BIAS-VARIANCE DECOMPOSITION FOR BOLTZMANN MACHINES

Anonymous authors
Paper under double-blind review

ABSTRACT

We achieve bias-variance decomposition for Boltzmann machines using an information geometric formulation. Our decomposition leads to an interesting phenomenon that the variance does not necessarily increase when more parameters are included in Boltzmann machines, while the bias always decreases. Our result gives a theoretical evidence of the generalization ability of deep learning architectures because it provides the possibility of increasing the representation power with avoiding the variance inflation.

1 INTRODUCTION

Why the deep learning architectures can generalize well despite their high representation power with a large number of parameters is one of crucial problems in theoretical deep learning analysis, and there are a number of attempts to solve the problem with focusing on several aspect such as sharpness and robustness (Dinh et al., 2017; Wu et al., 2017; Keskar et al., 2017; Neyshabur et al., 2017; Kawaguchi et al., 2017). However, the complete understanding of this phenomenon is not achieved yet due to the complex structure of deep learning architectures.

To theoretically analyze the generalizability of the architectures, in this paper, we focus on a family of Boltzmann machines (Ackley et al., 1985) including RBMs (Restricted Boltzmann Machines) (Hinton, 2002; Smolensky, 1986) and DBMs (Deep Boltzmann Machines) (Salakhutdinov & Hinton, 2012), the fundamental probabilistic model of deep learning (Hinton et al., 2006), and we firstly present *bias-variance decomposition for Boltzmann machines*. The key to achieve this analysis is to employ an *information geometric formulation* of a hierarchical probabilistic model, which was firstly explored by Amari (2001); Nakahara & Amari (2002); Nakahara et al. (2006). In particular, the recent advances of the formulation by Sugiyama et al. (2016; 2017) enables us to analytically obtain the Fisher information of parameters in Boltzmann machines, which is essential to give the lower bound of variances in bias-variance decomposition.

We show an interesting phenomenon revealed by our bias-variance decomposition: The variance does not necessarily increase while the bias always decreases when we include more parameters in Boltzmann machines, which is caused by its hierarchical structure. Our result indicates the possibility of designing a deep learning architecture that can reduce both of bias and variance, leading to better generalization ability with keeping the representation power.

The remainder of this paper is organized as follows: First we formulate Boltzmann machines using an information geometric formulation in Section 2, which includes the traditional Boltzmann machines (Section 2.2), arbitrary-order Boltzmann machines (Section 2.3), and Boltzmann machines with hidden nodes such as RBMs and DBMs (Section 2.4). Then we present the main result of this paper, bias-variance decomposition for Boltzmann machines, in Section 3 and discuss its property. Finally, we conclude with summarizing the contribution of this paper in Section 4.

2 FORMULATION

To theoretically analyze a family of Boltzmann machines (Ackley et al., 1985) including RBMs (Restricted Boltzmann Machines) (Hinton, 2002; Smolensky, 1986) and DBMs (Deep Boltzmann Machines) (Salakhutdinov & Hinton, 2012), we introduce an information geometric formulation of hierarchical probabilistic models that include the above family of Boltzmann machines.

2.1 PRELIMINARY: LOG-LINEAR MODEL

First we prepare a log-linear probabilistic model on a partial order structure, which has been introduced by Sugiyama et al. (2016; 2017). Let $(S, \leq)$ be a *partially ordered set*, or a *poset* (Gierz et al., 2003), where a *partial order* $\leq$ is a relation between elements in S satisfying the following three properties for all $x, y, z \in S$: (1) $x \leq x$ (reflexivity), (2) $x \leq y$, $y \leq x \Rightarrow x = y$ (antisymmetry), and (3) $x \leq y$, $y \leq z \Rightarrow x \leq z$ (transitivity). We assume that S is always finite and includes the least element (bottom) $\perp \in S$; that is, $\perp \leq x$ for all $x \in S$. We denote $S \setminus \{\perp\}$ by S^+.

We use two functions, the *zeta function* $\zeta \colon S \times S \to \{0, 1\}$ and the *Möbius function* $\mu \colon S \times S \to \mathbb{Z}$. The zeta function ζ is defined as $\zeta(s, x) = 1$ if $s \leq x$ and $\zeta(s, x) = 0$ otherwise, while the Möbius function μ is its convolutional inverse, that is,

$$\sum_{s \in S} \zeta(x, s)\mu(s, y) = \sum_{x \leq s \leq y} \mu(s, y) = \begin{cases} 1 & \text{if } x = y, \\ 0 & \text{otherwise,} \end{cases}$$

which is inductively defined as

$$\mu(x, y) = \begin{cases} 1 & \text{if } x = y, \\ -\sum_{x \leq s < y} \mu(x, s) & \text{if } x < y, \\ 0 & \text{otherwise.} \end{cases}$$

For any functions f, g, and h with the domain S such that

$$g(x) = \sum_{s \in S} \zeta(s, x) f(s) = \sum_{s \leq x} f(s), \quad h(x) = \sum_{s \in S} \zeta(x, s) f(s) = \sum_{s \geq x} f(s),$$

f is uniquely recovered using the Möbius function:

$$f(x) = \sum_{s \in S} \mu(s, x) g(s), \quad f(x) = \sum_{s \in S} \mu(x, s) h(s).$$

This is the *Möbius inversion formula* and is fundamental in enumerative combinatorics (Ito, 1993).

Sugiyama et al. (2017) introduced a *log-linear model* on S, which gives a discrete probability distribution with the structured outcome space $(S, \leq)$. Let P denote a probability distribution that assigns a probability $p(x)$ for each $x \in S$ satisfying $\sum_{x \in S} p(x) = 1$. Each probability $p(x)$ for $x \in S$ is defined as

$$\log p(x) \coloneqq \sum_{s \in S} \zeta(s, x)\theta(s) = \sum_{s \leq x} \theta(s). \tag{1}$$

From the Möbius inversion formula, θ is obtained as

$$\theta(x) = \sum_{s \in S} \mu(s, x) \log p(s). \tag{2}$$

In addition, we introduce $\eta \colon S \to \mathbb{R}$ as

$$\eta(x) \coloneqq \sum_{s \in S} \zeta(x, s) p(s) = \sum_{s \geq x} p(s), \tag{3}$$

$$p(x) = \sum_{s \in S} \mu(x, s)\eta(s). \tag{4}$$

The second equation is from the Möbius inversion formula. Sugiyama et al. (2017) showed that the set of distributions $\boldsymbol{\mathcal{S}} = \{P \mid 0 < p(x) < 1 \text{ and } \sum p(x) = 1\}$ always becomes the *dually flat Riemannian manifold*. This is why two functions θ and η are dual coordinate systems of $\boldsymbol{\mathcal{S}}$ connected with the Legendre transformation, that is,

$$\theta = \nabla\varphi(\eta), \quad \eta = \nabla\psi(\theta)$$

with two convex functions

$$\psi(\theta) \coloneqq -\theta(\perp) = -\log p(\perp), \quad \varphi(\eta) \coloneqq \sum_{x \in S} p(x) \log p(x).$$

Moreover, the Riemannian metric $g(\xi)$ ($\xi = \theta$ or η) such that

$$g(\theta) = \nabla\nabla\psi(\theta), \quad g(\eta) = \nabla\nabla\varphi(\eta),$$

which corresponds to the gradient of θ or η, is given as

$$g_{xy}(\theta) = \frac{\partial \eta(x)}{\partial \theta(y)} = \mathbf{E}\left[\frac{\log p(s)}{\theta(x)}\frac{\log p(s)}{\theta(y)}\right] = \sum_{s\in S}\zeta(x,s)\zeta(y,s)p(s) - \eta(x)\eta(y), \tag{5}$$

$$g_{xy}(\eta) = \frac{\partial \theta(x)}{\partial \eta(y)} = \mathbf{E}\left[\frac{\log p(s)}{\eta(x)}\frac{\log p(s)}{\eta(y)}\right] = \sum_{s\in S}\mu(s,x)\mu(s,y)p(s)^{-1}. \tag{6}$$

for all $x, y \in S^+$. Furthermore, $\boldsymbol{\mathcal{S}}$ is in the exponential family (Sugiyama et al., 2016), where θ coincides with the *natural parameter* and η with the *expectation parameter*.

Let us consider two types of submanifolds:

$$\boldsymbol{\mathcal{S}}_\alpha = \{\, P \in \boldsymbol{\mathcal{S}} \mid \theta(x) = \alpha(x) \text{ for all } x \in \mathrm{dom}(\alpha) \,\},$$
$$\boldsymbol{\mathcal{S}}_\beta = \{\, P \in \boldsymbol{\mathcal{S}} \mid \eta(x) = \beta(x) \text{ for all } x \in \mathrm{dom}(\beta) \,\}$$

specified by two functions α, β with $\mathrm{dom}(\alpha), \mathrm{dom}(\beta) \subseteq S^+$, where the former submanifold $\boldsymbol{\mathcal{S}}_\alpha$ has constraints on θ while the latter $\boldsymbol{\mathcal{S}}_\beta$ has those on η. It is known in information geometry that $\boldsymbol{\mathcal{S}}_\beta$ is e-flat and $\boldsymbol{\mathcal{S}}_\alpha$ is m-flat, respectively (Amari, 2016). Suppose that $\mathrm{dom}(\alpha) \cup \mathrm{dom}(\beta) = S^+$ and $\mathrm{dom}(\alpha) \cap \mathrm{dom}(\beta) = \emptyset$. Then the intersection $\boldsymbol{\mathcal{S}}_\alpha \cap \boldsymbol{\mathcal{S}}_\beta$ is always the singleton, that is, the distribution Q satisfying $Q \in \boldsymbol{\mathcal{S}}_\alpha$ and $Q \in \boldsymbol{\mathcal{S}}_\beta$ always uniquely exists, and the following *Pythagorean theorem* holds:

$$D_{\mathrm{KL}}(P, R) = D_{\mathrm{KL}}(P, Q) + D_{\mathrm{KL}}(Q, R), \tag{7}$$
$$D_{\mathrm{KL}}(R, P) = D_{\mathrm{KL}}(R, Q) + D_{\mathrm{KL}}(Q, P) \tag{8}$$

for any $P \in \boldsymbol{\mathcal{S}}_\alpha$ and $R \in \boldsymbol{\mathcal{S}}_\beta$.

2.2 Standard Boltzmann machines

A *Boltzmann machine* is represented as an undirected graph $G = (V, E)$ with a vertex set $V = \{1, 2, \ldots, n\}$ and an edge set $E \subseteq \{\{i, j\} \mid i, j \in V\}$. The energy function Φ:$\{0,1\}^n \to \mathbb{R}$ of the Boltzmann machine G is defined as

$$\Phi(\boldsymbol{x}; \boldsymbol{\theta}) = -\sum_{i\in V}\theta_i x_i - \sum_{\{i,j\}\in E}\theta_{ij}x_i x_j$$

for each $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n, \theta_{12}, \theta_{13}, \ldots, \theta_{n-1n})$ is a parameter vector for vertices (bias) and edges (weight) such that $\theta_{ij} = 0$ if $\{i, j\} \notin E$. The probability $p(\boldsymbol{x}; \boldsymbol{\theta})$ of the Boltzmann machine G is obtained for each $\boldsymbol{x} \in \{0,1\}^n$ as

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp(-\Phi(\boldsymbol{x}, \boldsymbol{\theta}))}{Z(\boldsymbol{\theta})} \tag{9}$$

with a partition function $Z(\boldsymbol{\theta})$ such that

$$Z(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}\in\{0,1\}^n}\exp(-\Phi(\boldsymbol{x}; \boldsymbol{\theta})) \tag{10}$$

to ensure the condition $\sum_{\boldsymbol{x}\in\{0,1\}^n} p(\boldsymbol{x}) = 1$.

It is clear that a Boltzmann machine is a special case of the log-linear model in Equation (1) with $S = 2^V$, the power set of V, and $\perp = \emptyset$. Let each $x \in S$ be the set of indices of "1" of $\boldsymbol{x} \in \{0,1\}^n$ and $\leq$ be the inclusion relation, that is, $x \leq y$ if and only if $x \subseteq y$. Suppose that

$$B = \left\{\, x \in S^+ \;\middle|\; |x| = 1 \text{ or } x \in E \,\right\}, \tag{11}$$

where $|x|$ is the cardinality of x, which we call a *parameter set*. The Boltzmann distribution in Equations (9) and (10) directly corresponds to the log-linear model in Equation (1):

$$\log p(x) = \sum_{s\in B}\zeta(s,x)\theta(s) - \psi(\theta), \quad \psi(\theta) = -\theta(\perp) = \log Z(\theta).$$

This means that the set of Boltzmann distributions $\boldsymbol{\mathcal{S}}(B)$ that can be represented by a parameter set B is a *submanifold* of $\boldsymbol{\mathcal{S}}$ given as

$$\boldsymbol{\mathcal{S}}(B) \coloneqq \{\, P \in \boldsymbol{\mathcal{S}} \mid \theta(x) = 0 \text{ for all } x \notin B \,\}.$$

Given an empirical distribution $\hat{P}$. Learning of a Boltzmann machine is to find the best approximation of $\hat{P}$ from the Boltzmann distributions $\boldsymbol{\mathcal{S}}(B)$, which is formulated as a minimization problem of the KL (Kullback–Leibler) divergence:

$$\min_{P_B \in \boldsymbol{\mathcal{S}}(B)} D_{\mathrm{KL}}(\hat{P}, P_B) = \min_{P_B \in \boldsymbol{\mathcal{S}}(B)} \sum_{s \in S} \hat{p}(s) \log \frac{\hat{p}(s)}{p_B(s)}.$$

This is equivalent to maximize the log-likelihood $L(P_B) = N \sum_{s \in S} \hat{p}(s) \log p_B(s)$ with the sample size N. Since we have

$$\begin{aligned}
\frac{\partial}{\partial \theta_B(x)} D_{\mathrm{KL}}(\hat{P}, P_B) &= \frac{\partial}{\partial \theta_B(x)} \sum_{s \in S} \hat{p}(s) \log p_B(s) \\
&= \frac{\partial}{\partial \theta_B(x)} \sum_{s \in S} \left(\hat{p}(s) \sum_{\perp < u \leq s} \theta_B(u) \right) - \frac{\partial}{\partial \theta_B(x)} \psi(\theta_B) \sum_{s \in S} \hat{p}(s) \\
&= \hat{\eta}(x) - \eta_B(x),
\end{aligned}$$

the KL divergence $D_{\mathrm{KL}}(\hat{P}, P_B)$ is minimized when $\hat{\eta}(x) = \eta_B(x)$ for all $x \in B$, which is well known as the *learning equation of Boltzmann machines*. Thus the minimizer $P_B \in \boldsymbol{\mathcal{S}}(B)$ of the KL divergence $D_{\mathrm{KL}}(\hat{P}, P_B)$ is the distribution given as

$$\begin{cases} \eta_B(x) = \hat{\eta}(x) & \text{if } x \in B \cup \{\perp\}, \\ \theta_B(x) = 0 & \text{otherwise.} \end{cases}$$

This distribution P_B is known as *m-projection* of $\hat{P}$ onto $\boldsymbol{\mathcal{S}}(B)$ (Sugiyama et al., 2017), which is unique and always exists as $\boldsymbol{\mathcal{S}}$ has the dually flat structure with respect to (θ, η).

2.3 Arbitrary-Order Boltzmann Machines

The parameter set B is fixed in Equation (11) in the traditional Boltzmann machines, but our log-linear formulation allows us to include or remove any element in $S^+ = 2^V \setminus \{\perp\}$ as a parameter. This attempt was partially studied by Sejnowski (1986); Min et al. (2014) that include higher order interactions to increase the representation power of Boltzmann machines.

Let $B_1, B_2, \ldots, B_m$ be a sequence of parameter sets such that

$$B_1 \subseteq B_2 \subseteq \cdots \subseteq B_{m-1} \subseteq B_m = S.$$

Since we have a hierarchy of submanifolds

$$\boldsymbol{\mathcal{S}}(B_1) \subseteq \boldsymbol{\mathcal{S}}(B_2) \subseteq \cdots \subseteq \boldsymbol{\mathcal{S}}(B_{m-1}) \subseteq \boldsymbol{\mathcal{S}}(B_m) = \boldsymbol{\mathcal{S}},$$

we obtain the decreasing sequence of KL divergences:

$$D_{\mathrm{KL}}(\hat{P}, P_{B_1}) \geq D_{\mathrm{KL}}(\hat{P}, P_{B_2}) \geq \cdots \geq D_{\mathrm{KL}}(\hat{P}, P_{B_{m-1}}) \geq D_{\mathrm{KL}}(\hat{P}, P_{B_m}) = 0,$$

where each $P_{B_i} = \operatorname{argmin}_{P \in \boldsymbol{\mathcal{S}}(B_i)} D_{\mathrm{KL}}(\hat{P}, P)$, the best approximation of $\hat{P}$ using B_i.

2.4 Boltzmann Machines with Hidden nodes

Another strategy to increase the representation power is to use *hidden nodes* (Le Roux & Bengio, 2008). A Boltzmann machine with hidden nodes is represented as $G = (V \cup H, E)$, where V and H correspond to visible and hidden nodes, respectively, and the resulting domain $S = 2^{V \cup H}$. In particular, restricted Boltzmann machines (RBMs) (Smolensky, 1986; Hinton, 2002) are often used in applications, where the edge set is given as

$$E = \{\, \{i, j\} \mid i \in V, j \in H \,\}$$

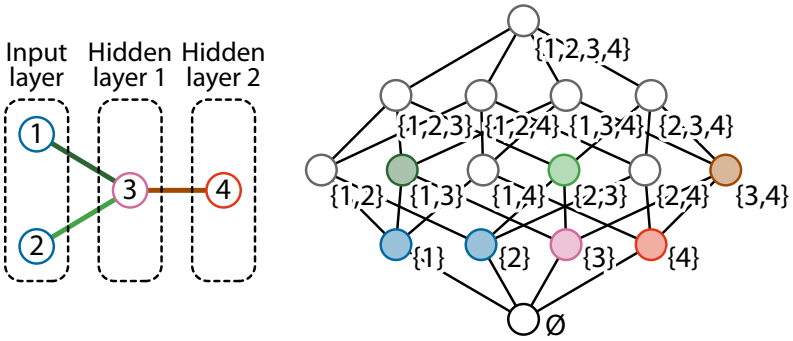

Figure 1: An example of a deep Boltzmann machine (left) with an input layer (visible nodes) $V = \{1, 2\}$ with two hidden layers $H_1 = \{3\}$ and $H_2 = \{4\}$, and the corresponding domain set $S^{V \cup H}$ (right). In the right-hand side, the colored objects $\{1\}, \{2\}, \{3\}, \{4\}, \{1, 2\}, \{2, 3\}$, and $\{3, 4\}$ denote the parameter set B, which correspond to nodes and edges of the DBM in the left-hand side.

Moreover, in a deep Boltzmann machine (DBM) (Salakhutdinov & Hinton, 2009; 2012), which is the beginning of the recent trend of deep learning (LeCun et al., 2015; Goodfellow et al., 2016), the hidden nodes H is divided into k disjoint subsets (layers) $H_1, H_2, \ldots, H_k$ and

$$E = \{ \{i, j\} \mid i \in H_{l-1}, j \in H_l, l \in \{1, \ldots, k\} \},$$

where $V = H_0$ for simplicity.

Let $S = 2^V$ and $S' = 2^{V \cup H}$ and $\boldsymbol{\mathcal{S}}$ and $\boldsymbol{\mathcal{S}}'$ be the set of distributions with the domains S and S', respectively. In both cases of RBMs and DBMs, we have

$$B = \{ x \in S' \mid |x| = 1 \text{ or } x \in E \},$$

(see Figure 1) and the set of Boltzmann distributions is obtained as

$$\boldsymbol{\mathcal{S}}'(B) = \left\{ P \in \boldsymbol{\mathcal{S}}' \;\middle|\; \theta(x) = 0 \text{ for all } x \notin B \right\}.$$

Since the objective of learning Boltzmann machines with hidden nodes is MLE (maximum likelihood estimation) with respect to the marginal probabilities of the visible part, the target empirical distribution $\hat{P} \in \boldsymbol{\mathcal{S}}$ is extended to the submanifold $\boldsymbol{\mathcal{S}}'(\hat{P})$ such that

$$\boldsymbol{\mathcal{S}}'(\hat{P}) = \left\{ P \in \boldsymbol{\mathcal{S}}' \;\middle|\; \eta(x) = \hat{\eta}(x) \text{ for all } x \in S \right\},$$

and the process of learning Boltzmann machines with hidden nodes is formulated as double minimization of the KL divergence such that

$$\min_{P \in \boldsymbol{\mathcal{S}}'(\hat{P}), P_B \in \boldsymbol{\mathcal{S}}'(B)} D_{\mathrm{KL}}(P, P_B). \tag{12}$$

Since two submanifolds $\boldsymbol{\mathcal{S}}'(B)$ and $\boldsymbol{\mathcal{S}}'(\hat{P})$ are e-flat and m-flat, respectively, it is known that the EM-algorithm can obtain the global optimum of Equation (12) (Amari, 2016, Section 8.1.3), which was first analyzed by Amari et al. (1992). Since this computation is infeasible due to combinatorial explosion of the domain $S' = 2^{V \cup H}$, a number of approximation methods such as Gibbs sampling have been proposed (Salakhutdinov & Hinton, 2012).

3 Bias-Variance Decomposition

Here we present the main result of this paper, *bias-variance decomposition for Boltzmann machines*. We focus on the expectation of the squared KL divergence $\mathbf{E}[D_{\mathrm{KL}}(P^*, \hat{P}_B)^2]$ from the true (unknown) distribution P^* to the MLE $\hat{P}_B$ of an empirical distribution $\hat{P}$ by a Boltzmann machine with a parameter set B, and decompose it using information geometric properties.

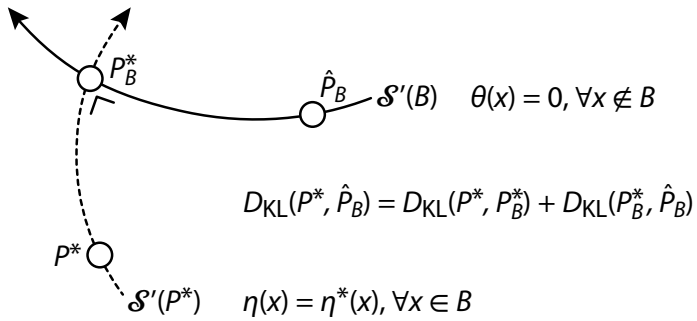

Figure 2: Illustration of the Pythagorean theorem.

Theorem 1 (Bias-variance decomposition of the KL divergence)**.** *Given a Boltzmann machine with a parameter set B. Let $P^* \in \boldsymbol{\mathcal{S}}$ be the true (unknown) distribution, $P_B^*, \hat{P}_B \in \boldsymbol{\mathcal{S}}(B)$ be the MLEs of P^* and an empirical distribution $\hat{P}$, respectively. If the Boltzmann machine includes hidden nodes, let P^* be a distribution in the submanifold $\boldsymbol{\mathcal{S}}'(P^*)$ achieved in Equation* (12) *and $P_B^*, \hat{P}_B \in \boldsymbol{\mathcal{S}}'(B)$. We have*

$$\mathbf{E}\Big[D_{\mathrm{KL}}(P^*, \hat{P}_B)^2\Big] = D_{\mathrm{KL}}(P^*, P_B^*)^2 + \mathbf{E}\Big[D_{\mathrm{KL}}(P_B^*, \hat{P}_B)^2\Big] \geq \underbrace{D_{\mathrm{KL}}(P^*, P_B^*)^2}_{bias^2} + \underbrace{\mathrm{var}(P_B^*, B)}_{variance},$$

$$\mathrm{var}(P_B^*, B) = \frac{1}{N} \sum_{x \in S} p_B^*(x)^{-1} \left(\sum_{s \in B} \mu(x, s) \eta_B^*(s) \right)^2,$$

where the equality holds when the sample size $N \to \infty$.

Proof. From the Pythagorean theorem illustrated in Figure 2,

$$\begin{aligned}
\mathbf{E}\Big[D_{\mathrm{KL}}(P^*, \hat{P}_B)^2\Big] &= \mathbf{E}\Big[\big(D_{\mathrm{KL}}(P^*, P_B^*) + D_{\mathrm{KL}}(P_B^*, \hat{P}_B)\big)^2\Big] \\
&= \mathbf{E}\Big[D_{\mathrm{KL}}(P^*, P_B^*)^2 + 2D_{\mathrm{KL}}(P^*, P_B^*) D_{\mathrm{KL}}(P_B^*, \hat{P}_B) + D_{\mathrm{KL}}(P_B^*, \hat{P}_B)^2\Big] \\
&= D_{\mathrm{KL}}(P^*, P_B^*)^2 + 2D_{\mathrm{KL}}(P^*, P_B^*) \mathbf{E}\Big[D_{\mathrm{KL}}(P_B^*, \hat{P}_B)\Big] + \mathbf{E}\Big[D_{\mathrm{KL}}(P_B^*, \hat{P}_B)^2\Big].
\end{aligned}$$

Since $\hat{\theta}_B(s)$ is an unbiased estimator of $\theta_B^*(s)$ for every $s \in S$, it holds that

$$\mathbf{E}\big[D_{\mathrm{KL}}(P_B^*, \hat{P}_B)\big] = \mathbf{E}\left[\sum_{x \in S} p_B^*(x) \log \frac{p_B^*(x)}{\hat{p}_B(x)}\right] = \sum_{x \in S} p_B^*(x) \sum_{s \leq x} \mathbf{E}\left[\theta_B^*(s) - \hat{\theta}_B(s)\right] = 0.$$

Hence we have

$$\mathbf{E}\Big[D_{\mathrm{KL}}(P^*, \hat{P}_B)^2\Big] = D_{\mathrm{KL}}(P^*, P_B^*)^2 + \mathbf{E}\Big[D_{\mathrm{KL}}(P_B^*, \hat{P}_B)^2\Big]. \tag{13}$$

The second term is

$$\begin{aligned}
&\mathbf{E}\Big[D_{\mathrm{KL}}(P_B^*, \hat{P}_B)^2\Big] \\
&= \mathbf{E}\left[\left(\sum_{x \in S} p_B^*(x) \log \frac{p_B^*(x)}{\hat{p}_B(x)}\right)^2\right] = \mathbf{E}\left[\sum_{x \in S} \sum_{y \in S} p_B^*(x) p_B^*(y) \log \frac{p_B^*(x)}{\hat{p}_B(x)} \log \frac{p_B^*(y)}{\hat{p}_B(y)}\right] \\
&= \mathbf{E}\left[\sum_{x \in S} \sum_{y \in S} p_B^*(x) p_B^*(y) \left(\sum_{s \in B, s \leq x} \Big(\theta_B^*(s) - \hat{\theta}_B(s)\Big) - \Big(\psi(\theta_B^*) - \psi(\hat{\theta}_B)\Big)\right)\right. \\
&\qquad\qquad \left.\left(\sum_{u \in B, u \leq y} \Big(\theta_B^*(u) - \hat{\theta}_B(u)\Big) - \Big(\psi(\theta_B^*) - \psi(\hat{\theta}_B)\Big)\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\left[\sum_{x\in S}\sum_{y\in S}\sum_{s\in B}\sum_{u\in B} p_B^*(x)p_B^*(y)\zeta(s,x)\zeta(u,y)\left(\theta_B^*(s)-\hat{\theta}_B(s)\right)\left(\theta_B^*(u)-\hat{\theta}_B(u)\right)\right] \\
&\qquad -2\,\mathbf{E}\left[\sum_{x\in S} p_B^*(x)\sum_{s\in B, s\leq x}\left(\theta_B^*(s)-\hat{\theta}_B(s)\right)\left(\psi(\theta_B^*)-\psi(\hat{\theta}_B)\right)\right] + \mathbf{E}\left[\left(\psi(\theta_B^*)-\psi(\hat{\theta}_B)\right)^2\right] \\
&= \sum_{x\in S}\sum_{y\in S}\sum_{s\in B}\sum_{u\in B} p_B^*(x)p_B^*(y)\zeta(s,x)\zeta(u,y)\mathbf{E}\left[\left(\theta_B^*(s)-\hat{\theta}_B(s)\right)\left(\theta_B^*(u)-\hat{\theta}_B(u)\right)\right] \\
&\quad -2\sum_{x\in S}\sum_{s\in B} p_B^*(x)\zeta(s,x)\mathbf{E}\left[\left(\theta_B^*(s)-\hat{\theta}_B(s)\right)\left(\psi(\theta_B^*)-\psi(\hat{\theta}_B)\right)\right] + \mathbf{E}\left[\left(\psi(\theta_B^*)-\psi(\hat{\theta}_B)\right)^2\right].
\end{aligned}
$$

Since $\theta(s)$ and $\psi(\theta) = -\theta(\perp)$ are orthogonal for all $s \in S$, that is, we have from Equation (5)

$$
\mathbf{E}\left[\frac{\partial \log p(s)}{\partial \theta(s)}\frac{\partial \log p(s)}{\partial \theta(\perp)}\right] = \sum_{s\in S}\zeta(x,s)p(s) - \eta(x) = \eta(x) - \eta(x) = 0,
$$

it follows that

$$
\begin{aligned}
\mathbf{E}\left[\left(\theta_B^*(s)-\hat{\theta}_B(s)\right)\left(\psi(\theta_B^*)-\psi(\hat{\theta}_B)\right)\right] &= \mathbf{E}\left[\hat{\theta}_B(s)\psi(\hat{\theta}_B)\right] - \theta_B^*(s)\psi(\theta_B^*) \\
&= \theta_B^*(s)\psi(\theta_B^*) - \theta_B^*(s)\psi(\theta_B^*) = 0, \\
\mathbf{E}\left[\left(\psi(\theta_B^*)-\psi(\hat{\theta}_B)\right)^2\right] &= \mathbf{E}\left[\psi(\hat{\theta}_B)^2\right] - \psi(\theta_B^*)^2 = \psi(\theta_B^*)^2 - \psi(\theta_B^*)^2 = 0.
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\mathbf{E}\Big[D_{\mathrm{KL}}(P_B^*, \hat{P}_B)^2\Big] \\
&= \sum_{x\in S}\sum_{y\in S}\sum_{s\in B}\sum_{u\in B} p_B^*(x)p_B^*(y)\zeta(s,x)\zeta(u,y)\mathbf{E}\left[\left(\theta_B^*(s)-\hat{\theta}_B(s)\right)\left(\theta_B^*(u)-\hat{\theta}_B(u)\right)\right].
\end{aligned}
$$

Using the Cramér–Rao bound,

$$
\mathbf{E}\left[\left(\theta_B^*(s)-\hat{\theta}_B(s)\right)\left(\theta_B^*(u)-\hat{\theta}_B(u)\right)\right] \geq \frac{1}{N}\sum_{w\in S}\mu(w,s)\mu(w,u)p_B^*(w)^{-1},
$$

where the Fisher information is given in Equation (6), we obtain

$$
\mathbf{E}\Big[D_{\mathrm{KL}}(P_B^*, \hat{P}_B)^2\Big] \geq \frac{1}{N}\sum_{x\in S}\sum_{y\in S}\sum_{s\in B}\sum_{u\in B}\sum_{w\in S} p_B^*(x)p_B^*(y)\zeta(s,x)\zeta(u,y)\mu(w,s)\mu(w,u)p_B^*(w)^{-1},
$$

where the equality holds if the sample size $N \to \infty$. Here we have

$$
\begin{aligned}
&\frac{1}{N}\sum_{x\in S}\sum_{y\in S}\sum_{s\in B}\sum_{u\in B}\sum_{w\in S} p_B^*(x)p_B^*(y)\zeta(s,x)\zeta(u,y)\mu(w,s)\mu(w,u)p_B^*(w)^{-1} \\
&= \frac{1}{N}\sum_{w\in S} p_B^*(w)^{-1}\sum_{s\in B}\sum_{u\in B}\mu(w,s)\mu(w,u)\sum_{x\in S}\zeta(s,x)p_B^*(x)\sum_{y\in S}\zeta(u,y)p_B^*(y) \\
&= \frac{1}{N}\sum_{w\in S} p_B^*(w)^{-1}\sum_{s\in B}\sum_{u\in B}\mu(w,s)\mu(w,u)\eta_B^*(s)\eta_B^*(u) \\
&= \frac{1}{N}\sum_{w\in S} p_B^*(w)^{-1}\left(\sum_{s\in B}\mu(w,s)\eta_B^*(s)\right)^2
\end{aligned}
$$

Therefore, from Equation (13), it follows that

$$
\mathbf{E}\Big[D_{\mathrm{KL}}(P^*, \hat{P}_B)^2\Big] = D_{\mathrm{KL}}(P^*, P_B^*)^2 + \mathbf{E}\Big[D_{\mathrm{KL}}(P_B^*, \hat{P}_B)^2\Big]
$$

$$\geq D_{\mathrm{KL}}(P^*, P_B^*)^2 + \frac{1}{N}\sum_{x\in S} p_B^*(x)^{-1}\left(\sum_{s\in B}\mu(x,s)\eta_B^*(s)\right)^2$$
$$= D_{\mathrm{KL}}(P^*, P_B^*)^2 + \mathrm{var}(P_B^*, B)$$

with the equality holding when $N \to \infty$. □

Let $B, B' \subseteq S' = 2^{V\cup H}$ such that $B \subseteq B'$, that is, B' has more parameters than B. Then it is trivial that the bias is always reduced, that is,

$$D_{\mathrm{KL}}(P^*, P_B^*) \geq D_{\mathrm{KL}}(P^*, P_{B'}^*)$$

because $\boldsymbol{\mathcal{S}}(B) \subseteq \boldsymbol{\mathcal{S}}(B')$. However, this monotonicity does not always hold for the variance. We illustrate this non-monotonicity in the following example. Let $S = 2^V$ with $V = \{1,2,3\}$ and assume that P^* is the uniform distribution, i.e., $p(x) = 1/8$ for all $x \in S$. Suppose that $B = \{\{1\}\}$ and $B' = \{\{1\},\{1,2\},\{1,3\}\}$. We have

$$\mathrm{var}(P^*, B)\cdot N = p(\emptyset)^{-1}\Big(\mu(\emptyset,\{1\})\eta(\{1\})\Big)^2 + p(\{1\})^{-1}\Big(\mu(\{1\},\{1\})\eta(\{1\})\Big)^2$$
$$= 8(-1\cdot 1/2)^2 + 8(1\cdot 1/2)^2 = 4,$$

and

$$\begin{aligned}
&\mathrm{var}(P^*, B')\cdot N \\
&= p(\emptyset)^{-1}\Big(\mu(\emptyset,\{1\})\eta(\{1\}) + \mu(\emptyset,\{1,2\})\eta(\{1,2\}) + \mu(\emptyset,\{1,3\})\eta(\{1,3\})\Big)^2 \\
&\quad + p(\{1\})^{-1}\Big(\mu(\{1\},\{1\})\eta(\{1\}) + \mu(\{1\},\{1,2\})\eta(\{1,2\}) + \mu(\{1\},\{1,3\})\eta(\{1,3\})\Big)^2 \\
&\quad + p(\{1,2\})^{-1}\Big(\mu(\{1,2\},\{1,2\})\eta(\{1,2\})\Big)^2 + p(\{1,3\})^{-1}\Big(\mu(\{1,3\},\{1,3\})\eta(\{1,3\})\Big)^2 \\
&= 8(-(1/2)+(1/4)+(1/4))^2 + 8((1/2)-(1/4)-(1/4))^2 + 8(1\cdot(1/4))^2 + 8(1\cdot(1/4))^2 \\
&= 0 + 0 + 1/2 + 1/2 = 1.
\end{aligned}$$

Thus the variance decreases when we include more parameters in B'. This interesting property, *non-monotonicity of the variance with respect to the growth of parameter sets*, comes from the hierarchical structure of S realized by the möbius function μ.

4 Conclusion

In this paper, we have firstly achieved bias-variance decomposition of the KL divergence for Boltzmann machines using the information geometric formulation of hierarchical probability distributions. Our model includes various types of Boltzmann machines such as the traditional Boltzmann machines, generalized Boltzmann machines with arbitrary order interactions, and Boltzmann machines hidden nodes such as RBMs and DBMs. Our bias-variance decomposition reveals the non-monotonicity of the variance with respect to increase of the number of parameters. This result indicates that it is possible to reduce both bias and variance when we include more parameters in the deep learning architectures. To solve the open problem of the generalizability of the deep learning architectures, our finding can be fundamental for further theoretical development.

References

D. H. Ackley, G. E. Hinton, and T. J. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive science*, 9(1):147–169, 1985.

S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, 2001.

S. Amari. *Information Geometry and Its Applications*. Springer, 2016.

S. Amari, K. Kurata, and H. Nagaoka. Information geometry of Boltzmann machines. *IEEE Transactions on Neural Networks*, 3(2):260–271, 1992.

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1019–1028, 2017.

G. Gierz, K. H. Hofmann, K. Keimel, J. D. Lawson, M. Mislove, and D. S. Scott. *Continuous Lattices and Domains*. Cambridge University Press, 2003.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

K. Ito (ed.). *Encyclopedic Dictionary of Mathematics*. The MIT Press, 2 edition, 1993.

K. Kawaguchi, L. P. Kaelbling, and Y. Bengio. Generalization in deep learning. *arXiv:1710.05468*, 2017.

N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *Proceedings of 5th International Conference on Learning Representations*, 2017.

N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. *Neural computation*, 20(6):1631–1649, 2008.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

M. R. Min, X. Ning, C. Cheng, and M. Gerstein. Interpretable sparse high-order Boltzmann machines. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pp. 614–622, 2014.

H. Nakahara and S. Amari. Information-geometric measure for neural spikes. *Neural Computation*, 14(10):2269–2316, 2002.

H. Nakahara, S. Amari, and B. J. Richmond. A comparison of descriptive models of a single spike train by information-geometric measure. *Neural computation*, 18(3):545–568, 2006.

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30*, 2017.

R. Salakhutdinov and G. E. Hinton. Deep Boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pp. 448–455, 2009.

R. Salakhutdinov and G. E. Hinton. An efficient learning procedure for deep Boltzmann machines. *Neural Computation*, 24(8):1967–2006, 2012.

T. J. Sejnowski. Higher-order Boltzmann machines. In *AIP Conference Proceedings*, volume 151, pp. 398–403, 1986.

P. Smolensky. Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart, J. L. McClelland, and PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, pp. 194–281. MIT Press, 1986.

M. Sugiyama, H. Nakahara, and K. Tsuda. Information decomposition on structured space. In *IEEE International Symposium on Information Theory*, pp. 575–579, 2016.

M. Sugiyama, H. Nakahara, and K. Tsuda. Tensor balancing on statistical manifold. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3270–3279, 2017.

L. Wu, Z. Zhu, and W. E. Towards understanding generalization of deep learning: Perspective of loss landscapes. In *ICML 2017 Workshop on Principled Approaches to Deep Learning*, 2017.