instance, via KL$(\hat{q}_m \parallel \hat{p}_m)$, to be small. Algebraic manipulation reveals that minimization of KL$(\hat{q}_m \parallel \hat{p}_m)$ is equivalent to the maximization of the following Shannon entropy term

$$-\sum_i q_\beta(x_i) \log q_\beta(x_i).$$

This closely related to the principle of maximum entropy learning [4, 5], which states that under the constraints of *testable information*, the best distribution that represents the current state of knowledge is the one with largest uncertainty, as measured by the Shannon entropy. In general, an evidence distribution with lower variance yields higher entropy, and in practice, a saturating ϕ-evidence function, as discussed above, encourages such low variance evidence distribution. In the model-selection setting, we can treat the expected log-evidence score as our testable information. More specifically, we want to use GLBO with saturating $\phi(u)$ evidence to reduce the variance of evidence distribution while maintaining a high log-evidence bound.

[1]Training samples do not necessarily overlap with $\tilde{\mathcal{D}}_m$.

H An asymptotic argument for the moving average estimator

Let f be a monotonically increasing concave function, $X = x_0 + \eta$, where η is a mean zero random variable with small absolute value. Using Taylor expansion, the bias of $\mathbb{E}_X[f(X)]$ wrt $f(\mathbb{E}[X])$ can be approximated as

$$\mathbb{E}_X[f(X)] - f(\mathbb{E}[X]) \approx f''(x_0)\text{var}[\eta],$$

which implies the bias diminishes linearly wrt the variance of η, when η is sufficiently small.

We also note that the moving average trick can be also applied to the K-sample estimate term

$$\hat{p}_K(x) = \frac{1}{K}\sum_{k=1}^{K}\frac{p_\alpha(x, Z_k)}{q_\beta(z_k|x)}.$$

Our objective then becomes

$$J_{\text{GLBO}}^{\text{ema},2}(x,t) = \psi^{-1}(\hat{h}_{\text{ema},2}(x,t)),$$

with $\hat{h}_{\text{ema},2}(x,t)$ iteratively defined as

$$\begin{aligned}\hat{h}_{\text{ema},2}(x,t) &= (1-w_t)\hat{h}_{\text{ema},2}(x,t-1)\\ &\quad + w_t h\left(\hat{p}_{\text{ema}}(x,t)\right),\\ \hat{p}_{\text{ema}}(x,t) &= (1-v_t)\hat{p}_{\text{ema}}(x,t-1)\\ &\quad + v_t\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p_\alpha(x,z_{t,k})}{q_\beta(z_{t,k}|x)}\right),\end{aligned}$$

where $w_t, v_t \in [0,1]$ are learning rates.

I Choice of temperature T and empirical performance

To examine how temperature T affects empirical performance, we vary T from 2^3 to 2^{10} on the log-scale using the AVB GLBO variant. Figure 2 summarizes the test log-evidence result for the MNIST dataset. The performance is maximized with a moderate temperature. As T shrinks further, the optimization becomes unstable in the current setting.

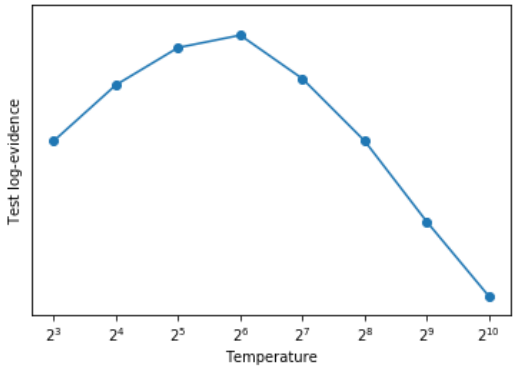


Figure SM 2: MNIST test log-evidence with different temperature.

J Relation to information theoretic model selection methods

In classical machine learning, model selection often relies on information-theoretic measures such as the *Akaike information criterion* (AIC) [1], *Bayesian information criterion* (BIC) [7], and *minimum description length* (MDL) [6]. For example, AIC uses an asymptotic argument to derive an information score based on the KL-divergence between the true and model distribution:

$$\text{AIC}(p(x;\theta),\mathcal{D}) \triangleq 2(\#(\alpha) - \log p_\alpha(\mathcal{D})),$$

where $\#(\cdot)$ denotes the counting measure. Models with smaller information score are preferred, often involving a tradeoff between *model complexity* ($\#\alpha$) and *model evidence* ($\log p_\alpha(x)$). Other information criteria share a similar rationale, which is closely related to the principle of Occam's razor and empirical learning theory. Modern practice in machine learning often employs over-parametrized learners, yielding superb performance not explained by conventional learning theory. The very concept of "model complexity" requires a major overhaul in this modern setting; we refer readers to the work of [8] for some recent advances. Our work focuses the model evidence part that suffers from the pathologies we discussed in earlier sections.

K Detailed experimental setups and additional results

K.1 Toy model

In Figure SM 1 we show the joint density of our toy distribution, and the approximate posterior $q_\beta(z)$ used to evaluate the theoretical bounds.

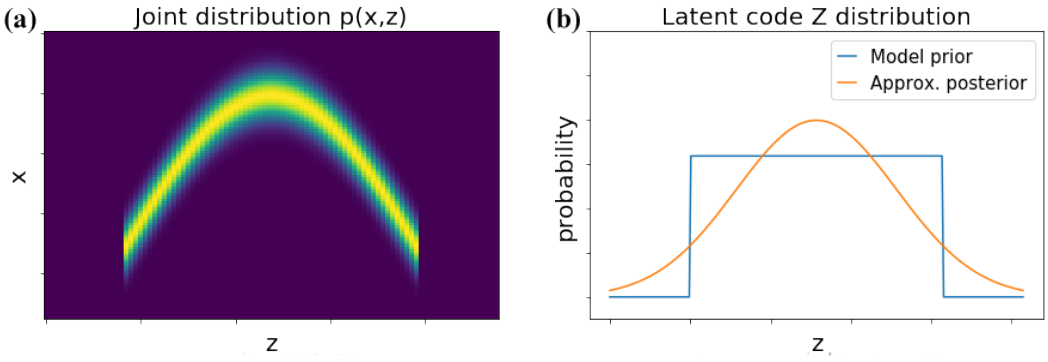


Figure SM 1: Toy distribution used in the theoretical bound experiment.

K.2 Adversarial Variational Bayesian (AVB) experiment

We have set temperature parameter $T = 100$ for the AVB MNIST experiments. A constant smoothing factor 0.3 is used for all experiments. For the AVB models, we have chosen the model with latent dimension 32. Other parameters parameters follow default settings. In Figure SM 2 we show the generated images and feature space interpolation on the MNIST and CelebA dataset.

K.2.1 The adaptive contrast (AC) trick

AC introduces an auxiliary distribution $\tilde{q}_\gamma(z|x)$ with known density expression, and further decompose $r_\beta(x, z)$ as

$$r_\beta(x, z) = \tilde{r}(x, z) + c(x, z),$$

where $\tilde{r}(x, z) \triangleq \log q_\beta(z|x) - \log \tilde{q}_\gamma(z|x)$ and $c(x, z) \triangleq \log \tilde{q}_\gamma(z|x) - \log p(z)$. Here $\tilde{r}(x, z)$ is similarly learned with a density ratio estimator by sampling from $q_\beta(z|x)$ and $\tilde{q}_\gamma(z|x)$, and $c(x, z)$ is computed directly. In our experiments, we use a mean and variance matched Gaussian for $\tilde{q}_\gamma(z|x)$.

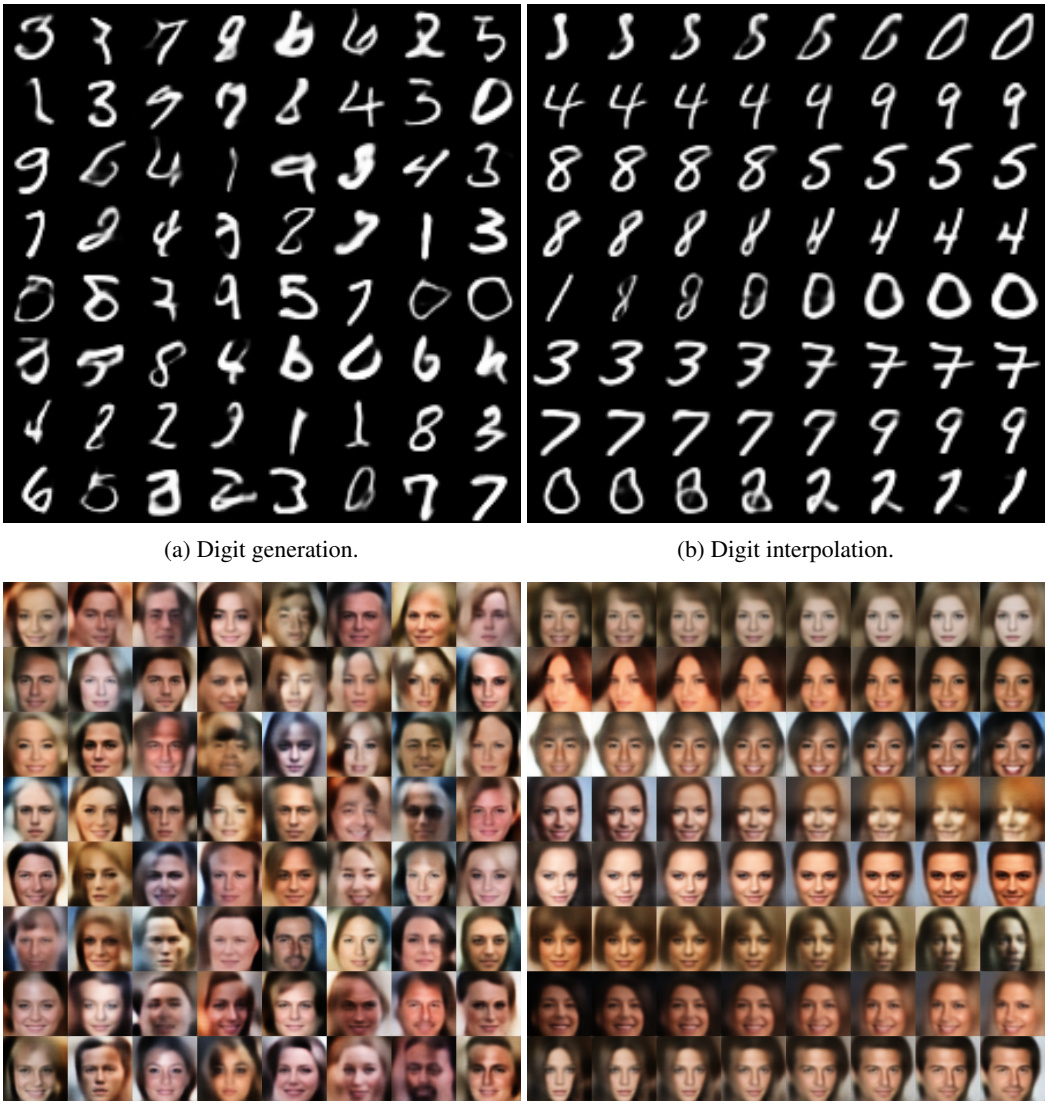
K.3 Convergence rate comparison experiment

In this experiment, we set the number of importance samples to $K = 5$. We used the following neural networks for encoder and decoder as described in Table SM 1.

K.4 Normalizing flow experiment

In our normalizing flow experiments, we consider the *Planar Flows* of the form

$$f_m(z_{m-1}, x; \beta) = z_{m-1} + (u_\beta(x)\cdot h(w_\beta(x)^T z_{m-1} + b_\beta(x))),$$



(a) Digit generation.

(b) Digit interpolation.

(c) Face generation.

(d) Face interpolation.

Figure SM 2: Visual inspection for GLBO.

where $u_\beta(x), w_\beta(x) \in \mathbb{R}^d, b_\beta(x) \in \mathbb{R}$ are functions of x parameterized by β, and $h(u)$ is a activation function, e.g. $\tanh(u)$. For PF we have

$$\log(\left|\det(\nabla_{z_{m-1}} z_m)\right|) = \left|1 + u_\beta^T h'(w_\beta^T z_{m-1} + b) w_\beta\right| .$$

We modified a publicly available implementation of NF from github [2] and set the number of flows to $M = 16$.

K.5 GLBO model selection experiment

In the model-selection experiment, we set $\ell_{\text{lower}} = -90$. This choice is made to make sure the ϕ-evidence score is well defined for all samples , and there is sufficient difference between the ϕ gradient of low-evidence and high-evidence samples.

[2] `https://github.com/abhisheksaurabh1985/vae_nf`

Encoder X to z	Decoder z to X
Input Image X	Input z random noise
4×4 conv. 32 lReLU, stride 2, BN	concat random noise
4×4 conv. 64 lReLU, stride 2, BN	MLP output 1024, lReLU, BN
4×4 conv. 128 lReLU, stride 2, BN	MLP output 8192, lReLU, BN
4×4 conv. 256 lReLU, stride 2, BN	
4×4 conv. 512 lReLU, stride 2, BN	5×5 deconv. 256 lReLU, stride 2, BN
MLP output 512, lReLU	5×5 deconv. 128 lReLU, stride 2, BN
MLP output dim of z, tanh	5×5 deconv. 64 lReLU, stride 2, BN
	5×5 deconv. 3 tanh, stride 2, BN

Table SM 1: Architecture of the models for VAE on CelebA. lReLU is the leaky ReLU with slope 0.1.

	Test RMSE	Test log-likelihood
Dataset	IWVI	IWVI
Boston	$2.85 \pm .42$	$-2.46 \pm .15$
Concrete	$5.17 \pm .32$	$-3.05 \pm .07$
Energy	$0.95. \pm .18$	$-1.66 \pm .05$
Kin8nm	$0.08 \pm .00$	$1.14 \pm .03$
Naval	$0.00 \pm .00$	$4.11 \pm .16$
CCPP	$4.03 \pm .14$	$-2.82 \pm .03$
Winequality	$0.62 \pm .03$	$-0.94 \pm .05$
Yacht	0.95 ± 0.25	$-2.69 \pm .01$
Protein	$4.50 \pm .08$	$-2.90 \pm .01$
Year	$8.81 \pm$ NA	$-3.62 \pm$ NA

Table SM 2: Test RMSE and log-likelihood results for Bayesian neural net regression.

K.6 Bayesian neural net regression (BNN) experiment

We benchmarked with single-layer neural nets are used in this experiment. 100 hidden units are used for 2 large datasets (Protein and Year Predict), and 50 hidden units are used for the other 8 small datasets. Standard Gaussian $\theta \sim \mathcal{N}(0, I)$ is used as our prior for the network weights, and we use Gaussian approximation $\mathcal{N}(\theta; \mu, \sigma^2)$ for the posterior. We repeat the experiments for 20 times on the small datasets, and for 5 times on Protein and 1 time for YearPredict due to computational considerations. The batch size is set to $1,000$ and 100 respectively for large and small datasets. The results for BNN trained with importance-weighted ELBO is reported in Table SM 2.

References

[1] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *ICLR*, 2016.

[3] Adji B Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David M Blei. Variational inference via chi upper bound minimization. In *NIPS*, 2017.

[4] Edwin T Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620, 1957.

[5] Edwin T Jaynes. Information theory and statistical mechanics. ii. *Physical Review*, 108(2):171, 1957.

[6] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[7] Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[8] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.