PAC-Learning for Energy-based Models

by

Xiang Zhang

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Computer Science

Courant Institute of Mathematical Sciences

New York University

May 2013

Professor Yann LeCun

Professor David Sontag

To my parents, with affection.

Acknowledgements

First I want to thank professor Yann LeCun for his advisement of this thesis. It was from his lectures that I saw the effectiveness of energy-based models and deep learning methods, which motivates me to start thinking about a theory for such well-engineered practice. Second, I want to to express my thanks to professor David Sontag who points out an error in my original thesis draft which would otherwise make this entire piece of work meaningless. Third, I would like to thank professor Mehryar Mohri who wrote such a concise book with treatment on the state-of-the-art theoretical results, which guides me through the first days of studying and exploration in the theoretical aspects of machine learning.

Abstract

In this thesis we prove that probably approximately correct (PAC) learning is guaranteed for the framework of energy-based models. Starting from the very basic inequalities, we establish our theory based on the existence of metric between hypothesis, to which the energy function is Lipschitz continuous. The result of the theory provides a new scheme of regularization called central regularization, which puts the effect of deep learning and feature learning in a new perspective. Experiments of this scheme shows that it achieved both good generalization error and testing error.

Contents

List of Figures

Introduction

The idea of energy-based models[11] is centered around the concept energy, which is the objective to be minimized during inference. This provides a very general framework for dealing with learning systems, and it immediately puts machine learning to the scope of mathematical optimization. In this thesis we provide a rigorous formulation of energy based models, with its three components – energy, discrimination and regularization – carefully studied.

The intention of PAC-learning[24] is to solve fundamental questions in learning such as what can be learnt and how many examples are needed to learn successfully. In the view of statistics, the formulation of PAC-learning is very much like a problem of seeking for a concentration inequality[3]. There are several of them, namely Chebyshev's inequality[3], Hoeffding's inequality[6], McDiarmid's inequality[17], and Bernstein's inequality[2]. They differ in terms of assumption and tightness of bounding.

Previous formulations of PAC-learning[9][25][18] use Hoeffding's inequality or McDiarmid's inequality to give polynomial bounds to sample size m, based on the assumption that both the risk and the hypotheses are in finite range. Similarly, in this thesis we formulate the bounds based on the assumption that the energy is in finite range.

If the hypotheses are parameterized by unbounded parameters, as in most energy-based models, it is likely that a maximum bound for all of the hypotheses does not exist. However, one can observe that if we are using minimization algorithms (empirical or structural), the hypotheses that the algorithm could explore are actually confined to a sublevel set of the hypothesis class at each step of the algorithm. This gives strong indication for combining algorithmic behaviour with

PAC analysis, which results in a novel concept *sublevel hypothesis class*.

The result of this thesis also provides a noval regularization approach, by the means that we establish our theory based on the existence of a metric between hypothesis, to which the loss function is Lipschitz continuous. The new scheme is to regularize around a center, which should be a good approximation of the true parameters of the best hypothesis. One good way of achieving such good approximation is by deep learning and feature learning techniques, for which this thesis provides some good experimental results.

For the purpose of conciseness, this thesis did not provide a complete example study of common used models and energy formulations used in energy-based models, but the reader can easily verify that all examples fit into this thesis by LeCun et al's tutorial paper[11].

Chapter 1

Elements of Energy-based Models

In this chapter we study the three elements of energy-based models: energy, discrimination and regularization. Following the convention of LeCun et al[11], energy is defined as the objective to be minimized during inference. In learning, the loss function would be more complicated than just the energy since mere energy neither distinguishes between correct and incorrect answers[13], nor provides effective generalization on the data the algorithm has not seen yet. Thus, the concepts of discrimination and regularization were introduced to justify these two necessities in a learning algorithm. A more comprehensive theory of machine learning was given in chapter 2, to study the theories behind regularization by examing different kinds of assumptions we could have over the energy and the hypothesis class.

1.1 Energy

The entire framework of energy-based models[11][13], by its name, is centered around the concept of *energy*. It captures dependencies by associating a scalar

energy (a measure of compatibility) to each configuration of the variables. Inference, i.e., making prediction or decision, consists in setting the value of observed variables and finding values of remaining variables that minimize the energy. As a result, learning consists in finding an energy function that associates low energies to correct values of the remaining variables, and higher energies to incorrect values. A loss functional, minimized during learning, is used to measure the quality of the available energy functions. Within this common inference/learning framework, the wide choices of energy functions and loss functionals allow for the design of many types of learning models, both probabilistic and non-probabilistic.

In order to better formulate our introduction, we place several definitions here, the first of which are the concept, hypothesis and decision.

Definition 1 (concept)**.** *A concept $c : \mathcal{X} \to \mathcal{Z}$ is a function from $\mathcal{X}$ to $\mathcal{Z}$, where $\mathcal{X}$ is the set of all possible inputs. A concept class C is a set of concepts.*

Definition 2 (hypothesis)**.** *A hypothesis h is also a concept. A hypothesis class H is a set of hypotheses.*

Definition 3 (decision)**.** *A decision is a functional $g : \{H \cup C\} \times \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y}$ is the set of all possible outputs.*

In machine learning, we are given samples $(x, y) \in \mathcal{X} \times \mathcal{Y}$ from which we wish to learn a hypothesis $h \in H$ of a given hypothesis class H to approximate the concept class C associated with a decision g. Usually the concept class C is not known explicitly during learning, but still it can be conceptually defined as follows.

Definition 4 (concept class associated with a decision)**.** *A concept class C_g associated with a decision g is a set of all possible concepts such that for any concept $c \in C_g$ and any input $x \in \mathcal{X}$, the value $y = g(c, x)$ is always the correct output.*

Although these definitions seem tedious, they will help us greatly in later illustration of the theorems. As we said before, energy is a function upon minimizing which we would be able to make decision. The definition below thus follows.

Definition 5 (energy)**.** *An energy is a functional* $E : \{H \cup C\} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$*, minimizing which could give a decision. There is a decision* g_E *associated with the energy* E *such that*

$$g_E(h, x) = \operatorname*{argmin}_{y \in \mathcal{Y}} E(h, x, y). \tag{1.1}$$

Definition 6 (well-defined energy)**.** *If an energy* E *is lower bounded at* 0*, that is*

$$\inf_{h \in \{H \cup C\},\ x \in \mathcal{X},\ y \in \mathcal{Y}} E(h, x, y) = 0, \tag{1.2}$$

then E *is well-defined. We can say* E *is badly-defined if* E *is not well-defined.*

The definition *well-defined energy* will be extensively useful in both the idea of discrimination and the proofs of the theorems in chapter 2. In fact, our entire theory resides in the existence of a well-defined energy. Given some badly-defined energy, it is usually possible to transform it to a well-defined energy with an equivalent decision.

Since we defined the decision associated with an energy, it is then often useful to define a concept class associated with an energy.

Definition 7 (concept class associated with an energy)**.** *The concept class* C_E *associated with an energy* E *is the concept class associated with* g_E*, which is the decision associated with* E*.*

One may wonder why unlike the previous theories, we defined *concept class* rather than just admit a single *concept*. This is actually self-explanatory from

the definitions above: a *concept* is not a mapping from the inputs to the outputs, rather, it is a mapping from the inputs to some intermediate values in $\mathcal{Z}$, and then taken by a decision g to make outputs. Thus, it is possible that multiple intermediate values in $\mathcal{Z}$ could associate with a particular output in $\mathcal{Y}$. Seen in this way, the intermediate value associated with an input of a concept is not unique, so it could not be a function. Using the definition *concept class* avoids this trouble, and later we will see it will make the definition of a metric between hypotheses and concepts easier. One example is binary classification, in which the decision function is usually $g(c, x) = \mathbf{sign}(c(x))$.

It is however possible that a concept c directly makes outputs, in which case the function g merely returns $c(x)$. In this case the *concept class* may refer to the possibility that there are multiple forms of c. One example is polynomial regression, in which conceptually speaking a higher-order polynomial function could be equivalent to a lower-order polynomial if its higher order coefficients are zero, but it may be referred to as a different *concept* than the lower-order one. This difference is usually important if we were to discuss the *concistency* between the hypothesis class and the concept class.

1.2 Discrimination and Margin

Rather than speaking of *discrimination* as a definition, it is better to say that it is an idea to measure the difference of energy between the incorrect outputs and the correct output. This is by virtue of the purpose of energy: it is to be minimized to make an inference. Thus, in learning, we may not want to choose a hypothesis that results in a flat or small-discrimination energy in the input

space. Rather, in machine learning we bear in our minds that it is perhaps best to maximize the discrimination. However, it is usually ill-posed or not always possible to maximize the discrimination to infinity. What we do in practice is try to optimize the discrimination to a target value, which is defined as the *margin*, or if we want the margin to be infinity, the optimization objective is penalized less as the discrimination goes larger. That being said, it is still possible for us to make some example definitions of discrimination. The reader should keep in mind that these definitions are not the unique ways.

Definition 8 (discrete discrmination)**.** *The discrimination associated with an energy E where $\mathcal{Y}$ is a discrete set can be defined as a functional $\rho : \{H \cup C\} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ in which*

$$\rho(h, x, y) = \min_{\tilde{y} \neq y} E(h, x, \tilde{y}) - E(h, x, y). \tag{1.3}$$

It is also possible to define the discrimination for a continuous $\mathcal{Y}$ if we contrain ourselves to a neighbourhood ϵ away from y.

Definition 9 (continuous discrimination)**.** *The ϵ-discrimination associated with an energy E where $\mathcal{Y}$ is a continuous set can be defined as a functional $\rho : \{H \cup C\} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ in which*

$$\rho(h, x, y) = \min_{\|\tilde{y} - y\| \geq \epsilon} E(h, x, \tilde{y}) - E(h, x, y). \tag{1.4}$$

It is worth noting that usually it would not be possible to calculate an optimization objective associated with a discrimination if $\mathcal{Y}$ is continuous, because it may consist of some integration over y. In such cases, we may want to get a discretization $\tilde{\mathcal{Y}}$ of the space of $\mathcal{Y}$ with respect to the given output y, such that the

optimization is computationally feasible. For cases like this, the discrimination is essentially discrete over set $\tilde{\mathcal{Y}}$.

As we have said, these are not the only ways to define discrimination. We will examine more kinds of discrimination as we discuss each of the algorithms along the way.

With the idea of discrimination, we could discuss a kind of transformation of energy such that it would bear with some form of discrimination. Spoken in a more mathematical way, the new energy $\mathcal{E}$ would be best if it could bear a margin of the discrimination ρ defined over an energy E, in which the margin is a target value of the discrimination if we would be able to minimize $\mathcal{E}$. In the context of our theory we will assume that $\mathcal{E}$ is a well-defined energy, such that its infimum is 0. Then, the following definition holds.

Definition 10 (margin)**.** *The margin of a well-defined energy $\mathcal{E}$ associated with a discrimination ρ defined with some other energy E is*

$$p = \min \{ m \,|\rho(h, x, y) \geq m \implies \mathcal{E}(h, x, y) = 0 \} . \tag{1.5}$$

Note that to make sense of the definition above, $\mathcal{E}(h, x, y)$ has to be a functional of $\rho(h, x, y)$. We wrote it in the above ways just to indicate that $\mathcal{E}(h, x, y)$ is also an energy.

Sometimes it is convenient to talk about a concept class associated with a marginal energy, in which each concept does not only gives the correct output for every energy term, but also embedding within itself the margin. For this we give the following definiton

Definition 11 (marginal concept class)**.** *The marginal concept class $C_{\mathcal{E}}$ of a well-*

defined energy $\mathcal{E}$ *with margin* m*, associated with a discrimination* $\rho(h, x, y)$ *defined with some other energy* E*, is a subclass of* C_E *satisfying* $\inf_{c \in C_\mathcal{E}} \rho(c, x, y) = m$*.*

Although the energies $\mathcal{E}$ and E should be different, it is still possible for them to be equivalent, that is, they produce the same decision g. Since $\mathcal{E}$ has the margin, it could be a good term to be used in a loss function. And since $\mathcal{E}$ is well-defined, all of our theories later could be applied to it.

1.3 Regularization

There has long been the discussion of generability in machine learning. A good generability indicates consistency between the empirical expectation of the energy and the true expectation, which means that it is of high probability that the empirical expectation of the energy calculated from a set of samples is close to the true expectation, assuming the samples are drawn in an independent and identical fashion from some distribution.

In machine learning the samples were assumed to be drawn from a true concept (the oracle) with or without some unbiased noise on the outputs, thus by definition the true expectation of the energy should be low. As a result, it is a good way to formulate a learning algorithm if we could have been able to minimize the true expectation. However, generally the true expectation is not computable, because in reality we are usually given just a finite number of samples. So we turn into minimizing the empirical expectation.

Unfortunately, such a minimization bears with the risk of overfitting, that is, the blindly minimized empirical expectation may be too low to represent the true expectation. Fortunately, there are statistical guarnatees to upper-bound the true

expectation with a high probability, using the empirical expectation plus some extra term. This extra term, in turn, will shed light on a regularization term to be added to the minimization objective to guarantee minimization of the true expectation, with a high probabiliy. This will be studied more extensively in the next chapter. For now, we will just define the regularization term as a functional of the hypotheses.

Definition 12 (regularization)**.** *A regularization term is a functional* $r : H \to \mathbb{R}$.

A machine learning problem in the scope of energy-based models could then be thought of as a minimizing procedure of the empirical expectation of the energy plus the regularizatoin term. Naturally, we will call this objective the *loss* of the model.

Definition 13 (loss)**.** *Given a set of samples* $S = \{(x, y)|(x, y) \in \mathcal{X} \times \mathcal{Y}\}$*, the loss is a functional* $L : H \to \mathbb{R}$ *defined as*

$$L(h) = \left[\frac{1}{|S|} \sum_{(x,y)\in S} E(h, x, y)\right] + \lambda \cdot r(h), \tag{1.6}$$

where E *is an energy and* r *is a regularization term.*

It is important to note that in pratice people always find some norm on the parameters of the hypothesis h can be a good regularization term. This is indeed the case, and our theory later will give a general proof of why it works. As of now, the elements of energy-based model are completely introduced. These elements will be used extensively in our later discussions of theories and practices.

1.4 An Example on Classification

To facilitate the reader in understanding the three elements above, we present two examples here which are (generalized) binary support vector machines and multi-class support vector machines, with comparisons between them side-by-side.

The hypothesis we would use for the binary class is a function $h_1 : \mathbb{R}^n \to \mathbb{R}$, whereas for a k-class multiclass hypothesis we use $h_{\mathrm{k}} : \mathbb{R}^n \to \mathbb{R}^k$, where $k \geq 2$. In the illustration below, we will also use the subscript 1 to represent binary classification and k for k-class multiclass classification. In the case that h_1 or h_k is parameterized linearly with respect to the input, our derivition below will naturally result in the classical binary or multi-class support vector machines[4][27].

Note that by definition 1 and 2, the construction above suggests that $\mathcal{X}_1 = \mathcal{X}_k = \mathbb{R}^n$, and $\mathcal{Z}_1 = \mathbb{R}, \mathcal{Z}_k = \mathbb{R}^k$. Notice that differently from any theory we had before, rather than producing class labels, a hypothesis produces real values that can later be used as a input to a decision function. However, as in definition 5, a decision is made upon minimizing some energy. Therefore, we first define the energy functionals for each case

$$E_1(h_1, x, y) = -yh_1(x), \ y \in \mathcal{Y}_1 = \{-1, +1\}, \tag{1.7}$$

$$E_k(h_k, x, y) = -(h_k(x))_y, \ y \in \mathcal{Y}_k = \{1, 2, \ldots, k\}, \tag{1.8}$$

where $(h(x))_y$ means the y-th element in the vector value $h(x)$. Notice the difference in which we have defined the output spaces $\mathcal{Y}_1$ and $\mathcal{Y}_k$. The energy

constructions above suggest the following decision functions

$$g_1(h_1, x) = \underset{y \in \mathcal{Y}_1}{\operatorname{argmin}} E_1(h_1, x, y) = \mathbf{sign}(h(x)), \tag{1.9}$$

$$g_k(h_k, x) = \underset{y \in \mathcal{Y}_k}{\operatorname{argmin}} E_k(h_k, x, y) = \underset{y \in \mathcal{Y}_k}{\operatorname{argmin}} - (h_k(x))_y. \tag{1.10}$$

The decision function for binary classification means that we choose the sign of $h_1(x)$, whereas for multiclass classification we choose the maximum component in the vector $h_k(x)$. These are all very standard choices – but resulted from our definition of energy functions.

However, neither of the energy functions above is well-defined. Both our learning problem formulation 13 and our main theorem 1 require that we have a well-defined energy. There are many equivalent forms of energy functions that can be used, including the Hinge loss function and Logistic loss function. They will produce support vector machines and (binary or multinomial) Logistic regression, repectively. As an example, we use Hinge function and its generalized form to construct the well-defined energy for binary and multi-class classification problems.

$$\mathcal{E}_1(h_1, x, y) = \max\left\{0, \frac{1}{2} - y h_1(x)\right\} \tag{1.11}$$

$$\mathcal{E}_k(h_k, x, y) = \sum_{\tilde{y} \in \mathcal{Y}_k \wedge \tilde{y} \neq y} \max\left\{0, 1 - [(h_k(x))_y - (h_k(x))_{\tilde{y}}]\right\} \tag{1.12}$$

It is easy to verify that the decision functions associated with the above well-defined energies are exactly the same as the decision functions g_1 and g_k associated with E_1 and E_k respectively.

The advantage of them being well-defined is not only that we can apply our later theorem 1 to it, but also that $\mathcal{E}_1$ and $\mathcal{E}_k$ bear a margin with respect to E_1

and E_k. Following definition definition 8, we can define the discriminations with respect to E_1 and E_k as

$$\begin{aligned}\rho_1(E_1, x, y) &= \min_{\tilde{y} \in \mathcal{Y}_1 \wedge \tilde{y} \neq y} E_1(h_1, x, \tilde{y}) - E_1(h_1, x, y) \\ &= E_1(h_1, x, -y) - E_1(h_1, x, y) \\ &= 2yh_1(x).\end{aligned} \tag{1.13}$$

$$\begin{aligned}\rho_k(E_k, x, y) &= \min_{\tilde{y} \in \mathcal{Y}_k \wedge \tilde{y} \neq y} E_k(h_k, x, \tilde{y}) - E_k(h_k, x, y) \\ &= \min_{\tilde{y} \in \mathcal{Y}_k \wedge \tilde{y} \neq y} (h_k(x))_y - (h_k(x))_{\tilde{y}}.\end{aligned} \tag{1.14}$$

Therefore, using equation 1.11 and 1.12, it easy to verify that there exist minimal $m_1 = 1, m_k = 1$ such that $\rho_1(E_1, x, y) \geq m_1 \implies \mathcal{E}_1(h_1, x, y) = 0$, and $\rho_k(E_k, x, y) \geq m_k \implies \mathcal{E}_k(h_k, x, y) = 0$. Therefore, by definition 10, the well-defined energies $\mathcal{E}_1$ and $\mathcal{E}_k$ both have margin 1. The concept of margin was original proposed to characterize how good a learning system is at distinguishing between the correct and the incorrect answers. The algorithm both having margin 1 means that by minimizing this well-defined energy, the algorithm will try its best to let the correct and incorrect answers differ for at least 1, in terms of the energy E_1 and E_k respectively. This is similar, but not equivalent, to the notion of geometrical margin proposed for binary support vector machines[4]. One thing to note is that, there have not been any other similarities in the notion of margin for multiclass support vector machines, except in the realm of energy-based models[11]. Therefore, we believe the notion of margin for energy-based models is a more general concept than the notion of geometrical margin.

As of now, the examples above illustrated the two first elements of energy-

based models – energy and discrimination, and their related concepts. To explain regularization, we need to first present our PAC-learning frameworks theorem 1 and theorem 2. The key result is theorem 2, which connects generalization error with a distance between a hypothesis and a concept class.

Proceeding from theorem 1 to theorem 2, as we will see later, requires that we can find a metric that our well-defined energy is Lipschitz continuous to. This is where our idea of marginal concept class – as in definition 11 – comes into play. Using the previous binary classification as example, marginal concept class $C_{\mathcal{E}_1}$ is a class of all concepts that produce correct results and it sasifies

$$\inf_{c_1 \in C_{\mathcal{E}_1}} \rho_1(c_1, x, y) = \inf_{c_1 \in C_{\mathcal{E}_1}} 2yc_1(x) = 1. \tag{1.15}$$

Therefore,

$$\begin{aligned} \mathcal{E}_1(h_1, x, y) &= \max\left\{0, \frac{1}{2} - yh_1(x)\right\} \\ &= \max\left\{0, \inf_{c_1 \in C_{\mathcal{E}_1}} yc_1(x) - yh_1(x)\right\} \\ &= \begin{cases} \inf_{c_1 \in C_{\mathcal{E}_1}} |c_1(x) - h_1(x)|, & \frac{1}{2} - yh_1(x) \geq 0, \\ 0, & \frac{1}{2} - yh_1(x) < 0. \end{cases} \end{aligned} \tag{1.16}$$

The first case is already a hypothesis distance. We need to study more closely for the second case. Notice that $\frac{1}{2} - yh_1(x) < 0$ suggests that the discrimination $\rho(h_1, x, y) \geq 1$. Using the definition of a marginal concept class, this means that the value of $h(x)$ is equal to the value of a concept $c_1(x), c_1 \in C_{\mathcal{E}_1}$, and such that

$\inf_{c_1 \in C_{\mathcal{E}_1}} |c_1(x) - h_1(x)| = 0$. Combining with the equation above, we know that

$$\mathcal{E}_1(h_1, x, y) = \inf_{c_1 \in C_{\mathcal{E}_1}} |c_1(x) - h_1(x)|. \tag{1.17}$$

As a result, the well-define energy $\mathcal{E}_1$ is Lipchitz continuous to a hypothesis distance as defined above, with Lipschitz constant 1. The case of multiclass classification can be similarly established. As a result, theorem 2 applies to the examples here and the explanation of regularization in chapter 3 holds.

Chapter 2

Probably Approximately Correct (PAC) Learning

In this chapter we introduce an extension of the traditional PAC-learning theory to energy-based models. All of the previous theories place generalization bounds on binary classification error, which does not quite fit into the framework of energy-based models. The reason is that energy-based models do not assume classification as its task. The measurement of error is placed directly using the loss. One advantage of this, as we will see in the next chapter, is that it gives connection of generalizaton bounds with algorithmic behaviour, by the fact that energy-based models do assume the minimization of a loss function as the algorithm.

2.1 Concentration Inequalities

We begin our chapter with the introduction of classical concentration inequalities[3]. These inequalities are used in the same way as all the previous theories, and so we

do not provide proofs on them. The reader could refer to our bibliography to find appropriate places where they were proved. The first one is Hoeffding's lemma, who will provide a good upper bound for the moment generating function used in Chernoff's bounding technique.

Corollary 1 (Hoeffding's lemma)**.** *Given $\epsilon > 0$, if X is an arbitrary random variable bounded in $[a, b]$, then for $t > 0$,*

$$\mathrm{e}^{t(X-\mathrm{E}[X])} \leq \mathrm{e}^{t^2(b-a)^2/8} \tag{2.1}$$

Then, using Chernoff's bounding technique, one can immediately verify the following Hoeffding's inequality[6].

Corollary 2 (Hoeffding's inequality)**.** *Given $\epsilon > 0$, if $X_1, \ldots, X_m$ are i.i.d. (independently and identically distributed) random variables bounded in $[a_i, b_i]$ respectively, then*

$$\Pr[S_m - \mathrm{E}[S_m] \geq \epsilon] \leq \mathrm{e}^{-2\epsilon^2/\sum_{i=1}^m (b_i-a_i)^2}, \tag{2.2}$$

$$\Pr[S_m - \mathrm{E}[S_m] \leq -\epsilon] \leq \mathrm{e}^{-2\epsilon^2/\sum_{i=1}^m (b_i-a_i)^2}, \tag{2.3}$$

where

$$S_m = \frac{1}{m}\sum_{i=1}^{m} X_i. \tag{2.4}$$

In this thesis the Hoeffding's inequality is not used. What was used in the subsequent establishment of bounds is actually the more powerful inequality by McDiarmid[17], as shown below. The advantage of this bound is that it generalizes Hoeffding's inequality to the case of a function parameterized by i.i.d. random varaibles.

Corollary 3 (McDiarmids Inequality)**.** *Let $X_1, \ldots, X_m$ be i.i.d. random variables from $\mathcal{X}$ and $f : \mathcal{X}^m \to \mathbb{R}$ verifying for all $i = 1, 2, \ldots, m$*

$$\sup_{x_1,\ldots,x_m,x_i'} |f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq M_i. \tag{2.5}$$

Then, for a given $\epsilon > 0$,

$$\Pr\left[|f(X_1, \ldots, X_m)| - \mathrm{E}[f(X_1, \ldots, X_m)| \geq \epsilon\right] \leq 2\mathrm{e}^{-2\epsilon^2/\sum_{i=1}^m c_i^2}. \tag{2.6}$$

These are all the concentration inequalities that are useful for this thesis. For a more detailed introduction of other concentration inequalities, the reader can refer to the survey by Boucheron et al.[3]

2.2 Stochastic Complexity

As the title suggests, this section concerns with the measurement of complexity for functions, which means the undesirability of the randomness of the output. However, we cannot just wave hands and make a concept without saying what is this notion of 'undesirability'. Our idea is that, intuitively, complexity of a system can be thought of as the correlation of its ability of producing output towards a noise distribution that characterizes the most undesired performance.

Clearly, such undesired performance should be problem dependent. For example, for real valued regression problems, the distribution characterizing the most undesired performance might be the Gaussian distributed centered at 0 with some variance (a.k.a. the Gaussian noise); on the other hand, for prediction of values in a range, white noise – uniform distribution in the range of output – may be the best

choice of undesired performance. In those particular machine learning applications which are to figure out how to do classification, the most undesired performance may be a discrete uniform distribution with respect to each classification output. In particular, for binary classification we have the salt and pepper noise.

In each of the examples above, the undesirability is identified by some probability distribution associated with the name 'noise'. Therefore, we generalize these probability distributions to the concept of noise distribution.

Definition 14 (Noise distribution)**.** *A distribution N is said to be a noise distribution if it satisfies the following requirements:*

1. *The distribution is symmetric:*

$$\forall \eta_0 \geq 0, \quad \Pr_{\eta \sim N}[\eta \geq \eta_0] = \Pr_{\eta \sim N}[\eta \leq -\eta_0]. \tag{2.7}$$

2. *The first (central) absolute moment exists:*

$$\mathrm{E}_{\eta \sim N}[|\eta - \mathrm{E}_{\eta' \sim N}[\eta']|] = \mathrm{E}[|\eta|] = \sigma_\eta^1 < \infty. \tag{2.8}$$

Requirement 1 is most intuitive, since we do not want our distribution to be biased to either positive or negative partof the parameter η. As a result of this requirement, the mean of a noise distribution is $\mathrm{E}_{\eta \sim N}[\eta] = 0$.

Requirement 2 is trivial if the noise distribution is bounded (e.g., white noise or any discrete uniform distribution) with its range satisfying some constraints. For real values, because of the central limit theorem, noise is most likely to be a Gaussian distribution who has some variance. Note that a Gaussian distribution also satysfies requirement 2.

Definition 15 (Empirical stochastic complexity)**.** *Let F be a family of functions mapping from $\mathcal{U}$ to $\mathbb{R}$, and $S = (u_1, u_2, \ldots, u_m)$ a fixed sample of size m with elements in $\mathcal{U}$. Then, the empirical stochastic complexity of F with respect to the sample S is defined as:*

$$\hat{\mathfrak{C}}_S(F) = \frac{1}{\sigma_\eta^1} \mathop{\mathrm{E}}_{\boldsymbol{\eta}} \left[\sup_{f \in F} \frac{1}{m} \sum_{i=1}^{m} \eta_i f(u_i) \right] \tag{2.9}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)^T$, with η_is independent random variables taking values from a noise distribution N. The random variables η_i are called noise variables.

The empirical stochastic complexity measures on average how well the function class F correlates with random noise on S. This describes the richness of the class F: richer or more complex class F can generate more kinds of $f(z_i)$'s and thus better correlate with the random noise, on average. There is a normalization term σ_η^1, which is a normalization term to different kinds of noise distributions, regardless of their first central absolute moments.

Definition 16 (Stochastic complexity)**.** *Let D denote the distribution from which the samples were drawn. For any integer $m \geq 1$, the stochastic complexity of F is the expectation of the empirical stochastic complexity over all samples of size m drawn according to D:*

$$\mathfrak{C}_m(F) = \mathop{\mathrm{E}}_{S \sim D^m} [\hat{\mathfrak{C}}_S(F)]. \tag{2.10}$$

Acure reader may discover that, if N characterizes the salt and pepper noise, the stochastic complexity is exactly the Rademacher complexity[1][18]. However, since energy-based models do not assume classification, there is no reason to restrain us to it. If N is Gaussian distribution, it is the Gaussian complexity[1] normalized by

its first absolute central moment. Notice that the first central absolute moment of the standard Gaussian distribution sohuld be $2/\pi$.

After these definition, we present a lemma which will be used later.

Lemma 1. *Let F be a family of functions mapping from $\mathcal{U}$ to $\mathbb{R}$, and assume $S = (u_1, u_2, \ldots, u_m)$ and $S' = (u'_1, u'_2, \ldots, u'_m)$ are samples of size m drawn i.i.d. from distribution D. Then, the following holds for variables $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_m)$ drawn i.i.d. from a noise distribution N:*

$$\mathop{\mathrm{E}}_{S,S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}(f(u'_i)-f(u_i))\right] \leq \frac{1}{\sigma^1_\eta}\mathop{\mathrm{E}}_{\boldsymbol{\eta},S,S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\eta_i(f(u'_i)-f(u_i))\right]. \quad (2.11)$$

Proof. Since N is a noise distribution, it can be known that it has first (central) absolute moment $\mathop{\mathrm{E}}_{\eta}[|\eta|] = \sigma^1_\eta$, as in requirement 2 of definition 16. Thus, the following derivation holds:

$$\begin{aligned}
\mathop{\mathrm{E}}_{S,S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}(f(u'_i)-f(u_i))\right] &= \frac{\sigma^1_\eta}{\sigma^1_\eta}\mathop{\mathrm{E}}_{S,S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}(f(u'_i)-f(u_i))\right] \\
&= \frac{1}{\sigma^1_\eta}\mathop{\mathrm{E}}_{S,S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma^1_{\eta_i}(f(u'_i)-f(u_i))\right] \\
&= \frac{1}{\sigma^1_\eta}\mathop{\mathrm{E}}_{S,S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\mathrm{E}[|\eta_i|](f(u'_i)-f(u_i))\right] \\
&= \frac{1}{\sigma^1_\eta}\mathop{\mathrm{E}}_{S,S'}\left[\sup_{f\in F}\mathop{\mathrm{E}}_{\boldsymbol{\eta}}\left[\frac{1}{m}\sum_{i=1}^{m}|\eta_i|(f(u'_i)-f(u_i))\right]\right] \\
&\leq \frac{1}{\sigma^1_\eta}\mathop{\mathrm{E}}_{\boldsymbol{\eta},S,S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}|\eta_i|(f(u'_i)-f(u_i))\right] \\
&= \frac{1}{\sigma^1_\eta}\mathop{\mathrm{E}}_{\boldsymbol{\eta},S,S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\eta_i(f(u'_i)-f(u_i))\right].
\end{aligned} \quad (2.12)$$

The second last line comes from the convexity of supremum. The last line comes

from that when $\eta_i > 0$, the associated summand remains unchanged; when $\eta_i < 0$, the associated summand flips signs, which is equivalent to swapping z_i and z_i' between S and S'. Since we are taking expectation over all possible S and S', this swap does not affect the overall expectation. The lemma is therefore proved. □

2.3 Generalization Bounds

In energy-based models, we do not assume that the prescribed hypothesis class is finite. In particular, in almost all cases the hypothesis is parameterized by a parameter w in the real multi-dimensional space. As a result, although we can easily construct an approximation bound for each of the hypotheses in the class, it is not guaranteed that when the algorithm explores these hypotheses it could achieve good generalization. Thus, by the fact that we are dealing with all hypotheses in the class, it is better if we can bound the maximum difference between the true expected energy and the emirical expectation, i.e.,

$$\phi_S(H) = \sup_{h \in H} \left\{ \mathop{\mathrm{E}}_{(x,y)\sim D} [E(h,x,y)] - \frac{1}{|S|} \sum_{(x,y)\in S} E(h,x,y) \right\} \tag{2.13}$$

where D is the distribution from which samples (x, y) were drawn.

To use McDiarmid's inequality on $\phi(S)$, it is necessary to identify a bound on the change of $\phi(S)$ if only one point is changed. If we assume that an upper bound on the energy is M, then it follows the lemma below.

Lemma 2. *Let F be a family of non-negative functions mapping from $\mathcal{U}$ to $\mathbb{R}$ with its values upper bounded by some value $M \geq 0$. Assume $S = (u_1, u_2, \ldots, u_m)$ is a sample of size m drawn i.i.d. from distribution D. Define the following functionals*

$\psi_S(f)$ and $\phi_S(F)$:

$$\psi_S(f) = \mathop{\mathrm{E}}_{u\sim D}[f(u)] - \frac{1}{m}\sum_{i=1}^{m} f(u_i), \quad \phi_S(F) = \sup_{f\in F}\ \psi_S(f), \tag{2.14}$$

where it verifies that $\phi_S(F) > 0$. Then, it holds that

$$\sup_{S,S'}\ |\phi_S(F) - \phi_{S'}(F)| \leq \frac{M}{m}, \tag{2.15}$$

where S' differs with S at only one point.

Proof. Denote $S' = (u'_1, u'_2, \ldots, u'_k, \ldots, u'_m) = (u_1, u_2, \ldots, u'_k, \ldots, u_m)$, such that the samples S and S' differ at u_k. Without loss of generality, let us assume that $\phi_S(F) \geq \phi_{S'}(F)$. Denote $f^* = \operatorname*{argmax}_{f\in F} \psi_S(f)$, we know that $\phi_S(F) = \psi_S(f^*)$, and $\phi_{S'}(F) = \operatorname*{argmax}_{f\in F} \psi_{S'}(f) \geq \psi_{S'}(f^*)$. As a result,

$$\begin{aligned}
|\phi_S(F) - \phi_{S'}(F)| &= \phi_S(F) - \phi_{S'}(F) \\
&= \psi_S(f^*) - \phi_{S'}(F) \\
&\leq \psi_S(f^*) - \psi_{S'}(f^*) \\
&= |\psi_S(f^*) - \psi_{S'}(f^*)| \\
&\leq \sup_{f\in F} |\psi_S(f) - \psi_{S'}(f)| \\
&= \sup_{f\in F} \left| \left[\mathop{\mathrm{E}}_{u\sim D}[f(u)] - \frac{1}{m}\sum_{i=1}^{m} f(u_i) \right] - \left[\mathop{\mathrm{E}}_{u\sim D}[f(u)] - \frac{1}{m}\sum_{i=1}^{m} f(u'_i) \right] \right| \\
&= \sup_{f\in F} \frac{1}{m} |f(u'_k) - f(u_k)| \\
&\leq \frac{M}{m}
\end{aligned} \tag{2.16}$$

A simlar result can be obtained if we assume $\phi_S(F) \leq \phi_{S'}(F)$. The lemma is

therefore proved. □

Note that by replacing u with (x, y) and $f(u)$ with $E(h, x, y)$ we recover the definition of $\phi_S(H)$ as in equation 2.13. The lemma above provides a direct way to apply McDiarmid's inequality (as in corollary 3). This will give us the theorem below, which is the sole generalization bound of this thesis.

Lemma 3 (Approximation bound for a function family with finite bound)**.** *Let F be a family of non-negative functions mapping from $\mathcal{U}$ to $\mathbb{R}$ upper bounded by some value $M \geq 0$. Assume $S = (u_1, u_2, \ldots, u_m)$ is a sample of size m drawn i.i.d. from distribution D. Define functional $\phi_S(F)$ as*

$$\phi_S(F) = \sup_{f \in F} \mathop{\mathrm{E}}_{S \sim D^m}[f(u)] - \frac{1}{m} \sum_{i=1}^{m} f(u_i) \tag{2.17}$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for $\phi_S(F)$:

$$\phi_S(F) \leq 2\mathfrak{C}_m(F) + M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{2.18}$$

Proof. With lemma 2, applying $\phi_S(F)$ to McDiarmid's inequality (corollary 3), we obtain

$$\Pr\left[\left|\phi_S(F) - \mathop{\mathrm{E}}_{S \sim D^m} \phi_S(F)\right| \geq \epsilon\right] \leq 2\mathrm{e}^{-2\epsilon^2/(M^2/m)}. \tag{2.19}$$

Setting the right hand to be δ we get $\epsilon = M\sqrt{\frac{\log(\frac{2}{\delta})}{2m}}$. Therefore, with probability at least $1 - \delta$,

$$\phi_S(F) \leq \mathop{\mathrm{E}}_{S \sim D^m}[\phi_S(F)] + M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{2.20}$$

What remains is to bound $\mathop{\mathrm{E}}_{S\sim D^m}[\phi_S(F)]$. Denote $\hat{\mathrm{E}}_S[f] = \frac{1}{m}\sum_{i=1}^m f(u_i)$, we have

$$\begin{aligned}
\mathop{\mathrm{E}}_{S\sim D^m}[\phi_S(F)] &= \mathop{\mathrm{E}}_{S\sim D^m}\left[\sup_{f\in F}\mathrm{E}(f) - \hat{\mathrm{E}}_S(f)\right] \\
&= \mathop{\mathrm{E}}_{S\sim D^m}\left[\sup_{f\in F}\mathop{\mathrm{E}}_{S'\sim D^m}\left[\hat{\mathrm{E}}_{S'}(f) - \hat{\mathrm{E}}_S(f)\right]\right] \\
&\le \mathop{\mathrm{E}}_{S',S\sim D^m}\left[\sup_{f\in F}\hat{\mathrm{E}}_{S'}(f) - \hat{\mathrm{E}}_S(f)\right] \\
&= \mathop{\mathrm{E}}_{S',S\sim D^m}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^m (f(u_i') - f(u_i))\right],
\end{aligned} \tag{2.21}$$

in which the third line came from Jensen's inequality for the convex supremum function. Apply lemma 1 to the right hand side of the inequality above, we get

$$\begin{aligned}
\mathop{\mathrm{E}}_{S\sim D^m}[\phi_S(F)] &\le \mathop{\mathrm{E}}_{S',S\sim D^m}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^m (f(u_i') - f(u_i))\right] \\
&\le \frac{1}{\sigma_\eta^1}\mathop{\mathrm{E}}_{\boldsymbol{\eta},S',S}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^m \eta_i(f(u_i') - f(u_i))\right] \\
&\le \frac{1}{\sigma_\eta^1}\mathop{\mathrm{E}}_{\boldsymbol{\eta},S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^m \eta_i f(u_i')\right] + \frac{1}{\sigma_\eta^1}\mathop{\mathrm{E}}_{\boldsymbol{\eta},S}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^m -\eta_i f(u_i)\right] \\
&\le \frac{1}{\sigma_\eta^1}\mathop{\mathrm{E}}_{\boldsymbol{\eta},S'}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^m \eta_i f(u_i')\right] + \frac{1}{\sigma_\eta^1}\mathop{\mathrm{E}}_{\boldsymbol{\eta},S}\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^m \eta_i f(u_i)\right] \\
&= 2\mathfrak{C}_m(F),
\end{aligned} \tag{2.22}$$

where $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_m)$ is a sample of noise drawn i.i.d. from noise distribution N. The theorem is therefore proved. □

As before, if we replace u by (x, y) and $f(u)$ with $\mathcal{E}(h, x, y)$, we can get the generalization bound on a well-defined energy, as follows.

Theorem 1 (Approximation bound with finite bound for energy)**.** *For a well-*

defined energy $\mathcal{E}(h, x, y)$ *over hypothesis class* H*, input set* $\mathcal{X}$ *and output set* $\mathcal{Y}$*, if it has an upper bound* $M > 0$*, then with probability at least* $1 - \delta$*, the following holds for all hypothesis* $h \in H$*:*

$$\mathop{\mathrm{E}}_{(x,y)\sim D}[\mathcal{E}(h, x, y)] \leq \frac{1}{m} \sum_{(x,y)\in S} \mathcal{E}(h, x, y) + 2\mathfrak{C}_m(F) + M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}, \tag{2.23}$$

where the function family F *is defined as*

$$F = \{\mathcal{E}(h, x, y) | h \in H\}, \tag{2.24}$$

D *is a distribution on the samples* (x, y)*, and* S *is a set of samples of size* m *drawn i.i.d. from* D*.*

The theorem provides a generalization bound for energy-based models with well-defined (non-negative) and bounded energy. The result suggests that merely minimizing the empirical expectation of energy may not give a good approximation of the true expectation. Instead, we need to figure out the appropriate term $r(h)$ for a structural energy minimization algorithm, based on the algorithmic effect on the two parameters $\mathfrak{C}_m(F)$ and M. As we will discuss in the following chapter, when a minimization algorithm proceeds, it will confine the hypothesis class H that the algorithm is able to explore dynamically, by the optimization objective. This suggests a dynamic treatment of the hypothesis class H with consideration of minimization algorithms.

2.4 Comparison to Previous Theories

To let the reader better understand our theory, here we provide a comaprison of our theory to the previous PAC-learning frameworks[9][18][25] and those PAC-Bayes models[16][20][21][23]. The difference is two-fold, for which one fold is a difference in how to get to our result – theorem 1, and the other fold is a difference in how we proceed from here to a better explanation of regularization.

The reader may recognise many different versions of theorems in the realm of learning theory that look just like theorem 1 – which is true since we used pretty much the same concentration inequalities to get here – but the assumptions are hugely different. All previous PAC-learning and PAC-Bayes frameworks assume binary classification problem and bound the classification error, while we assume a well-defined energy that is used as part of the optimization problem in learning and bound it directly to lay the foundation of a more direct way of explainin regularization. The theorem works for not only binary classification, but also regression, multi-class classification and even probability estimation – anything that can be formulated in energy-based models. Also, almost any learning algorithm that gets to the form of a mathematical optimization problem can usually be formulated in energy-based models, thanks to the large number of examples and tricks given in LeCun et al's tutorial[11].

Another part of the reason why we can get here is the definition of hypothesis – we separted the definitions of hypothesis and decision, unlike the previous theories where the hypothesis produces binary results directly. We disagree with the later approach because it lacks one layer of thinking – as a function, a linear hypothesis produces real values, rather than binary classification labels. This is is also the intrinsic advantage of energy-based models, which makes it a generalized framework

to so many different forms of learning problems, and therefore our theory applies to them too.

Before we talk about the difference of how we proceed from here, another important difference is in the idea of stochastic complexity. Previous PAC-learning theories use Rademacher complexity[1][18] which is a degenerated case of our stochastic complexity, as previously shown. This complexity measurement is meaningful for functions that produces binary values – such as 0-1 loss used in previous theories – but it does not make sense to the definition of energy which is usually a real value. Therefore, our generalized stochastic complexity helps to make better intuition.

A theory is much less meaningful if it is useless for practice. Therefore, where we go from here is a an even more crucial question. The general routine of previous PAC-learning theories is to ignore the third term in equation 2.23, since M was characterized as a value prescribed to the hypothesis class. More specifically, since the previous theories bound the 0-1 loss, M in this case is precisely 1[25]. The hope of the third term falls upon the number of samples m, that is, when m is is large enough it can be guaranteed to be small. Then, the second term, for which the previous theories use Rademacher complexity, is bounded by a sequence of complicated concepts such as growth functions and Vapnik-Chervonenkis dimension (VC-dimension)[26], with the help of Massart's lemma[15] and Sauer's lemma[19]. The result is a bound on the generalization error proportional to $O(\sqrt{(d + \log 1/\delta)/m})$, in which d is the VC-dimension[9][18][25].

At the beginning this was quite exciting, since a numerical bound naturally falls on the possibility of figuring out what is the VC-dimension d for a given hypothesis class. However, in practice this rarely works except for a very few simple cases of hypotheses. There are two reasons for this disapointing fact. The obvious one is

that it is difficult and challenging for anyone to characterize the VC-dimension for some general class of hypothesis functions; the amount of mathematical analysis to get VC-dimension is quite intimidating. The other reason, not so obvious, is that the assumption that VC-dimension – a static value – does not change during learning is problematic. This is particularly true if a learning algorithm could be an iterative optimization procedure – like that in our case of energy-based models, since at each step the algorithm will have some guarantee to not go beyond some changing value of the optimization objective, and therefore the set of hypotheses that the algorithm is able to explore afterwards is actually confined to those that produce a smaller objective value than this upper bound.

The later problem is dealt with in chapter 3 of our thesis, for which this dynamic subset of hypotheses was named *sublevel hypothesis class*. To do that, we begin with a bound on a *hypothesis distance* that the energy function is Lipschitz continuous to, and bound the stochastic complexity and the upper bound M with the complexity and upper bound of this distance measurement. What helped us to get there is a generalized Talagran's contraction lemma, presented in Talagrand et al's excellent introductory book to Banach-space statistics[14]. The bound on sublevel hypothesis class therefore falls into the possiblity of bounding the upper bound of this hypothesis distance, since its stochastic complexity could also be hopefully bounded by this upper bound. As a result, the notions of complicated upper bounds for the Radamacher complexity such as growth function and VC-dimension are not used, therefore eliminating the difficulty in figuring them out.

Just like many distance measrements, the hypothesis distance satisfies the triangular inequality. In chapter 3 we will use this triangular inequality twice. The

first application reveals a similar idea of bias-variance trade-off[5], which we believe is the first time that such similarity appeared in a PAC-learning framework. The second application reveals the reason why metric-based regularization works, such as l_p regularization used extensively in machine learning practices today. Further more, our revelation extends the metric-based regularization to the new form of a centered metric around the parameter that was learnt using prior knowledge unsupervisedly. This form matches with the recent ideas of unsupervised pre-training, which is a part of the excitement in deep learning. We therefore conducted several experiments to validate this new way of regularization, and it produced both better generalization error and testing error compared to raw norm-based regularization.

Our last words will compare between the explanation of regularization from our framework and PAC-Bayes models. The PAC-Bayes framework[16][20][21][23] is particularly interesting because it combines frequentists' idea of generalization error with the prior and posterior distribution of hypothesis. Similarly to previous PAC-learning theories, the PAC-Bayes framework applies only to binary classificatioin problem, to which our theory holds an advantage. However, PAC-Bayes also provides an explanation of regularization by the fact that the generalization error is bounded by the a KL-divergence defined with prior distribution of hypotheses. It can explain our new way of regularization, although only for binary classification problems. We are quite interested in extending PAC-Bayes framework to energy-based models, to see whether some results in combining algorithm behaviour could also be achieved. This could be a piece of good work for the future.

So far, neither the previous theories nor our current theory can give an efficient numerical bound for general learning problems, except for some very limited cases. But as the datasets used in machine learning grows larger and larger, a numerical

bound is quite appealing but less necessary, compared to a better explanation of regularization. Therefore, this thesis does not concern with a numerical bound. Without further ado, the next chapter introduces the latter parts of our theory.

Chapter 3

Explanation of Regularization

In this chapter, we provide an explanation for the technique of regularization used in many machine learning techniques. Specifically, we provide a noval explanation for role of regularization by considering the effect of optimization algorithms in the learning process. This gives one advantage that is provided by the formulation of energy-based models, as in equation 1.6, that the we immediately put ourselves into seeking the solution of a minimization algorithm. This provides the possibility of connecting the complexity and bound measurements dynamically with respect to how the algorithm proceeds. In turn, this thought also provides a way to design the optimization objective – the regularization term – in a theoretically founded way.

3.1 Hypothesis Metric

Previously we have a bound for a well-defined energy. This may not be a practical bound, since the energy function may be too complicated to derive anything meaning for its complexity. Instead, if we can give an approximation bound using

a measurement of the distance between the hypothesis and the concept class, it may be a better bound in the sense that this distance is not only conceptually intuitive, but also helpful in our later derivation of any norm-based regularization used extensively in machine learning literature.

We begin our discussion with a definition of this distance measurement, which is dependent on a metric between a hypothesis and a concept, when a sample is given. Note that since hypotheses and concepts are essentially the same thing, the metric is mathematically symmetrical.

Definition 17 (Hypothesis distance)**.** *Given a sample input x, assume we associate a metric $\kappa(c(x), h(x))$ between a concept and a hypothesis. Then, the distance between a hypothesis h and the concept class is defined as $k(C(x), h(x)) = \inf_{c \in C} \kappa(c(u), h(u))$*

The motivation for us to introduce this definition of hypothesis distance is that we want to provide a generalization bound for it, to offer easier access to the implications of theorem 1 associated with the algorithmic behaviour in a learning process. Therefore, we need to bound the two terms $\mathcal{C}_m(F)$ and the upper bound M. This puts us in need of a relationship between a well-defined energy and a hypothesis distance. For this, our assumption is that the energy provides a functional Lipschitz continuous guarantee with respect to the distance.

Definition 18 (Lipschitz energy)**.** *If a well-defined energy $\mathcal{E}(h, x, y)$ satisfies the inequality*

$$\|\mathcal{E}(h_1, x, y) - \mathcal{E}(h_2, x, y)\| \leq \mathcal{L} \, \|k(C(x), h_1(x)) - k(C(x), h_2(x))\| \,, \tag{3.1}$$

where C is a prescribed concept class.

The relationship between the stochastic complexity of the energy function class and that of the hypothesis distance can be characterized by the following lemma.

Lemma 4. *Let $\Phi : \mathbb{R} \to \mathbb{R}$ be an $\mathcal{L}$-Lipschitz function. Then, for any function class K of real-valued functions, the following inequalities hold:*

$$\begin{aligned} \hat{\mathfrak{C}}_S(\Phi \circ K) &\leq \mathcal{L} \cdot \hat{\mathfrak{C}}_S(K) \\ \mathfrak{C}_m(\Phi \circ K) &\leq \mathcal{L} \cdot \mathfrak{C}_m(K) \end{aligned} \tag{3.2}$$

where S is sample of size m drawn i.i.d. from distribution D.

Proof. First of all, observe that if the first inequality holds for any S of sample size m, then the second holds since it is just an expectation of the first one. Thus, the rest of this proof concerns the first inequality only. Let's fix a sample $S = (x_1, x_2, \ldots, x_m)$. Then, by definition

$$\begin{aligned} \hat{\mathfrak{C}}_S(\Phi \circ K) &= \frac{1}{\sigma_\eta^1 m} \mathop{\mathrm{E}}_{\boldsymbol{\eta}} \left[\sup_{k \in K} \sum_{i=1}^{m} \eta_i (\Phi \circ k)(x_i) \right] \\ &= \frac{1}{\sigma_\eta^1 m} \mathop{\mathrm{E}}_{\eta_1, \ldots \eta_{m-1}} \left[\mathop{\mathrm{E}}_{\eta_m} \left[\sup_{k \in K} u_{m-1}(k) + \eta_m (\Phi \circ k)(x_m) \right] \right] \end{aligned} \tag{3.3}$$

where $u_{m-1} = \sum_{i=1}^{m-1} \eta_i (\Phi \circ k)(x_i)$. Also, by definition of E_{η_m} and symmetry property of noise distribution,

$$\begin{aligned} &(1-\epsilon) \mathop{\mathrm{E}}_{\eta_m} \left[\sup_{k \in K} u_{m-1}(k) + \eta_m (\Phi \circ k)(x_m) \right] \\ &= (1-\epsilon) \mathop{\mathrm{E}}_{\eta_m} \left[\frac{1}{2} \sup_{k \in K} [u_{m-1}(k) + |\eta_m| (\Phi \circ k)(x_m)] + \frac{1}{2} \sup_{k \in K} [u_{m-1}(k) - |\eta_m| (\Phi \circ k)(x_m)] \right]. \end{aligned} \tag{3.4}$$

By definition of the supremum, there is some $\delta > 0$ such that for any $0 < \epsilon < \delta$

and $\eta_m \in \mathbb{R}$, there exist $k_1, k_2 \in K$ such that

$$\begin{aligned} & u_{m-1}(k_1) + |\eta_m|(\Phi \circ k_1)(x_m) \geq (1-\epsilon)\left[\sup_{k\in K} u_{m-1}(k) + |\eta_m|(\Phi \circ k)(x_m)\right] \\ \text{and} \quad & u_{m-1}(k_2) - |\eta_m|(\Phi \circ k_2)(x_m) \geq (1-\epsilon)\left[\sup_{k\in K} u_{m-1}(k) - |\eta_m|(\Phi \circ k)(x_m)\right]. \end{aligned} \tag{3.5}$$

Now, let us bound the right hand side of equation 3.4 by fixing an η_m. Using the the inequality for supremum, for any $0 < \epsilon < \delta$,

$$\begin{aligned} & (1-\epsilon)\left[\frac{1}{2}\sup_{k\in K}[u_{m-1}(k) + |\eta_m|(\Phi \circ k)(x_m)] + \frac{1}{2}\sup_{k\in K}[u_{m-1}(k) - |\eta_m|(\Phi \circ k)(x_m)]\right]. \\ & \leq \frac{1}{2}[u_{m-1}(k_1) + |\eta_m|(\Phi \circ k_1)(x_m)] + \frac{1}{2}[u_{m-1}(k_2) - |\eta_m|(\Phi \circ k_2)(x_m)] \end{aligned} \tag{3.6}$$

Let $s = \mathbf{sign}(k_1(x_m) - k_2(x_m))$. Then, the previous inequality implies

$$\begin{aligned} & (1-\epsilon)\left[\frac{1}{2}\sup_{k\in K}[u_{m-1}(k) + |\eta_m|(\Phi \circ k)(x_m)] + \frac{1}{2}\sup_{k\in K}[u_{m-1}(k) - |\eta_m|(\Phi \circ k)(x_m)]\right]. \\ & \leq \frac{1}{2}[u_{m-1}(k_1) + |\eta_m|(\Phi \circ k_1)(x_m)] + \frac{1}{2}[u_{m-1}(k_2) - |\eta_m|(\Phi \circ k_2)(x_m)] \\ & \leq \frac{1}{2}[u_{m-1}(k_1) + u_{m-1}(k_2) + s\mathcal{L}|\eta_m| \cdot (k_1(x_m) - k_2(x_m))] \quad \text{(Lischitz property)} \\ & = \frac{1}{2}[u_{m-1}(k_1) + s\mathcal{L}|\eta_m| \cdot k_1(x_m)] + \frac{1}{2}[u_{m-1}(k_2) - s\mathcal{L}|\eta_m| \cdot k_2(x_m)] \\ & \leq \frac{1}{2}\sup_{k\in K}[u_{m-1}(k) + s\mathcal{L}|\eta_m| \cdot k(x_m)] + \frac{1}{2}\sup_{k\in K}[u_{m-1}(k) - s\mathcal{L}|\eta_m| \cdot k(x_m)]. \end{aligned} \tag{3.7}$$

As a result,

$$
\begin{aligned}
&(1-\epsilon)\mathop{\mathrm{E}}_{\eta_m}\left[\sup_{k\in K} u_{m-1}(k)+\eta_m(\Phi\circ k)(x_m)\right] \\
&=(1-\epsilon)\mathop{\mathrm{E}}_{\eta_m}\left[\frac{1}{2}\sup_{k\in K}[u_{m-1}(k)+|\eta_m|(\Phi\circ k)(x_m)]+\frac{1}{2}\sup_{k\in K}[u_{m-1}(k)-|\eta_m|(\Phi\circ k)(x_m)]\right] \\
&\leq \mathop{\mathrm{E}}_{\eta_m}\left[\frac{1}{2}\sup_{k\in K}[u_{m-1}(k)+s\mathcal{L}|\eta_m|\cdot k(x_m)]+\frac{1}{2}\sup_{k\in K}[u_{m-1}(k)-s\mathcal{L}|\eta_m|\cdot k(x_m)]\right] \\
&=\mathop{\mathrm{E}}_{\eta_m}\left[\sup_{k\in K} u_{m-1}(h)+\mathcal{L}\cdot\eta_m k(x_m)\right],
\end{aligned}
\tag{3.8}
$$

where the last line again utilizes the definition of expectation and the symmetry property of noise distribution. Since the inequality holds for all $0<\epsilon<\delta$, we have

$$
\mathop{\mathrm{E}}_{\eta_m}\left[\sup_{k\in K} u_{m-1}(k)+\eta_m(\Phi\circ k)(x_m)\right]\leq \mathop{\mathrm{E}}_{\eta_m}\left[\sup_{k\in K} u_{m-1}(h)+\mathcal{L}\cdot\eta_m k(x_m)\right]. \tag{3.9}
$$

Proceeding in the same way for all other η_is ($i\neq m$) proves the lemma. □

It can be easily recognised that if $\hat{\mathfrak{C}}_S$ is the Rademacher complexity, the lemma above is identical to the Talagrand's lemma[14][18] as in Mohri et al's book[18]. Before we introduce the complete bound for hypothesis distance, we need another lemma to bound the maximum of a well-defined energy function, as follows. The proof is very easily verified using the non-negativity of a well-defined energy.

Lemma 5. *Assume a well-defined energy $\mathcal{E}$ is $\mathcal{L}$-Lipschitz with respect to a hypothesis distance k. If the maximum of the hypothesis distance is M, then the maximum of $\mathcal{E}$ is bounded by $\mathcal{L}\cdot M$.*

Following the two lemmas above, applying theorem 1 we now can give the thorem below for a bound on a well-defined energy, using the stochastic complexity and maximum bound of the hypothesis distance.

Theorem 2 (Approximation bound with finite bound for hypothesis distance)**.** *For a well-defined energy $\mathcal{E}(h, x, y)$ over hypothesis class H, input set $\mathcal{X}$ and output set $\mathcal{Y}$, if it is $\mathcal{L}$-Lipscthiz to a metric k who has an upper bound $M > 0$, then with probability at least $1 - \delta$, the following holds for all hypothesis $h \in H$:*

$$\mathop{\mathrm{E}}_{(x,y)\sim D}[\mathcal{E}(h, x, y)] \leq \frac{1}{m} \sum_{(x,y)\in S} \mathcal{E}(h, x, y) + 2\mathcal{L} \cdot \mathfrak{C}_m(K) + \mathcal{L} \cdot M\sqrt{\frac{\log \frac{2}{\delta}}{2m}}, \tag{3.10}$$

where the function family K is defined as

$$K = \{k(C(x), h(x))|h \in H\}, \tag{3.11}$$

D is a distribution on the samples (x, y), and S is a set of samples of size m drawn i.i.d. from D.

The theorem above provides us a tool for bounding generalizatoin errors using the hypothesis distance, other than the energy functions, for that in practice the distance is always easier to be used for deriving feasible regularization terms for the structural loss minimization algorithm. Later part of this thesis will show how a regularization term can be derived generally.

Similarly to the bound on energy, the theorem above states that in order to get good generalization on true expectation, we have to reduce the stochastic complexity and the upper-bound of the metric. This will follow with our innovative idea in this thesis, that is rather than thinking of the stochastic complexity and the upper-bound of the metric as static, we think of them as dynamically changing as the algorithm proceeds. Thus, compared to the previous PAC-learning theories[9][25][18] and PAC-Bayes theories [16][20][21][23], this thesis may provide

a different view of how does the statistical guarantees work and thus a more direct way of stipulate regularization.

3.2 Sublevel Hypothesis Class

Even though we have the previous bounds on energy and hypothesis distance, a difficulty appears since we cannot easily ask the question of whether there is an upper of the stochastic complexity or whether M exists, if no additional prior assumptions are given. For that, we introduce the concept sublevel hypothesis class to characterize the property of a minimization learning algorithm, that at a certain step the objective never goes beyond some limitation. This seemingly limited scope of thinking includes all that can be formulated in energy-based models – indeed, all energy-based models are minimization problems with an objective in the form of equation 1.6.

More specifically, if we assume a guarantee from an iterative optimization algorithm that at a certain step the objective value never goes beyond the previous step, then when the algorithm proceeds, it keeps confining the scope of hypotheses to those ones that can only have smaller objective values. However, this guarantee can be loosened as long as the algorithm converges, since we can than assume at a certain time during the algorithm's computation, if we were to inspect the parameters we can derive from the algorithm that the objective never goes beyond some limitation that is converging – this limitation does not have to be the current objective value. To keep our illustration simple and direct, we use the first way of thinking henceforth.

The reader may expect that, at this point, we may proceed into the inspection

of a mathematical formulation of such guaranteed minimization property. But the truth is that we do not need to – with the previous observations at hand, we could just start to design an objective function that keep bounding the stochastic complexity and the maximum value. In turn, this is precisely the idea of regularization. But this is certainly not the whole story – we did not show what kind of regularization term could be used to bound the stochastic complexity and the maximum value yet. One thing for sure, this will certainly make the bounds tighter, unlike the previous theories in which the hypothesis class is kept static.

To begin our journey, let us first define the notion of sublevel hypothesis class with respect to a regularization term $r(h)$, with first thinking of minimization guarantee.

Definition 19 (Sublevel hypothesis class)**.** *Assuming we are using a multi-objective optimization algorithm, whose objective is* $(\frac{1}{m}\sum_{(x,y)\in S} E(h,x,y), r(h))$*. If the algorithm is currently at* h_0*, then the sublevel hypothesis class* H_0 *is a subclass of the prescribed hypothesis class* H *defined as*

$$H_0 = \left\{ h \in H \,\middle|\, \frac{1}{m}\sum_{(x,y)\in S} E(h,x,y) \leq \frac{1}{m}\sum_{(x,y)\in S} E(h_0,x,y), \text{ and } r(h) \leq r(h_0) \right\}, \tag{3.12}$$

in which S *is sample of size* m *drawn i.i.d. from some distribution* D*.*

The first term, needless to say, guarantees minimization of the first term on the right hand side of equation 3.10. What we would hope is to a way to construct $r(h)$ such that the second and the third terms – stochastic complexity and the upper bound M – could be minimized, so as to get a guarantee on the minimization of its left hand side – the true expectation. Notice that scalarization of both the

objectives in the definition above gives the optimization problem in equation 1.6.

Fortunately, the stochastic complexity $\mathcal{C}_m(K)$ can be upper bounded by M. What left to us is to hopefully bound M using $r(h)$, such that the algorithm provides good generalization. This follows with the theorem below, but the reader should be cautious that this theorem is merely a hope in designing $r(h)$. It will not provide a sensible bound if we replace the stochastic bound in equation 3.10, since the left-hand is always smaller than $\mathcal{L} \cdot M$.

Theorem 3 (Upper bound of stochastic complexity)**.** *Let F be a family of non-negative functions mapping from $\mathcal{U}$ to $\mathbb{R}$ upper bounded by some value $M \geq 0$. Then the following holds*

$$\begin{aligned} \hat{\mathfrak{C}}_S(F) &\leq \frac{M}{2}, \\ \mathfrak{C}_m(F) &\leq \frac{M}{2}. \end{aligned} \tag{3.13}$$

Proof. Define indicator function

$$\mathbb{1}\{\eta \geq 0\} = \begin{cases} 0, & \text{if } \eta \geq 0; \\ 1, & \text{otherwise.} \end{cases} \tag{3.14}$$

Then, by symmetry of noise distribution, we know that $\mathrm{E}_{\eta \sim N}\left[\mathbb{1}\{\eta \geq 0\}|\eta|\right] = \sigma_\eta^1/2$. Also note that by non-negativity of the function family F, we know that $\eta f(u) \leq$

$\mathbb{1}\{\eta \geq 0\}|\eta|M$. Thus, the following derivation holds

$$\begin{aligned}
\hat{\mathfrak{C}}_S(F) &= \frac{1}{\sigma_\eta^1} \mathrm{E}_{\boldsymbol{\eta}} \left[\sup_{f \in F} \frac{1}{m} \sum_{i=1}^{m} \eta_i f(u_i) \right] \\
&\leq \frac{1}{\sigma_\eta^1} \mathrm{E}_{\boldsymbol{\eta}} \left[\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{\eta \geq 0\} |\eta_i| M \right] \\
&= \frac{M}{\sigma_\eta^1} \mathrm{E}_{\eta} \left[\mathbb{1}\{\eta \geq 0\} |\eta| \right] \\
&= \frac{M}{2}.
\end{aligned} \tag{3.15}$$

The same inequality holds for $\mathfrak{C}_m(F)$ as well since it is an expection over $\hat{\mathfrak{C}}_S(F)$. $\square$

As of now, all the ideas of this thesis regarding PAC-learning has been introduced. The result indicates that the ability of generalization can hopefully depend on M – the upper bound of the hypothesis distance with respect to which a well-defined energy is $\mathcal{L}$-Lipschitz to – in a sublevel hypothesis class that is determined by the current minimization algorithm state h_0. This gives a very important tool for deriving the regularization term when the loss function of a problem is determined – minimizing a function of parameters that upper bounds M of the sublevel hypothesis class. The next section discuss specific implications of this idea.

One other issue to be noted is that M does not only function as a hope for bounding the generalization error. Since it is the upper bound of a distance measurement between a hypothesis and the concept class, the effect of itself being small is that the sublevel hypothesis class with the algorithm achived great consistency with the approximation of the concept class. This is another fold of the story, although relatively trivial, regarding how crucial M is for the theory on energy-based models.

3.3 Distance Decomposition

From the previous sections, we have recognised that the hope of generalizatioin for an energy-based model solely depends on M. However, we may still want to derive some more specific upper bound on M so that a distribution-independent regularization term $r(h)$ can be given. In this section we will study the case when κ is some norm, and the hypotheses are parameterized by a vector $w \in \mathbb{R}^m$ such that $h(x) = h(w, x)$. This encopes almost all energy-based models used in practice.

Before we talk about the distance k, we need a tool which is the triangle inequality for hypothesis distance.

Theorem 4 (Triangle inequality of hypothesis distance)**.**

$$\forall h_1, h_2 \in H, k(C(x), h_1(x)) \leq k(C(x), h_2(x)) + \kappa(h_1(x), h_2(x)).$$

Proof. Let $c^* = \operatorname{argmin}_{c \in C} \kappa(c(u), h_2(u))$, then

$$\begin{aligned} k(C(x), h_1(x)) &= \inf_{c \in C} \kappa(c(u), h_1(u)) \\ &\leq \kappa(c^*(u), h_1(u)) \\ &\leq \kappa(c^*(u), h_2(u)) + \kappa(h_1(u), h_2(u)) \\ &= \inf_{c \in C} \kappa(c(u), h_2(u)) + \kappa(h_1(u), h_2(u)) \\ &= k(C(x), h_2(x)) + \kappa(h_1(x), h_2(x)). \end{aligned} \tag{3.16}$$

□

Let k be the hypothesis distance that the energy is $\mathcal{L}$-Lispchitz to, and define

$h^* = \text{argmin}_{h \in H_0} \max_x \{k(C(x), h(x))\}$, the triangle inequality states that

$$k(C(x), h(x)) \leq k(C(x), h^*(x)) + \kappa(h(x), h^*(x)). \tag{3.17}$$

This equation is particularly interesting, since $K(C(x), h^*(x))$ is like a distance between the current sublevel hypothesis class and the target concept class – measured in terms of the maximum over input space – which pretty much like a measurement of the consistency betweeen sublevel hypothesis class H_0 and the concept class C. Taking the maximum of both sides over x, we get

$$M = \max_x \{K(C(x), h(x))\} \leq \max_x \{K(C(x), h^*(x))\} + \max_x \{\kappa(h(x), h^*(x))\}. \tag{3.18}$$

If we match this to the bias-variance trade-off in many machine learning texts[5], then the two terms on the right hand side of the inequalities behaves very much like a bias and a variance, respectively.

If we prescribe the algorithm a large-capacity sublevel hypothesis class, then $K(C(x), h^*(x))$ is usually small. The problem of generalization then falls into how can we make $\max_x \{\kappa(h(x), h^*(x))\}$ small enough. If the hypotheses are parameterized by a parameter w and define $h(w^*, x) = h^*(x)$, Then what we would hope for is a Lipchitz-like inequality such has $\kappa(h(w, x), h(w^*, x)) \leq \mathcal{L}(x)\kappa_w(w, w^*)$, in which κ_w is a metric for the parameters w. This will give the bounding form

$$\max_x \{\kappa(h(x), h^*(x))\} \leq \max_x \{\mathcal{L}(x)\}\kappa_w(w, w^*), \tag{3.19}$$

such that $\kappa_w(w, w^*)$ could be used directly as the regularization term $r(w)$, if w^* is known.

However, in reality there is no way to know about the true w^*. There are two ways to deal with this – either using again the triangle inequality

$$\kappa_w(w, w^*) \leq \kappa_w(w, 0) + \kappa_w(w^*, 0), \tag{3.20}$$

such that we regularize using $\kappa_w(w, 0)$, or using some approximation w^c of w^* such that $\kappa_w(w^*, w^c)$ is small and again by triangle inequality we regularize using $\kappa_w(w, w^c)$. The former explains all the norm (and metric) based regularization used in literature, and the latter explains a possible role unsupervised learning especially deep learning could play – their learnt parameters using prior knowledge could be good choices of w^c. The next chapter will provide experimental results using the latter idea.

However, despite all the excitement we had above, we had not solved the problem in seeking $\mathcal{L}(x)$ in equation 3.19 in the first place. The feasibility of such a decomposition is probably dependent on the specific definition of the metric κ. Here we provide a theorem for the case that κ is some norm, which is perhaps the most used form of regularization in literature.

Theorem 5 (Metric decomposition for norm)**.** *For (sub-)differentiable hypothesis functions $h(w, x) \in \mathbb{R}^n$ with parameters $w \in \mathbb{R}^m$, the following inequality decomposition holds*

$$\forall x, \quad \kappa(h(w, x), h(w^*, x)) \leq \kappa_w(w, w^*) \cdot \mathcal{L}(x) \tag{3.21}$$

for metrics k and k_w defined as

$$\kappa(h(w, x), h(w^*, x)) = \|h(w, x) - h(w^*, x)\|_q^q, \quad \kappa_w(w, w^*) = \|w - w^*\|_p^q, \tag{3.22}$$

and $\mathcal{L}(x)$ *defined as*

$$\forall x, \quad \mathcal{L}(x) = \sum_{i=1}^{n} [l_i(x)]^q, \qquad \textit{with } l_i(x) = \sup_w \|\nabla_w h_i(w, x)\|_r \,, \tag{3.23}$$

where

$$\frac{1}{p} + \frac{1}{r} = 1. \tag{3.24}$$

Proof. With the definition of $l_i(x)$, assuming the hypothesis functions are differentiable, first we prove that

$$|h_i(w, x) - h_i(w^*, x)| \leq l_i(x)\|w - w^*\|_p. \tag{3.25}$$

By the mean value theorem, it is easy to know that there exist some $0 \leq \alpha \leq 1$ such that

$$\forall x, \quad h_i(w, x) - h_i(w^*, x) = \nabla_w h((1-\alpha)w + \alpha w^*) \cdot (w^* - w). \tag{3.26}$$

Thus, by Hölder's inequality[7], if $1/p + 1/r = 1$ we know that $\forall x$,

$$\begin{aligned} |h_i(w, x) - h_i(w^*, x)| &= |\nabla_w h((1-\alpha)w + \alpha w^*) \cdot (w^* - w)| \\ &\leq \|\nabla_w h((1-\alpha)w + \alpha w^*)\|_r \|w^* - w\|_p \\ &\leq l(x)\|w - w^*\|_p. \end{aligned} \tag{3.27}$$

As a result, $k(h(w,x),h(w^*,x))$ can be bounded as

$$\begin{aligned}
\kappa(h(w,x),h(w^*,x)) &= \|h(w,x)-h(w^*,x)\|_q^q \\
&= \sum_{i=1}^{n} |h_i(w,x)-h_i(w^*,x)|^q \\
&\leq \sum_{i=1}^{n} [l(x)\|w-w^*\|_r]^q \\
&= \|w-w^*\|_r^q \cdot \left(\sum_{i=1}^{n}[l(x)]^q\right) \\
&= \|w-w^*\|_r^q \cdot \mathcal{L}(x) \\
&= \kappa_w(w,w^*)\cdot\mathcal{L}(x).
\end{aligned} \tag{3.28}$$

The theorem also holds for sub-differentiable hypothesis class by extending the mean value theorem to subgradients. □

The theorem tells us that, if the energy is $\mathcal{L}$-Lipschitz to the q-th power of a q-th norm, a good regularization parameter should be chosen as $r(h) = r(w) = \|w-w^c\|_2^q$, in which w^c is an approximation of the best parameter and if it is not available, we can let $w^c = 0$.

It is interesting to know that regularization with the l_2 norm of the training parameter seems to always give a good bound on generalization error. However, one should admit that because of certain asymptotic equivalence of norms, we can choose arbitrary norm-based regularization depending on some other prior requirements – such as sparsity from l_1 norm. As far as this theory concerns, any norm will be good candidates for regularization.

Another interesting fact from the proof of the theorem is that the dimension of the output and the dimension of the parameter both determine the complexity to

some extend, since $\mathcal{L}(x)$ is bounded with a sum of n items of $[l_i(x)]^q$, and $l_i(x)$ is also bounded with a sum of the squares of each dimension of $\nabla_w h_i(w, x)$. Thus, in designing a learning system, it is generally not a good idea to use a large dimension for the output and a large amount of parameters.

As of now, we have provided a complete theory of PAC-learning for energy-based models. The decision on whether a learning algorithm could be bounded falls onto the possibility of finding a hypothesis distance which the loss function is L-Lipschitz to. Then, if the metric is in the form of a norm, l_2 regularization (or any other norm-based regularization) can be applied to it for the purpose of achieving good generalization.

3.4 Central Regularization

There was a particular noval and interesting idea follows from our introduction to a theory for energy-based models, that is to use an approximation vector w^c to approximate the best answer w^*, and then regularize with the following regularizer:

$$r(h) = r(w) = \|w - w^c\|_2^2. \tag{3.29}$$

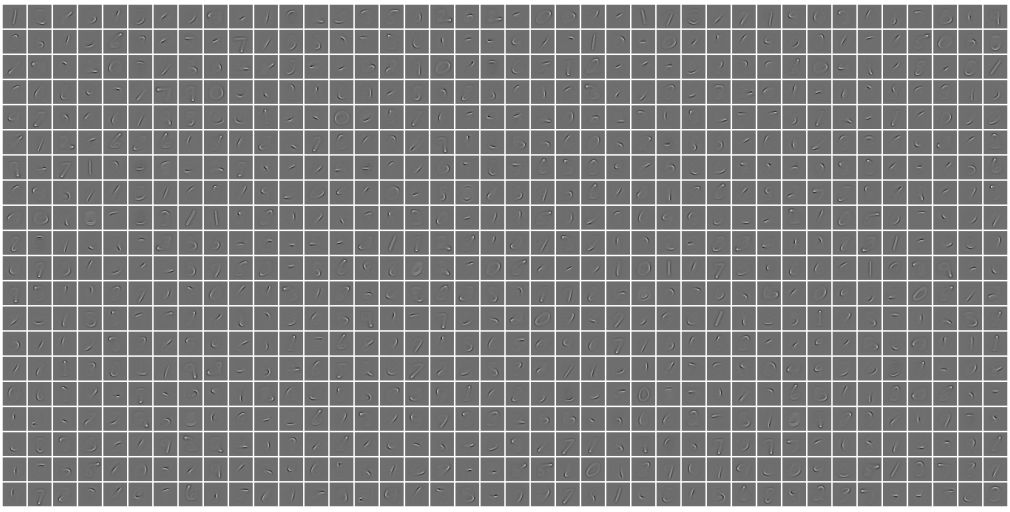
We call this definition of the regularizer central regularization.

Following the recent developments of deep learning and feature learning, we have already got a set of nicely working algorithms that use various prior knowledge for identifying w^c, for different kinds of data. In this thesis we will experiment on one of them – predictive sparse decomposition[8]. It uses a alternating direction method in conjunction with fast iterative shrinkage thresholding (FISTA)[8] to

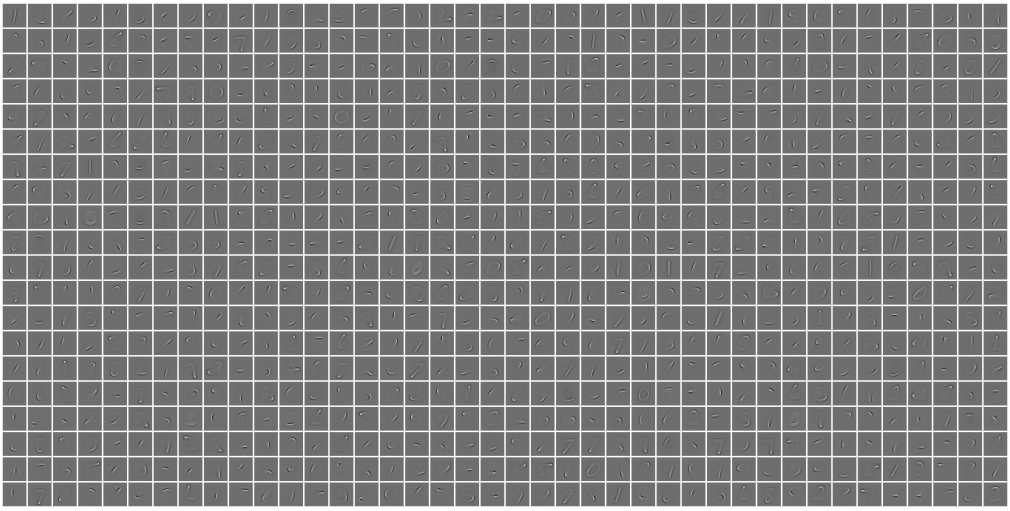
optimize the following objective

$$E(W, V, x, z) = \|y - Wz\|_2^2 + \lambda\|z\|_1 + \|f(V, x) - z\|_2^2, \tag{3.30}$$

in which W is the dictionary and $f(V, x)$ is an encoder.



(a) Decoders



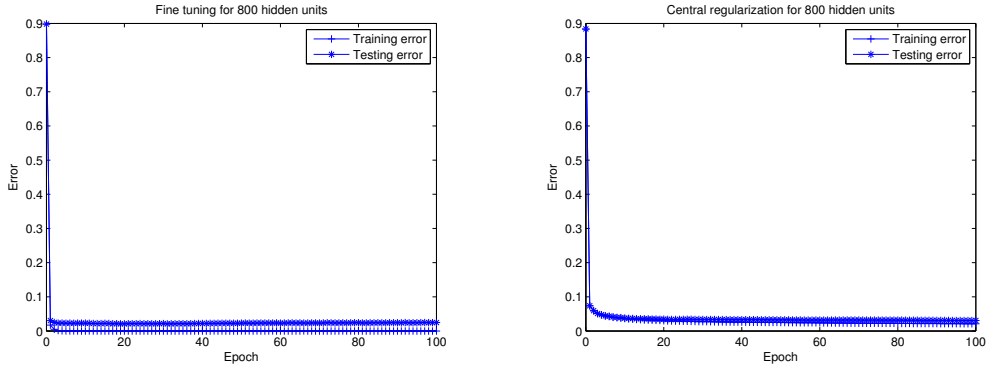
(b) Encoders

Figure 3.1: The pretrained predictive sparse decomposition autoencoder

For our purpose we use the following smoothed version of rectified linear unit

$$f(V, x) = \log\{1 + \exp(Vx)\}. \tag{3.31}$$

The dataset used is a 32×32 expansion of the MNIST dataset[12]. It contains 60,000 training images and 10,000 testing images of handwritten digits. Figure 3.1 shows the visualization of W (the decoder) and V (the encoder) learnt using predictive sparse decomposition for 800 dictionary entries. The output of the encoder $f(V, x)$ will then be connected to a linear layer and a cross-entropy loss to train a 2-layer neural network model.



(a) Fine tuning (b) Central regularization

Figure 3.2: Comparison between fine tuning and central regularization

Our first experiment is between fine tuning and central regularization. In the experiment, we initialize the weights of the first layer using the learnt encoder weights V, and then optimize the neural network altogether. Central regularization added a term $r(w) = \|w - w^c\|$ with $w^c = V$ for the first layer. The regularization parameter λ is 0.05. From figure 3.2 we can see that while achieving similar performance, central regularized model observe closer values for the training erro and testing error. This is exactly what to be expected from any regularization

method – good generalization.

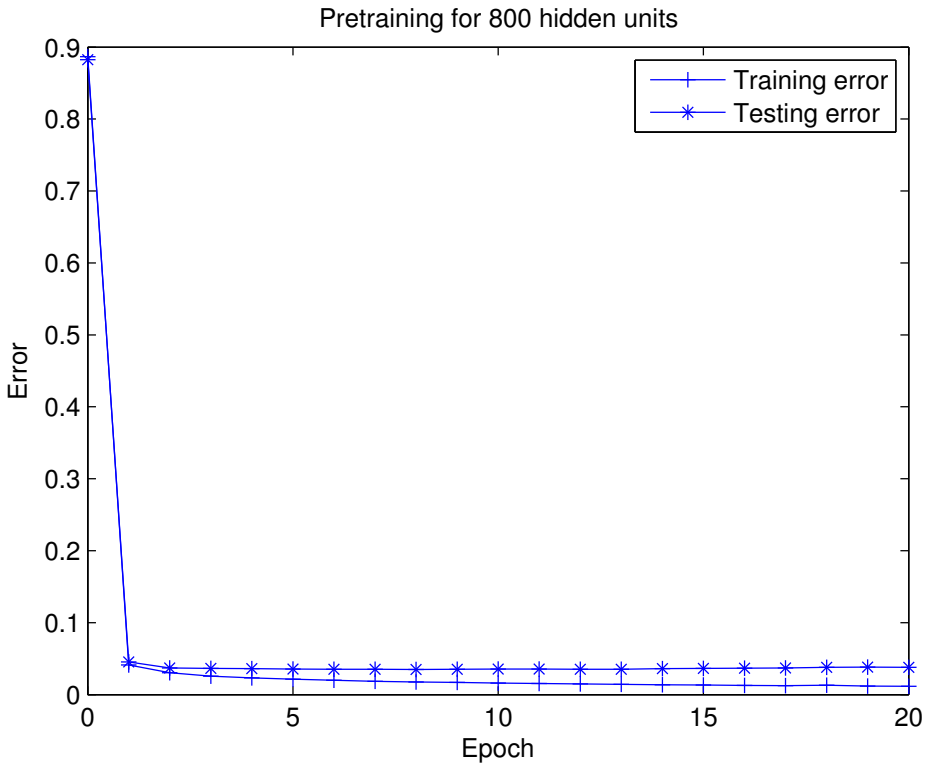
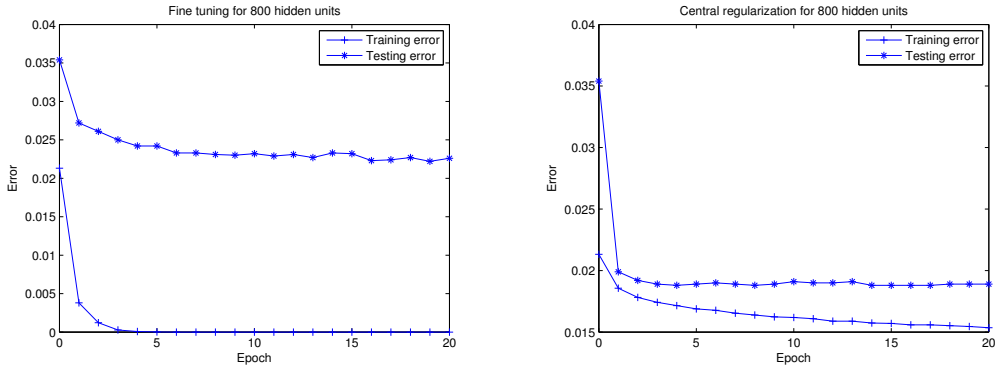


Figure 3.3: Pretraining performance

Our second experiment is concerning what would happen if we embed a pretraining stage in the algorithm. That is, we keep the first layer fixed, pretrain only the second layer and then pass this whole structure to fine tuning and central regularization. This provides comaprisons of fine tuning and central regularization in the algorithm's near-optimal stage, from a common starting point. Figure 3.3 is the result of this pretraining stage, which makes a common ground of 3.5% testing error and 2.1% training error for the comparison.

Figure 3.4 shows the comparison between fine tuning and central regularization after pretraining. The common ground given by the pretraining stage provides a much nicer comparison here. Central regularization not only makes the curves



(a) Fine tuning (b) Central regularization

Figure 3.4: Comparison between fine tuning and central regularization after pre-training

between training error and testing error closer, but also provides a lower testing error. This suggests that a good property of regularization – better generalization will make the testing error small. The testing error we achieved after 20 epoches of central regularization is about 1.9%, although worse than those state-of-the art ones trained with distortions[22], but better than similar architectures trained with merely norm-based regularization[10].

Chapter 4

Conclusion

This thesis studies the PAC-learning theory of energy-based models. We began our introduction with the three elements of energy-based models – energy, discrimination and regularization. Energy is defined as the objective to be minimized during inference, and discrimination measures the difference of energy between the incorrect outputs and the correct outputs. In a learning algorithm we wish to achieved small loss on an energy associated with some margin, and add regularization to the optimization objective to ensure small generalization error. But how to design a regularization term is not clear, unless some theory on generalization error can be provided.

The theory we chose to expand ourselves on the probably approximately correct (PAC) learning theory. Starting with concentration inequalities and a definition of complexity, we establish that the generalization error is bounded by an upper bound of the energy. To ensure that this upper bound exists, we make an observation that at each step of an optimization algorithm, the hypothesis class that the algorithm is able to explore is confined by its current objective value. This com-

bined together the PAC-learning model and algorithmic behaviour of energy-based models.

Then, if the energy is Lipschitz continuous to some metric between hypothesis, we can measure this generalization error bound interms of the distance between the current hypothesis and the target class. A first triangle inequality reveals a similar idea of bias-variance trade-off, and a second one reveals why we do metric (or norm) based regularziation. A novel idea of central regularization is also provided, and our experiments show good practical results of this regularization scheme.

The future of this work may include a study of the theory based on finite variance and mean of the energy rather than an upper bound, and the possibility of progressive central regularization when a deep learning model and a supervised model are trained together in parallel.

Bibliography

[1] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, 3:463–482, Mar. 2003.

[2] S. N. Bernstein. On a modification of chebyshev's inequality and of the error formula of laplace. *Math ḿatique des Annales Scientifiques des Institutions Savantes de lUkraine*, 1924.

[3] S. Boucheron, G. Lugosi, and O. Bousquet. Concentration inequalities. In O. Bousquet, U. von Luxburg, and G. Rätsch, editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 208–240. Springer, 2003.

[4] C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, Sept. 1995.

[5] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Comput.*, 4(1):1–58, Jan. 1992.

[6] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[7] O. Hölder. Üeber einen mittelwertsatz. *Nachrichten von der Königl. Gesellschaft der Wissenschaften und der Georg-Augusts-Universität zu Göttingen*, 2:38–47, 1889.

[8] K. Kavukcuoglu, M. Ranzato, and Y. LeCun. Fast inference in sparse coding algorithms with applications to object recognition. *CoRR*, abs/1010.3467, 2010.

[9] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, MA, USA, 1994.

[10] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[11] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. In G. Bakir, T. Hofman, B. Schölkopf, A. Smola, and B. Taskar, editors, *Predicting Structured Data*. MIT Press, 2006.

[12] Y. Lecun and C. Cortes. The MNIST database of handwritten digits.

[13] Y. LeCun and F. Huang. Loss functions for discriminative training of energy-based models. In *Proc. of the 10-th International Workshop on Artificial Intelligence and Statistics (AIStats'05)*, 2005.

[14] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

[15] P. Massart. Some Applications of Concentration Inequalities to Statistics. *Annales de la Faculté des Sciences de Toulouse*, IX(2):245–303, 2000.

[16] D. A. McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, COLT' 98, pages 230–234, New York, NY, USA, 1998. ACM.

[17] C. McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics*, London Mathematical Society Lecture Note Series 141, pages 148–188. Cambridge University Press, 1989.

[18] M. Mohri, A. Rostamizadeh, and T. Ameet. *Foundations of Machine Learning*. MIT Press, Cambridge, Massachusetts, 2012.

[19] N. Sauer. On the density of families of sets. *J. Comb. Theory, Ser. A*, 13(1):145–147, 1972.

[20] M. Seeger. Pac-bayesian generalization error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3:233–269, 2002.

[21] J. Shawe-Taylor and R. C. Williamson. A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, COLT '97, pages 2–9, New York, NY, USA, 1997. ACM.

[22] P. Y. Simard, R. Szeliski, J. Benaloh, J. Couvreur, and I. Calinov. Using character recognition and segmentation to tell computer from humans. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition - Volume 1*, ICDAR '03, pages 418–, Washington, DC, USA, 2003. IEEE Computer Society.

[23] J. S. Taylor, P. Bartlett, R. C. Williamson, and M. Anthony. Structural Risk Minimization over Data-Dependent Hierarchies. *IEEE Trans. Inf. Theory*, 44(5):1926–1940, 1998.

[24] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, Nov. 1984.

[25] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

[26] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[27] J. Weston and C. Watkins. Multi-class support vector machines, 1998.