ON FEATURE DIVERSITY IN ENERGY-BASED MODELS

Firas Laakom
Faculty of Information Technology
Tampere University
Tampere, Finland
firas.laakom@tuni.fi

Jenni Raitoharju
Programme for Environmental Information
Finnish Environment Institute
Jyväskylä, Finland
jenni.raitoharju@syke.fi

Alexandros Iosifidis
Department of Electrical and Computer Engineering
Aarhus University
Aarhus, Denmark
ai@ece.au.dk

Moncef Gabbouj
Faculty of Information Technology
Tampere University
Tampere, Finland
moncef.gabbouj@tuni.fi

ABSTRACT

Energy-based learning is a powerful learning paradigm that encapsulates various discriminative and generative approaches. An energy-based model (EBM) is typically formed of one (or many) inner-models which learn a combination of the different features to generate an energy mapping for each input configuration. In this paper, we focus on the diversity of the produced feature set. We extend the probably approximately correct (PAC) theory of EBMs and analyze the effect of the diversity on the performance of EBMs. We derive generalization bounds for various learning contexts, i.e., regression, classification, and implicit regression, with different energy functions and we show that indeed increasing the diversity of the feature set can consistently decrease the gap between the true and empirical expectation of the energy and boosts the performance of the model.

1 INTRODUCTION

The energy-based learning paradigm was first proposed by LeCun et al. (2006) as an alternative to probabilistic graphical models (Koller & Friedman, 2009). As their name suggests, energy-based models (EBMs) map each input 'configuration' to a single scalar, called the 'energy'. In the learning phase, the parameters of the model are optimized to associate the desired configurations with small energy values and the undesired ones with higher energy values (Kumar et al., 2019; Song & Ermon, 2019; Yu et al., 2020; Nash & Durkan, 2019; Meng et al., 2020; Arbel et al., 2021). In the inference phase, given an incomplete input configuration, the energy surface is explored to find the remaining variables which yield the lowest energy. EBMs encapsulate solutions to several supervised (LeCun et al., 2006; Fang & Liu, 2016) and unsupervised learning problems (Ranzato et al., 2007b; Haarnoja et al., 2017; Parshakova et al., 2019; Deng et al., 2020; Bakhtin et al., 2021) and provide a common theoretical framework for many learning models, including traditional discriminative (Zhai et al., 2016; Grathwohl et al., 2019; Li et al., 2020; LeCun et al., 2006; Teh et al., 2003) and generative (Zhao et al., 2016; Dai et al., 2017; Ranzato et al., 2007a; Che et al., 2020; Khalifa et al., 2020; Arbel et al., 2021) approaches.

Formally, let us denote the energy function by $E(W, \boldsymbol{X}, \boldsymbol{Y})$, where W represents the model parameters to be optimized during training and $\boldsymbol{X}, \boldsymbol{Y}$ are sets of variables. Figure 1 illustrates how classification, regression, and implicit regression can be expressed as EBMs. In Figure 1 (a), a regression scenario is presented. The input $\boldsymbol{X}$, e.g., an image, is transformed using an inner model $G_W(\boldsymbol{X})$ and its distance, D, to the second input $\boldsymbol{Y}$ is computed yielding the energy function. A valid energy function in this case can be the L_1 or the L_2 distance. In the binary classification case (Figure 1 (b)), the energy can be defined as $E(W, \boldsymbol{X}, \boldsymbol{Y}) = -\boldsymbol{Y} G_W(\boldsymbol{X})$. In the inference phase, given an input $\boldsymbol{X}$, the label $\boldsymbol{Y}^*$ can be obtained by solving the following optimization problem:

$$\boldsymbol{Y}^* = \arg\min_{\boldsymbol{Y}} E(W, \boldsymbol{X}, \boldsymbol{Y}). \tag{1}$$

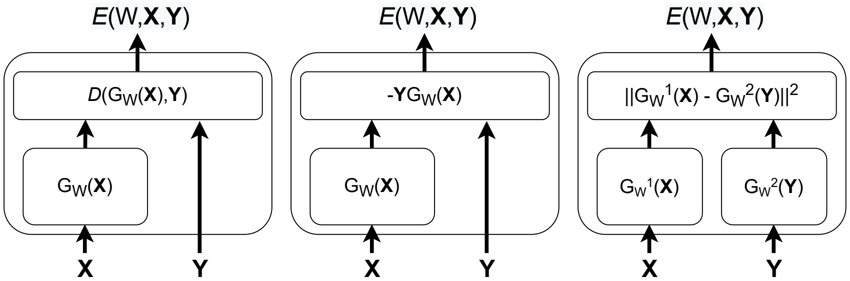

Figure 1: An illustration of energy-based models used to solve (a) a regression problem (b) a binary classification problem (c) an implicit regression problem.

An EBM typically relies on an inner model, i.e., $G_w(\boldsymbol{X})$, to generate the desired energy landscape (LeCun et al., 2006). Depending on the problem at hand, this function can be constructed as a linear projection, a kernel method, or a neural network and its parameters are optimized in a data-driven manner in the training phase. Formally, $G_w(\boldsymbol{X})$ can be written as

$$G_W(\boldsymbol{X}) = \sum_{i}^{D} w_i \phi_i(\boldsymbol{X}), \tag{2}$$

where $\{\phi_1(\cdot), \cdots, \phi_D(\cdot)\}$ is the feature set, which can be hand-crafted, separately trained from unlabeled data (Zhang & LeCun, 2017), or modeled by a neural network and optimized in the training phase of the EBM model (Du & Mordatch, 2019). In the rest of the paper, we assume that the inner models G_W defined in the energy-based learning system (Figure 1) are obtained as a weighted sum of different features as expressed in equation 2.

In (Zhang, 2013), it was shown that simply minimizing the empirical energy over the training data does not theoretically guarantee the minimization of the expected value of the true energy. Thus, developing and motivating novel regularization techniques is required (Zhang & LeCun, 2017). We argue that the quality of this feature set, i.e., $\{\phi_1(\cdot), \cdots, \phi_D(\cdot)\}$, plays a critical role in the overall performance of the global model. In this work, we extend the theoretical analysis of (Zhang, 2013) and focus on the 'diversity' of this set and its effect on the generalization ability of the EBM models. Intuitively, it is clear that a less correlated set of intermediate representations is richer and thus able to capture more complex patterns in the input. Thus, it is important to avoid redundant features for achieving a better performance. However, a theoretical analysis is missing. We start by quantifying the diversity of a set. To this end, we introduce ϑ-diversity:

Definition 1. *(ϑ-diversity) A set of feature functions, $\{\phi_1(\cdot), \cdots, \phi_D(\cdot)\}$ is called ϑ-diverse, if there exists a constant $\vartheta \in \mathbb{R}$, such that for every input $\boldsymbol{X}$ we have*

$$\sum_{i \neq j}^{D} (\phi_i(\boldsymbol{X}) - \phi_j(\boldsymbol{X}))^2 > \vartheta \tag{3}$$

with a high probability τ.

Intuitively, if two feature maps $\phi_i(\cdot)$ and $\phi_j(\cdot)$ are different, then with high probability they have different outputs for the same input. However, if for example the features are extracted using a neural network with a ReLU activation function, then there is a high probability that some of the features associated with the input will be zero. Thus, defining a lower bound for the pair-wise diversity directly is impractical. To this end, we quantify diversity as the lower-bound over the sum of the pair-wise distances of the feature maps as expressed in equation 3. ϑ measures the diversity of a set.

In machine learning context, diversity has been explored in ensemble learning (Li et al., 2012; Yu et al., 2011), sampling (Derezinski et al., 2019; Bıyık et al., 2019; Gartrell et al., 2019), ranking (Yang et al., 2019; Gan et al., 2020), pruning (Kondo & Yamauchi, 2014; He et al., 2019; Singh et al., 2020; Lee et al., 2020), and neural networks (Xie et al., 2015; 2017). In Xie et al. (2015; 2017), it was shown theoretically and experimentally that employing a diversity strategy over the weights of a neural network using the mutual angles improves the generalization ability of the model. In this work, we explore a new line of research, where diversity is defined over the feature maps directly, using the ϑ-diversity, in the context of energy-based learning. We theoretically study the generalization ability of EBMs in different learning contexts, i.e., regression, classification, implicit regression, and we derive new generalization bounds using the ϑ-diversity providing theoretical guarantees that a diverse set of features indeed improves the generalization ability of the model. The contributions of this paper can be summarized as follows:

- We explore a new line of research, where diversity is defined over the features representing the input data and not over the model's parameters. To this end, we introduce ϑ-diversity as a quantification of the diversity of a given feature set.
- We extend the theoretical analysis (Zhang, 2013) and study the effect of the diversity of the feature set on the generalization of the energy-based models (EBMs).
- We derive approximation bounds for the expectation of the true energy in different learning contexts, i.e., regression, classification, and implicit regression, using different energy functions. Our analysis consistently shows that increasing the diversity of the feature set can boost the performance of an energy based model.

2 PAC-LEARNING OF EBMS WITH ϑ-DIVERSITY

In this section, we derive a qualitative justification for ϑ-diversity using probably approximately correct (PAC) learning (Valiant, 1984). The PAC-based theory for standard energy based models has been established in (Zhang, 2013). Based on the Rademacher complexity (Bartlett & Mendelson, 2002), several EBMs learning guarantees have been shown. In Lemma 1, we present the principal PAC-learning bound for energy functions with finite outputs.

Definition 2. *(Bartlett & Mendelson, 2002) For a given dataset with m samples $\boldsymbol{S} = \{\boldsymbol{x}_i, y_i\}_{i=1}^m$ generated by a distribution $\mathcal{D}$ and for a model space $\mathcal{F} : \mathcal{X} \rightarrow \mathbb{R}$ with a single dimensional output, the empirical Rademacher complexity $\mathcal{R}_m(\mathcal{F})$ of the set $\mathcal{F}$ is defined as follows:*

$$\mathcal{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{N} \sigma_i f(\boldsymbol{x}_i) \right], \tag{4}$$

where the Rademacher variables $\sigma = \{\sigma_1, \cdots, \sigma_N\}$ are independent uniform random variables in $\{-1, 1\}$.

Lemma 1. *(Zhang, 2013) For a well-defined energy function $E(h, \boldsymbol{x}, \boldsymbol{y})$ over hypothesis class $\mathcal{H}$, input set $\mathcal{X}$ and output set $\mathcal{Y}$ (LeCun et al., 2006), the following holds for all h in $\mathcal{H}$ with a probability of at least $1 - \delta$*

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}[E(h, \boldsymbol{x}, \boldsymbol{y})] \leq \frac{1}{m} \sum_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{S}} E(h, \boldsymbol{x}, \boldsymbol{y}) + 2\mathcal{R}_m(\mathcal{E}) + M\sqrt{\frac{\log(2/\delta)}{2m}}, \tag{5}$$

where $\mathcal{E}$ is the energy function class defined as $\mathcal{E} = \{E(h, \boldsymbol{x}, \boldsymbol{y}) | h \in \mathcal{H}\}$, $\mathcal{R}_m(\mathcal{E})$ is its Rademacher complexity, and M is the upper bound of $\mathcal{E}$.

Lemma 1 provides a generalization bound for energy-based models with well-defined (non-negative) and bounded energy. The expected energy is bounded using the sum of three terms: The first term is the empirical expectation of energy over the training data, the second term depends on the Rademacher complexity of the energy class, and the third term involves the number of the training data m and the upper-bound of the energy function M. This shows that merely minimizing the empirical expectation of energy, i.e., the first term, may not yield a good approximation of the true expectation. In (Zhang & LeCun, 2017), it has been shown that regularization using unlabeled data reduces the second and third terms, thus, leading to better generalization. In this work, we express these two terms using the ϑ-diversity and show that employing a diversity strategy may also decrease the gap between the true and empirical expectation of the energy. In Section 2.1, we consider the special case of regression and derive two bounds relative to two energy functions based on L_1 and L_2 distances. In Section 2.2, we derive the bound relative to the binary classification task using as energy function $E(\boldsymbol{W}, \boldsymbol{x}, y) = -\mathrm{y}G_{\boldsymbol{W}}(\boldsymbol{x})$ (LeCun et al., 2006). In Section 2.3, we consider the case of implicit regression, which encapsulates different learning problems such as metric learning, generative models, and denoising (LeCun et al., 2006). For this case, we use the L_2 distance between the inner models as the energy function.

2.1 Regression Task

Regression can be formulated as an energy-based learning problem (Figure 1 (a)) using the inner model $G_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x})$. We also suppose that the feature set is well-defined over the input domain $\mathcal{X}$, i.e., $\forall \boldsymbol{x} \in \mathcal{X}\ ||\Phi(\boldsymbol{x})||_2 \leq A$. The two valid energy functions which can be used for regression are: $E_2(\boldsymbol{W}, \boldsymbol{x}, y) = \frac{1}{2}||G_{\boldsymbol{W}}(\boldsymbol{x}) - y||_2^2$ and $E_1(\boldsymbol{W}, \boldsymbol{x}, y) = ||G_{\boldsymbol{W}}(\boldsymbol{x}) - y||_1$ (LeCun et al., 2006). Theorem 1 and Theorem 2 express the special cases of Lemma 1 using the ϑ-diversity of the feature set $\{\phi_1(\cdot), \cdots, \phi_D(\cdot)\}$.

Theorem 1. *For the energy function $E(h, \boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}||G_{\boldsymbol{W}}(\boldsymbol{x}) - y||_2^2$, over the input set $\mathcal{X} \in \mathbb{R}^N$, hypothesis class $\mathcal{H} = \{G_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x}) \mid \Phi \in \mathcal{F},\ \forall \boldsymbol{x}\ ||\Phi(\boldsymbol{x})||_2 \leq A\}$, and output set $\mathcal{Y} \subset \mathbb{R}$, if the feature set $\{\phi_1(\cdot), \cdots, \phi_D(\cdot)\}$ is ϑ-diverse with a probability τ, then with a probability of at least $(1-\delta)\tau$, the following holds for all h in $\mathcal{H}$*

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}[E(h, \boldsymbol{x}, \boldsymbol{y})] \leq \frac{1}{m} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \boldsymbol{S}} E(h, \boldsymbol{x}, \boldsymbol{y}) + 8D||\boldsymbol{w}||_\infty (||\boldsymbol{w}||_\infty \sqrt{DA^2 - \vartheta^2} + B) \mathcal{R}_m(\mathcal{F})$$
$$+ (||\boldsymbol{w}||_\infty \sqrt{DA^2 - \vartheta^2} + B)^2 \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (6)$$

where B is the upper-bound of $\mathcal{Y}$, i.e., $y \leq B, \forall y \in \mathcal{Y}$.

Theorem 2. *For the energy function $E(h, \boldsymbol{x}, \boldsymbol{y}) = ||G_{\boldsymbol{W}}(\boldsymbol{x}) - y||_1$, over the input set $\mathcal{X} \in \mathbb{R}^N$, hypothesis class $\mathcal{H} = \{G_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x}) \mid \Phi \in \mathcal{F},\ \forall \boldsymbol{x}\ ||\Phi(\boldsymbol{x})||_2 \leq A\}$, and output set $\mathcal{Y} \subset \mathbb{R}$, if the feature set $\{\phi_1(\cdot), \cdots, \phi_D(\cdot)\}$ is ϑ-diverse with a probability τ, then with a probability of at least $(1-\delta)\tau$, the following holds for all h in $\mathcal{H}$*

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}[E(h, \boldsymbol{x}, \boldsymbol{y})] \leq \frac{1}{m} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \boldsymbol{S}} E(h, \boldsymbol{x}, \boldsymbol{y}) + 4D||\boldsymbol{w}||_\infty \mathcal{R}_m(\mathcal{F})$$
$$+ 2(||\boldsymbol{w}||_\infty \sqrt{DA^2 - \vartheta^2} + B) \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (7)$$

where B is the upper-bound of $\mathcal{Y}$, i.e., $y \leq B, \forall y \in \mathcal{Y}$.

The proofs are available in the Appendix. We note that, in Theorem 1 and Theorem 2, we consistently find that the bound of the true expectation of the energy is a decreasing function with respect to ϑ. This proves that that for the regression task employing a diversity strategy can improve the generalization performance of the energy-based model.

2.2 Two-Class Classifier

Here, we consider the problem of binary classification, as illustrated in Figure 1 (b). Using the same assumption as in regression for the inner model, i.e., $G_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x})$, energy function of $E(\boldsymbol{W}, \boldsymbol{x}, y) = -\mathrm{y} G_{\boldsymbol{W}}(\boldsymbol{x})$ (LeCun et al., 2006), and the ϑ-diversity of the feature set,we express Lemma 1 for this specific configuration in Theorem 3.

Theorem 3. *For the energy function $E(h, \boldsymbol{x}, \boldsymbol{y}) = -\mathrm{y} G_{\boldsymbol{W}}(\boldsymbol{x})$, over the input set $\mathcal{X} \in \mathbb{R}^N$, hypothesis class $\mathcal{H} = \{G_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x}) \mid \Phi \in \mathcal{F},\ \forall \boldsymbol{x}\ ||\Phi(\boldsymbol{x})||_2 \leq A\}$, and output set $\mathcal{Y} \subset \mathbb{R}$, if the feature set $\{\phi_1(\cdot), \cdots, \phi_D(\cdot)\}$ is ϑ-diverse with a probability τ, then with a probability of at least $(1-\delta)\tau$, the following holds for all h in $\mathcal{H}$*

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}[E(h, \boldsymbol{x}, \boldsymbol{y})] \leq \frac{1}{m} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \boldsymbol{S}} E(h, \boldsymbol{x}, \boldsymbol{y}) + 4D||\boldsymbol{w}||_\infty \mathcal{R}_m(\mathcal{F})$$
$$+ ||\boldsymbol{w}||_\infty \sqrt{DA^2 - \vartheta^2} \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (8)$$

The proof is available in the Appendix. Similar to the regression task, we note that the upper-bound of the true expectation is a decreasing function with respect to the diversity term. Thus, a more diverse feature set, i.e., higher ϑ, has a lower upper-bound for the true energy.

2.3 Implicit Regression

In this section, we consider the problem of implicit regression. This is a general formulation of a different set of problems such as metric learning, where the goal is to learn a distance function between two domains, image denoising, or object detection as illustrated in (LeCun et al., 2006). This form of EBM (Figure 1 (c)) has two inner models, $G^1_{\boldsymbol{W}}(\cdot)$ and $G^2_{\boldsymbol{W}}(\cdot)$, which can be equal or different according to the problem at hand. Here, we consider the general case, where the two models correspond to two different combinations of different features, i.e., $G^1_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D^{(1)}} w^1_i \phi^1_i(\boldsymbol{x})$ and $G^{(2)}_{\boldsymbol{W}}(\boldsymbol{y}) = \sum_{i=1}^{D^{(2)}} w^2_i \phi^2_i(\boldsymbol{y})$. Thus, we have a different ϑ-diversity term for each set. The final result is presented in Theorem 4.

Theorem 4. *For the energy function $E(h, \boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}||G^{(1)}_{\boldsymbol{W}}(\boldsymbol{x}) - G^{(2)}_{\boldsymbol{W}}(\boldsymbol{y})||_2^2$, over the input set $\mathcal{X} \in \mathbb{R}^N$, hypothesis class $\mathcal{H} = \{G^{(1)}_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D^{(1)}} w^{(1)}_i \phi^{(1)}_i(\boldsymbol{x}) = \boldsymbol{w}^{(1)^T} \Phi^{(1)}(\boldsymbol{x}), G^{(2)}_{\boldsymbol{W}}(\boldsymbol{y}) = \sum_{i=1}^{D^{(2)}} w^{(2)}_i \phi^{(2)}_i(\boldsymbol{y}) = \boldsymbol{w}^{(2)^T} \Phi^{(2)}(\boldsymbol{y}) \mid \Phi^{(1)} \in \mathcal{F}_1,\ \Phi^{(2)} \in \mathcal{F}_2,\ \forall \boldsymbol{x}\ ||\Phi^{(1)}(\boldsymbol{x})||_2 \leq A^{(1)},\ \forall \boldsymbol{y}\ ||\Phi^{(2)}(\boldsymbol{y})||_2 \leq A^{(2)}\}$, and output set $\mathcal{Y} \subset \mathbb{R}^N$, if the feature set $\{\phi^{(1)}_1(\cdot), \cdots, \phi^{(1)}_{D^{(1)}}(\cdot)\}$ is $\vartheta^{(1)}$-diverse with a probability τ_1 and the feature set $\{\phi^{(2)}_1(\cdot), \cdots, \phi^{(2)}_{D^{(2)}}(\cdot)\}$ is $\vartheta^{(2)}$-diverse with a probability τ_2, then with a probability of at least $(1-\delta)\tau_1\tau_2$, the following holds for all h in $\mathcal{H}$*

$$\begin{aligned}\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}[E(h,\boldsymbol{x},\boldsymbol{y})] \leq \frac{1}{m} \sum_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{S}} & E(h,\boldsymbol{x},\boldsymbol{y}) \\ & + 8(\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2})\big(D^{(1)}||\boldsymbol{w}^{(1)}||_\infty \mathcal{R}_m(\mathcal{F}_1) + D^{(2)}||\boldsymbol{w}^{(2)}||_\infty \mathcal{R}_m(\mathcal{F}_2)\big) \\ & \qquad + 2\big(\mathcal{J}_1 + \mathcal{J}_2\big)\sqrt{\frac{\log(2/\delta)}{2m}}, \end{aligned} \tag{9}$$

where $\mathcal{J}_1 = ||\boldsymbol{w}^{(1)}||^2_\infty \Big(D^{(1)} {A^{(1)}}^2 - {\vartheta^{(1)}}^2\Big)$ and $\mathcal{J}_2 = ||\boldsymbol{w}^{(2)}||^2_\infty \Big(D^{(2)} {A^{(2)}}^2 - {\vartheta^{(2)}}^2\Big)$.

The proof of Theorem 4 is available in the Appendix. The upper-bound of the energy model depends on the diversity variable of both feature sets. Moreover, we note that the bound for the implicit regression decreases proportionally to ϑ^2, as opposed to the classification case for example, where the bound is proportional to ϑ.

We note that the theory developed in our paper (Theorems 1 to 4) is agnostic to the loss function (LeCun et al., 2006) or the optimization strategy used (Kumar et al., 2019; Song & Ermon, 2019; Yu et al., 2020). We show that increasing the diversity of the features consistently decreases the upper-bound of the true expectation of the energy and, thus, can boost the generalization performance of the energy-based model. We note that our analysis is independent of how the features are obtained, e.g., handcrafted or optimized. In fact, in the recent state-of-the-art EBMs (Khalifa et al., 2020; Bakhtin et al., 2021; Nash & Durkan, 2019; Yu et al., 2020), the features are typically parameterized using a deep learning model and optimized during the training. Thus, our theory suggests the use of a diversity strategy, for example in the form of a regularization as in (Cogswell et al., 2016), to avoid learning redundant features can improve the performance of the model and decrease the gap between the expectation of the true and the empirical energy.

3 Conclusion

The energy-based learning is a powerful learning paradigm that encapsulates various discriminative and generative systems. An EBM is typically formed of one (or many) inner models which learn a combination of different features to generate an energy mapping for each input configuration. In this paper, we introduced the feature diversity concept, i.e., ϑ-diversity, and we used it to extend the PAC theory of EBMs. We derived different generalization bounds for various learning contexts, i.e., regression, classification, and implicit regression, with different energy functions and we consistently found that increasing the diversity of the feature set can improve the approximation error of the true expectation of the energy function. We also note that our theory is independent of the loss function or the training strategy used to optimize the parameters of the EBM.

Future directions include developing practical strategies to promote the diversity of the feature set in case the features are optimized following a data-driven process, like the training phase of a neural network.

REFERENCES

Michael Arbel, Liang Zhou, and Arthur Gretton. Generalized energy based models. In *International Conference on Learning Representations*, 2021.

Anton Bakhtin, Yuntian Deng, Sam Gross, Myle Ott, Marc'Aurelio Ranzato, and Arthur Szlam. Residual energy-based models for text. *Journal of Machine Learning Research*, 22(40):1–41, 2021.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, pp. 463–482, 2002.

Erdem Bıyık, Kenneth Wang, Nima Anari, and Dorsa Sadigh. Batch active learning using determinantal point processes. *arXiv preprint arXiv:1906.07975*, 2019.

Tong Che, Ruixiang Zhang, Jascha Sohl-Dickstein, Hugo Larochelle, Liam Paull, Yuan Cao, and Yoshua Bengio. Your gan is secretly an energy-based model and you should use discriminator driven latent sampling. *arXiv preprint arXiv:2003.06060*, 2020.

Michael Cogswell, Faruk Ahmed, Ross B. Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. In *International Conference on Learning Representations*, 2016.

Zihang Dai, Amjad Almahairi, Philip Bachman, Eduard Hovy, and Aaron Courville. Calibrating energy-based generative adversarial networks. *arXiv preprint arXiv:1702.01691*, 2017.

Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. Residual energy-based models for text generation. *arXiv preprint arXiv:2004.11714*, 2020.

Michal Derezinski, Daniele Calandriello, and Michal Valko. Exact sampling of determinantal point processes with sublinear time preprocessing. In *Advances in Neural Information Processing Systems*, pp. 11546–11558, 2019.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Yi Fang and Mengwen Liu. A unified energy-based framework for learning to rank. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, pp. 171–180, 2016.

Lu Gan, Diana Nurbakova, Léa Laporte, and Sylvie Calabretto. Enhancing recommendation diversity using determinantal point processes on knowledge graphs. In *Conference on Research and Development in Information Retrieval*, pp. 2001–2004, 2020.

Mike Gartrell, Victor-Emmanuel Brunel, Elvis Dohmatob, and Syrine Krichene. Learning nonsymmetric determinantal point processes. In *Advances in Neural Information Processing Systems*, pp. 6718–6728, 2019.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pp. 1352–1361. PMLR, 2017.

Yang He, Ping Liu, Ziwei Wang, Zhilan Hu, and Yi Yang. Filter pruning via geometric median for deep convolutional neural networks acceleration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4340–4349, 2019.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation. *arXiv preprint arXiv:2012.11635*, 2020.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Yusuke Kondo and Koichiro Yamauchi. A dynamic pruning strategy for incremental learning on a budget. In *International Conference on Neural Information Processing*, pp. 295–303. Springer, 2014.

Rithesh Kumar, Sherjil Ozair, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*, 2019.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Seunghyun Lee, Byeongho Heo, Jung-Woo Ha, and Byung Cheol Song. Filter pruning and re-initialization via latent space clustering. *IEEE Access*, 8:189587–189597, 2020.

Nan Li, Yang Yu, and Zhi-Hua Zhou. Diversity regularized ensemble pruning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 330–345, 2012.

Shuang Li, Yilun Du, Gido M van de Ven, Antonio Torralba, and Igor Mordatch. Energy-based models for continual learning. *arXiv preprint arXiv:2011.12216*, 2020.

Chenlin Meng, Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. Autoregressive score matching. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020*, 2020.

Charlie Nash and Conor Durkan. Autoregressive energy machines. In *International Conference on Machine Learning*, pp. 1735–1744. PMLR, 2019.

Tetiana Parshakova, Jean-Marc Andreoli, and Marc Dymetman. Distributional reinforcement learning for energy-based sequential models. *arXiv preprint arXiv:1912.08517*, 2019.

Marc Ranzato, Christopher Poultney, Sumit Chopra, Yann LeCun, et al. Efficient learning of sparse representations with an energy-based model. *Advances in neural information processing systems*, 19:1137, 2007a.

Marc'Aurelio Ranzato, Y-Lan Boureau, Sumit Chopra, and Yann LeCun. A unified energy-based framework for unsupervised learning. In *Artificial Intelligence and Statistics*, pp. 371–379. PMLR, 2007b.

Pravendra Singh, Vinay Kumar Verma, Piyush Rai, and Vinay Namboodiri. Leveraging filter correlations for deep model compression. In *The IEEE Winter Conference on Applications of Computer Vision*, pp. 835–844, 2020.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf`.

Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4(Dec):1235–1260, 2003.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Michael M Wolf. Mathematical foundations of supervised learning, 2018.

Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, pp. 1216–1224. PMLR, 2017.

Pengtao Xie, Yuntian Deng, and Eric Xing. On the generalization error bounds of neural networks under diversity-inducing mutual angular regularization. *arXiv preprint arXiv:1511.07110*, 2015.

Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. Balanced ranking with diversity constraints. In *International Joint Conference on Artificial Intelligence*, pp. 6035–6042, 2019.

Lantao Yu, Yang Song, Jiaming Song, and Stefano Ermon. Training deep energy-based models with f-divergence minimization. In *International Conference on Machine Learning*, pp. 10957–10967. PMLR, 2020.

Yang Yu, Yu-Feng Li, and Zhi-Hua Zhou. Diversity regularized machine. In *International Joint Conference on Artificial Intelligence*, 2011.

Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *International Conference on Machine Learning*, pp. 1100–1109. PMLR, 2016.

Xiang Zhang. *Pac-learning for energy-based models*. PhD thesis, Citeseer, 2013.

Xiang Zhang and Yann LeCun. Universum prescription: Regularization using unlabeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

4 Appendix

4.1 Proof of Theorem 1

Lemma 2. *With a probability of at least τ, we have*

$$\sup_{\boldsymbol{x},h} |h(\boldsymbol{x})| \leq \sqrt{\mathcal{J}}, \tag{10}$$

where $\mathcal{J} = ||\boldsymbol{w}||_\infty^2 \Big(DA^2 - \vartheta^2 \Big)$ *and* $A = \sup_{\boldsymbol{x}} ||\phi(\boldsymbol{x})||_2$,

Proof.

$$\begin{aligned} h^2(\boldsymbol{x}) = \left(\sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) \right)^2 &\leq \left(\sum_{i=1}^{D} ||\boldsymbol{w}||_\infty \phi_m(\boldsymbol{x}) \right)^2 = ||\boldsymbol{w}||_\infty^2 \left(\sum_{i=1}^{D} \phi_i(\boldsymbol{x}) \right)^2 \\ &= ||\boldsymbol{w}||_\infty^2 \left(\sum_{i,j} \phi_i(\boldsymbol{x}) \phi_j(\boldsymbol{x}) \right) = ||\boldsymbol{w}||_\infty^2 \left(\sum_i \phi_i(\boldsymbol{x})^2 + \sum_{i \neq j} \phi_i(\boldsymbol{x}) \phi_j(\boldsymbol{x}) \right) \end{aligned} \tag{11}$$

We have $||\Phi(\boldsymbol{x})||_2 \leq A$. For the first term in equation 11, we have $\sum_m \phi_m(\boldsymbol{x})^2 \leq A^2$. By using the identity $\phi_m(\boldsymbol{x})\phi_n(\boldsymbol{x}) = \frac{1}{2} \left(\phi_m(\boldsymbol{x})^2 + \phi_n(\boldsymbol{x})^2 - (\phi_m(\boldsymbol{x}) - \phi_n(\boldsymbol{x}))^2 \right)$, the second term can be rewritten as

$$\sum_{m \neq n} \phi_m(\boldsymbol{x}) \phi_n(\boldsymbol{x}) = \frac{1}{2} \sum_{m \neq n} \left(\phi_m(\boldsymbol{x})^2 + \phi_n(\boldsymbol{x})^2 - \Big(\phi_m(\boldsymbol{x}) - \phi_n(\boldsymbol{x}) \Big)^2 \right). \tag{12}$$

In addition, we have with a probability τ, $\frac{1}{2} \sum_{m \neq n} ||\phi_m(\boldsymbol{x}) - \phi_n(\boldsymbol{x})||_2 \geq \vartheta$. Thus, we have with a probability at least τ:

$$\sum_{m \neq n} \phi_m(\boldsymbol{x}) \phi_n(\boldsymbol{x}) \leq \frac{1}{2} (2(D-1)A^2 - 2\vartheta^2) = (D-1)A^2 - \vartheta^2. \tag{13}$$

By putting everything back to equation 11, we have with a probability τ,

$$h^2(\boldsymbol{x}) \leq ||\boldsymbol{w}||_\infty^2 \Big(A^2 + (D-1)A^2 - \vartheta^2 \Big) = ||\boldsymbol{w}||_\infty^2 \Big(DA^2 - \vartheta^2 \Big) = \mathcal{J}. \tag{14}$$

Thus, with a probability τ,

$$\sup_{\boldsymbol{x},h} |h(\boldsymbol{x})| \leq \sqrt{\sup_{\boldsymbol{x},h} h(\boldsymbol{x})^2} \leq \sqrt{\mathcal{J}}. \tag{15}$$

□

Lemma 3. *With a probability of at least τ, we have*

$$\sup_{\boldsymbol{x},y,f} |E(h(\boldsymbol{x}), y)| \leq (\sqrt{\mathcal{J}} + B)^2. \tag{16}$$

Proof. We have $\sup_{\boldsymbol{x},y,h} |h(\boldsymbol{x}) - y| \leq 2 \sup_{\boldsymbol{x},y,h} (|h(\boldsymbol{x})| + |y|) = 2(\sqrt{\mathcal{J}} + B)$. Thus $sup_{x,y,h} |E(h(x), y)| \leq (\sqrt{\mathcal{J}} + B)^2$. □

Lemma 4. *With a probability of at least τ, we have*

$$\mathcal{R}_m(\mathcal{E}) \leq 4D ||\boldsymbol{w}||_\infty (\sqrt{\mathcal{J}} + B) \mathcal{R}_m(\mathcal{F}) \tag{17}$$

Proof. Using the decomposition property of the Rademacher complexity (if ϕ is a L-Lipschitz function, then $\mathcal{R}_m(\phi(\mathcal{A})) \leq L\mathcal{R}_m(\mathcal{A})$) and given that $E(\cdot, y) = ||. - y||^2$ is K-Lipschitz with a constant $K = sup_{\boldsymbol{x},y,h} ||h(\boldsymbol{x}) - y|| \leq 2(\sqrt{\mathcal{J}} + B)$, we have $\mathcal{R}_m(\mathcal{E}) \leq K\mathcal{R}_m(\mathcal{F}) \leq 2(\sqrt{\mathcal{J}} + B)\mathcal{R}_m(\mathcal{H})$, where $\mathcal{H} = \{G_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) \mid ||\boldsymbol{w}||_1 \leq D ||\boldsymbol{w}||_\infty \}$. Next, similar to the proof of Theorem 2.10 in (Wolf, 2018), we note that $\sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) \in (D ||\boldsymbol{w}||_\infty) conv(\mathcal{F} + -(\mathcal{F})) := \mathcal{G}$, where $conv$ denotes the convex hull and $\mathcal{F}$ is the set of ϕ functions. Thus, $\mathcal{R}_m(\mathcal{H}) \leq \mathcal{R}_m(\mathcal{G}) = D ||\boldsymbol{w}||_\infty \mathcal{R}_m(conv(\mathcal{F} + (-\mathcal{F})) = D ||\boldsymbol{w}||_\infty \mathcal{R}_m(\mathcal{F} + (-\mathcal{F})) = 2D ||\boldsymbol{w}||_\infty \mathcal{R}_m(\mathcal{F})$. □

Theorem 1 For the energy function $E(h, \boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}||G_{\boldsymbol{W}}(\boldsymbol{x}) - y||_2^2$, over the input set $\mathcal{X} \in \mathbb{R}^N$, hypothesis class $\mathcal{H} = \{G_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x}) \mid \forall \boldsymbol{x} \; ||\Phi(\boldsymbol{x})||_2 \leq A\}$, and output set $\mathcal{Y} \subset \mathbb{R}$, if the feature set $\{\phi_1(\cdot), \cdots, \phi_D(\cdot)\}$ is ϑ-diverse with a probability τ, then with a probability of at least $(1-\delta)\tau$, the following holds for all h in $\mathcal{H}$

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}[E(h, \boldsymbol{x}, \boldsymbol{y})] \leq \frac{1}{m} \sum_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{S}} E(h, \boldsymbol{x}, \boldsymbol{y}) + 8D||\boldsymbol{w}||_\infty (||\boldsymbol{w}||_\infty \sqrt{DA^2 - \vartheta^2} + B)\mathcal{R}_m(\mathcal{F})$$
$$+ (||\boldsymbol{w}||_\infty \sqrt{DA^2 - \vartheta^2} + B)^2 \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (18)$$

where B is the upper-bound of $\mathcal{Y}$, i.e., $y \leq B, \forall y \in \mathcal{Y}$,

Proof. We replace the variables in Lemma 1 using Lemma 3 and Lemma 4. □

4.2 Proof of Theorem 2

Lemma 5. *With a probability of at least τ, we have*

$$\sup_{\boldsymbol{x},y,f} |E(h(\boldsymbol{x}), y)| \leq 2(\sqrt{\mathcal{J}} + B). \quad (19)$$

Proof. We have $\sup_{\boldsymbol{x},y,h} |h(\boldsymbol{x}) - y| \leq 2 \sup_{\boldsymbol{x},y,h}(|h(\boldsymbol{x})| + |y|) = 2(\sqrt{\mathcal{J}} + B)$. □

Lemma 6. *With a probability of at least τ, we have*

$$\mathcal{R}_m(\mathcal{E}) \leq 2D||\boldsymbol{w}||_\infty \mathcal{R}_m(\mathcal{F}) \quad (20)$$

Proof. $|.|$ is 1-Lipschitz, Thus $\mathcal{R}_m(\mathcal{E}) \leq \mathcal{R}_m(\mathcal{H})$. □

Theorem 2 For the energy function $E(h, \boldsymbol{x}, \boldsymbol{y}) = ||G_{\boldsymbol{W}}(\boldsymbol{x}) - y||_1$, over the input set $\mathcal{X} \in \mathbb{R}^N$, hypothesis class $\mathcal{H} = \{G_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D} w_i \phi_i(\boldsymbol{x}) = \boldsymbol{w}^T \Phi(\boldsymbol{x}) \mid \forall \boldsymbol{x} \; ||\Phi(\boldsymbol{x})||_2 \leq A\}$, and output set $\mathcal{Y} \subset \mathbb{R}$, if the feature set $\{\phi_1(\cdot), \cdots, \phi_D(\cdot)\}$ is ϑ-diverse with a probability τ, then with a probability of at least $(1-\delta)\tau$, the following holds for all h in $\mathcal{H}$

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}[E(h, \boldsymbol{x}, \boldsymbol{y})] \leq \frac{1}{m} \sum_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{S}} E(h, \boldsymbol{x}, \boldsymbol{y}) + 4D||\boldsymbol{w}||_\infty \mathcal{R}_m(\mathcal{F})$$
$$+ 2(||\boldsymbol{w}||_\infty \sqrt{DA^2 - \vartheta^2} + B) \sqrt{\frac{\log(2/\delta)}{2m}}, \quad (21)$$

Proof. We replace the variables in Lemma 1 using Lemma 5 and Lemma 6. □

4.3 Proof of Theorem 3

Lemma 7. *With a probability of at least τ, we have*

$$\sup_{\boldsymbol{x},y,f} |E(h(\boldsymbol{x}), y)| \leq \sqrt{\mathcal{J}} \quad (22)$$

Proof. We have $sup - \mathrm{y} G_{\boldsymbol{W}}(\boldsymbol{x}) \leq sup|G_{\boldsymbol{W}}(\boldsymbol{x})| \leq \sqrt{\mathcal{J}}$ □

Lemma 8. *With a probability of at least τ, we have*

$$\mathcal{R}_m(\mathcal{E}) \leq 2D||\boldsymbol{w}||_\infty \mathcal{R}_m(\mathcal{F}) \quad (23)$$

Proof. We note that for $y \in \{-1, 1\}$, σ and $-y\sigma$ follow the same distribution. Thus, we have $\mathcal{R}_m(\mathcal{E}) = \mathcal{R}_m(\mathcal{H})$. Next, we note that $\mathcal{R}_m(\mathcal{H}) \leq 2D||\boldsymbol{w}||_\infty \mathcal{R}_m(\mathcal{F})$ □

Theorem 3 For a well-defined energy function $E(h, \boldsymbol{x}, \boldsymbol{y})$ (LeCun et al., 2006), over hypothesis class $\mathcal{H}$, input set $\mathcal{X}$ and output set $\mathcal{Y}$, if it has upper-bound M, then with a probability of at least $1-\delta$, the following holds for all h in $\mathcal{H}$

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}[E(h,\boldsymbol{x},\boldsymbol{y})] \leq \frac{1}{m}\sum_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{S}} E(h,\boldsymbol{x},\boldsymbol{y}) + 4D||\boldsymbol{w}||_\infty \mathcal{R}_m(\mathcal{F}) + ||\boldsymbol{w}||_\infty \sqrt{DA^2-\vartheta^2}\sqrt{\frac{\log(2/\delta)}{2m}}, \quad (24)$$

Proof. We replace the variables in Lemma 1 using Lemma 7 and Lemma 8. □

4.4 Proof of Theorem 4

Lemma 9. *With a probability of at least $\tau_1\tau_2$, we have*

$$\sup_{\boldsymbol{x},y,f} |E(h(\boldsymbol{x}),y)| \leq 2\Big(\mathcal{J}_1 + \mathcal{J}_2\Big) \quad (25)$$

Proof. We have $||G^{(1)}_{\boldsymbol{W}}(\boldsymbol{x}) - G^{(2)}_{\boldsymbol{W}}(\boldsymbol{y})||_2^2 \leq 2(||G^{(1)}_{\boldsymbol{W}}(\boldsymbol{x})||_2^2 + ||G^{(2)}_{\boldsymbol{W}}(\boldsymbol{y})||_2^2)$. Similar to Theorem 1, we have $\sup ||G^{(1)}_{\boldsymbol{W}}(\boldsymbol{x})||_2^2 \leq ||\boldsymbol{w}^{(1)}||_\infty^2 \Big(D^{(1)}{A^{(1)}}^2 - {\vartheta^{(1)}}^2\Big) = \mathcal{J}_1$ and $\sup ||G^{(2)}_{\boldsymbol{W}}(\boldsymbol{y})||_2^2 \leq ||\boldsymbol{w}^{(2)}||_\infty^2 \Big(D^{(2)}{A^{(2)}}^2 - {\vartheta^{(2)}}^2\Big) = \mathcal{J}_2$ □

Lemma 10. *With a probability of at least $\tau_1\tau_2$, we have*

$$\mathcal{R}_m(\mathcal{E}) \leq 4(\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2})\big(D^{(1)}||\boldsymbol{w}^{(1)}||_\infty \mathcal{R}_m(\mathcal{F}_1) + D^{(2)}||\boldsymbol{w}^{(2)}||_\infty \mathcal{R}_m(\mathcal{F}_2)\big) \quad (26)$$

Proof. Let f be the square function, i.e., $f(x) = \frac{1}{2}x^2$ and $\mathcal{E}_0 = \{G^{(1)}_{\boldsymbol{W}}(x) - G^{(2)}_{\boldsymbol{W}}(y) \mid x \in \mathcal{X}, y \in \mathcal{Y}\}$. We have $\mathcal{E} = f(\mathcal{E}_0 + (-\mathcal{E}_0))$. f is Lipschitz over the input space, with a constant L bounded by $\sup_{x,\boldsymbol{W}} G^{(1)}_{\boldsymbol{W}}(x) + \sup_{y,\boldsymbol{W}} G^{(2)}_{\boldsymbol{W}}(y) \leq \sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2}$. Thus, we have $\mathcal{R}_m(\mathcal{E}) \leq (\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2})\mathcal{R}_m(\mathcal{E}_0 + (-\mathcal{E}_0)) \leq 2(\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2})\mathcal{R}_m(\mathcal{E}_0)$. Next, we note that $\mathcal{R}_m(\mathcal{E}_0) = \mathcal{R}_m(\mathcal{H}_1 + (-\mathcal{H}_2)) = \mathcal{R}_m(\mathcal{H}_1) + \mathcal{R}_m(\mathcal{H}_2)$. Using same as technique as in Lemma 4, we have $\mathcal{R}_m(\mathcal{H}_1) \leq 2D^{(1)}||\boldsymbol{w}^{(1)}||_\infty \mathcal{R}_m(\mathcal{F}_1)$ and $\mathcal{R}_m(\mathcal{H}_2) \leq 2D^{(2)}||\boldsymbol{w}^{(2)}||_\infty \mathcal{R}_m(\mathcal{F}_2)$

□

Theorem 4 For the energy function $E(h,\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}||G^{(1)}_{\boldsymbol{W}}(\boldsymbol{x}) - G^{(2)}_{\boldsymbol{W}}(\boldsymbol{y})||_2^2$, over the input set $\mathcal{X} \in \mathbb{R}^N$, hypothesis class $\mathcal{H} = \{G^{(1)}_{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{i=1}^{D^{(1)}} w_i^{(1)}\phi_i^{(1)}(\boldsymbol{x}) = {\boldsymbol{w}^{(1)}}^T\Phi^{(1)}(\boldsymbol{x}), G^{(2)}_{\boldsymbol{W}}(\boldsymbol{y}) = \sum_{i=1}^{D^{(2)}} w_i^{(2)}\phi_i^{(2)}(\boldsymbol{y}) = {\boldsymbol{w}^{(2)}}^T\Phi^{(2)}(\boldsymbol{y}) \mid \Phi^{(1)} \in \mathcal{F}_1,\ \Phi^{(2)} \in \mathcal{F}_2,\ \forall \boldsymbol{x}\ ||\Phi^{(1)}(\boldsymbol{x})||_2 \leq A^{(1)},\ \forall \boldsymbol{y}\ ||\Phi^{(2)}(\boldsymbol{y})||_2 \leq A^{(2)}\}$, and output set $\mathcal{Y} \subset \mathbb{R}^N$, if the feature set $\{\phi_1^{(1)}(\cdot), \cdots, \phi_{D^{(1)}}^{(1)}(\cdot)\}$ is $\vartheta^{(1)}$-diverse with a probability τ_1 and the feature set $\{\phi_1^{(2)}(\cdot), \cdots, \phi_{D^{(2)}}^{(2)}(\cdot)\}$ is $\vartheta^{(2)}$-diverse with a probability τ_2, then with a probability of at least $(1-\delta)\tau_1\tau_2$, the following holds for all h in $\mathcal{H}$

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \boldsymbol{D}}[E(h,\boldsymbol{x},\boldsymbol{y})] \leq \frac{1}{m}\sum_{(\boldsymbol{x},\boldsymbol{y})\in \boldsymbol{S}} E(h,\boldsymbol{x},\boldsymbol{y}) + 8(\sqrt{\mathcal{J}_1} + \sqrt{\mathcal{J}_2})\big(D^{(1)}||\boldsymbol{w}^{(1)}||_\infty \mathcal{R}_m(\mathcal{F}_1) + D^{(2)}||\boldsymbol{w}^{(2)}||_\infty \mathcal{R}_m(\mathcal{F}_2)\big) + 2\big(\mathcal{J}_1 + \mathcal{J}_2\big)\sqrt{\frac{\log(2/\delta)}{2m}}, \quad (27)$$

where $\mathcal{J}_1 = ||\boldsymbol{w}^{(1)}||_\infty^2 \Big(D^{(1)}{A^{(1)}}^2 - {\vartheta^{(1)}}^2\Big)$ and $\mathcal{J}_2 = ||\boldsymbol{w}^{(2)}||_\infty^2 \Big(D^{(2)}{A^{(2)}}^2 - {\vartheta^{(2)}}^2\Big)$.

Proof. We replace the variables in Lemma 1 using Lemma 9 and Lemma 10. □