LETTER Communicated by Kenji Fukumizu

Dynamics of Learning Near Singularities in Layered Networks

Haikun Wei
weihaikun@brain.riken.jp
RIKEN Brain Science Institute, Saitama, 3510198, Japan, Southeast University, Nanjing, 210096, China, and Kyushu Institute of Technology, Kitakyushu 8080196, Japan

Jun Zhang
junz@umich.edu
RIKEN Brain Science Institute, Saitama, 3510198, Japan, and University of Michigan, Ann Arbor, MI 48109, U.S.A.

Florent Cousseau
florent@mns.k.u-tokyo.ac.jp
RIKEN Brain Science Institute, Saitama, 3510198, Japan, and University of Tokyo, Chiba, 2778561, Japan

Tomoko Ozeki
tozeki@tokai.ac.jp
RIKEN Brain Science Institute, Saitama, 3510198, Japan, and Tokai University, Kanagawa, 2591292, Japan

Shun-ichi Amari
amari@brain.riken.jp
RIKEN Brain Science Institute, Saitama, 3510198, Japan

We explicitly analyze the trajectories of learning near singularities in hierarchical networks, such as multilayer perceptrons and radial basis function networks, which include permutation symmetry of hidden nodes, and show their general properties. Such symmetry induces singularities in their parameter space, where the Fisher information matrix degenerates and odd learning behaviors, especially the existence of plateaus in gradient descent learning, arise due to the geometric structure of singularity. We plot dynamic vector fields to demonstrate the universal trajectories of learning near singularities. The singularity induces two types of plateaus, the on-singularity plateau and the near-singularity plateau, depending on the stability of the singularity and the initial parameters of learning. The results presented in this letter are universally applicable to a wide class of hierarchical models. Detailed stability analysis of the dynamics

of learning in radial basis function networks and multilayer perceptrons will be presented in separate work.

1 Introduction

There is a family of layered neural models with a number of hidden units. All of these hidden units share the same activation function, and the output unit linearly combines the signals from them. Typical examples are the multilayer perceptrons (MLPs) and radial basis function (RBF) networks. The gaussian mixture model of probability density also belongs to this class.

These models possess permutation symmetry such that the input-output behavior is invariant under the permutation of hidden units (Chen, Lu, & Hecht-Nielsen, 1993). That is, if we exchange the parameters of two hidden units, together with the weights they apply to the output units, the input-output map of the network is unchanged. When the activation function of the hidden units is an odd or even function, there exists another type of symmetry caused by the sign change invariance. The symmetry causes the model parameters to be unidentifiable. When two hidden units overlap and have identical parameters, there are many equivalent ways of partitioning their output weights such that the total network input-output behavior remains invariant. This defines a continuous region in the parameter space in which all models share the same input-output relation, although their model parameters are different.

The geometric structure of the parameter space of a hierarchical model has been studied from the point of view of unidentifiability by many researchers (Sussmann, 1992; Chen et al., 1993; Kurková & Kainen, 1994; Fukumizu, 1996). It was noted that the symmetry gives rise to a singular structure in the parameter space. When a model is disturbed by noise, it becomes a statistical model, in which the Fisher information matrix plays the role of a Riemannian metric in the space (Amari & Nagaoka, 2000). The Fisher information matrix degenerates on the subsets of unidentifiability, where the space collapses. More strongly, when we consider the space of input-output behaviors, which can be obtained by dividing the parameter space into equivalence classes by the input-output equivalence relation, then the resultant space has algebraic singularities (Amari, Park, & Ozeki, 2006).

The conventional method of statistical inference relies on the regularity conditions, which include the existence of a nonsingular Fisher information matrix, and the Cramér-Rao paradigm holds. The maximum-likelihood estimator (MLE) is asymptotically efficient, and its accuracy (measured by the error-covariance matrix) converges to zero with the order of n^{-1} as the number n of observed examples increases. However, this does not hold at a singularity. The asymptotic optimality of MLE is not guaranteed, nor does the n^{-1} convergence of the parameters hold. This is also known in the

statistical community (e.g., Hartigan, 1985; Dacunha-Castelle & Gassiat, 1997; Amari, Park, & Ozeki, 2002). Hagiwara (2002) and Fukumizu (1996, 2003) gave a statistical evaluation of the estimation of the parameters in such singular models. Watanabe (2001a & 2001b) and Watanabe and Amari (2003) and his colleagues gave a detailed analysis of the behaviors of estimators from the Bayesian point of view where techniques from modern algebraic geometry were used effectively.

The dynamics of learning in layered models was studied in Amari (1967) and in Heskes and Kappen (1991) for regular cases. However, it is known that layered models have strange behaviors different from those in regular statistical models. A statistical-physical approach has clarified that this phenomenon arises from symmetry (see, e.g., Riegler & Biehl, 1995; Saad & Solla, 1995a, 1995b; Biehl, Riegler, & Wöhler, 1996; Inoue, Park, & Okada, 2003, 2004; Park, Inoue, & Okada, 2003, 2005; see also Biehl & Schwarze, 1995; Biehl & Caticha, 2002; Freeman & Saad, 1997a, 1997b; Huh, Oh, & Kang, 2000). These studies show that the plateau phenomenon is ubiquitous in such hierarchical systems in the thermodynamical limit, where the number of examples and the number of parameters increase in proportionality. The plateau phenomenon has been widely observed in simulation studies of such systems.

It has been gradually recognized that the plateau phenomenon arises from the geometric singularity of the parameter space. Fukumizu and Amari (2000) gave a detailed analysis of the shape of the cost function in the neighborhood of a singularity and elucidated that the plateau reflects the random walk of model parameters on the singularity. More detailed theoretical studies have appeared in Amari and Ozeki (2001) and in Cousseau, Ozeki, and Amari (in press), while Park et al. (2003, 2005) gave simulation results. The odd behaviors of estimation and learning in such hierarchical models are reviewed in Amari et al. (2006) in detail.

Theoretical analysis of the trajectories of dynamics of learning near singularities was first studied in the case of a gaussian mixture in Amari et al. (2006). For the special case where the teacher is on a singularity, similar trajectories were also found for MLPs (Cousseau et al., in press) and RBF networks (Wei & Amari, 2006). Following these analyses on the stability and dynamical flows near a singularity, this letter presents two new results. First, we prove that such dynamical flows (trajectories of learning) are common and universal in various kinds of hierarchical models, regardless of whether the teacher is on a singularity, that is, whether the model is redundant. Second, we elucidate the ubiquitous mechanism underlying the plateau phenomenon in online gradient learning in such hierarchical systems, including MLPs and RBF networks.

This letter outlines a general framework under which singularities and plateaus in learning can be treated. Detailed stability analyses of RBF networks and MLPs will be given in separate work, in which more specific effects of singularities on learning behaviors will be presented in detail.

The rest of the letter is organized as follows. Section 2 summarizes the paradigm of learning for hierarchical models, including MLPs, RBF networks, and gaussian mixtures. The general online learning equations and singular region are given first. Then a new coordinate system is introduced, in which the learning equation and singularity have new and transparent expressions for later analyses. In section 3, we derive the trajectories of learning near singularity and plot the dynamic vector fields based on the stability analysis. Section 4 is devoted to a detailed explanation of how singularity gives rise to on-singularity and near-singularity plateaus. The relation between the plateau phenomenon and trajectories of learning is revealed in this section. Section 5 contains conclusions and discussions.

2 Learning Paradigm in Hierarchical Networks

2.1 Two-Layer Networks. The hierarchical neural models discussed in this letter have a two-layer structure in which the units of the first layer (hidden layer) receive input signals $\mathbf{x} = [x_1, x_2, \ldots, x_n] \in R^n$, where $\mathbf{x}$ is a column vector.

The output of the ith hidden neuron (also called hidden unit i) in the hierarchical model is $\phi(\mathbf{x}, \mathbf{J}_i)$, where vector $\mathbf{J}_i \in R^m$ represents the parameters for specifying the activation function ϕ, and m denotes the number of parameters. The last layer has one output unit that computes a linear combination of the activations of the hidden units, so the input-output mapping of the model (or the output function) with k hidden units is written as

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=1}^{k} w_i \phi(\mathbf{x}, \mathbf{J}_i), \tag{2.1}$$

where w_i denotes the weight from hidden unit i to the output and vector $\boldsymbol{\theta} = [\mathbf{J}_1, w_1, \ldots, \mathbf{J}_k, w_k] \in R^{(m+1)k}$ represents all the parameters in the model. These parameters are modifiable, and learning is carried out by modifying $\boldsymbol{\theta}$. Note that m (dimension of $\mathbf{J}_i$) can be different according to the different neural models.

We show three typical cases of the two-layer hierarchical models having the same structure:

1. **Radial basis function (RBF) networks**. The RBF network has the activation function

$$\phi(\mathbf{x}, \mathbf{J}_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma_i^2}\right), \tag{2.2}$$

where $\mathbf{J}_i = [\boldsymbol{\mu}_i, \sigma_i]$, σ_i is the width parameter that controls the spread of the function around the center $\boldsymbol{\mu}_i \in R^n$, and $\|\cdot\|$ is the Euclidean norm. Note that here, $\mathbf{J}_i \in R^{n+1}$, that is, $m = n + 1$.

2. **Multilayer perceptrons (MLPs).** The MLP has the sigmoidal activation function

$$\phi(\mathbf{x}, \mathbf{J}_i) = \tanh\left(\mathbf{J}_i^T \mathbf{x}\right), \tag{2.3}$$

where superscript T denotes transposition of a vector. Here, one may use the modified error function for ease of theoretical analyses (Cousseau et al., in press):

$$\phi(\mathbf{x}, \mathbf{J}_i) = erf\left(\mathbf{J}_i^T \mathbf{x}\right), \tag{2.4}$$

$$erf(u) = \sqrt{\frac{2}{\pi}} \int_0^u \exp\left(-\frac{t^2}{2}\right) dt. \tag{2.5}$$

Note that for MLPs, if $\mathbf{J}_i$ includes a bias term (threshold term), then $\mathbf{J}_i \in R^{n+1}$ and $m = n + 1$; otherwise $\mathbf{J}_i \in R^n$.

3. **Gaussian mixtures.** Here, the probability density function $p(\mathbf{x})$ of input $\mathbf{x}$ is calculated by

$$p(\mathbf{x}) = \sum_{i=1}^{k} w_i \phi(\mathbf{x}, \mathbf{J}_i), \tag{2.6}$$

where ϕ is the gaussian function of the form 2.2, and $\mathbf{J}_i \in R^{n+1}$. However, the region of the parameters $w_1, \ldots, w_k$ is restricted by

$$w_i \geq 0, \quad \sum_{i=1}^{k} w_i = 1. \tag{2.7}$$

This case was analyzed in Amari et al. (2006). The population coding with multiple stimuli was also analyzed in this form by Amari and Nakahara (2005).

2.2 Loss Function and Stochastic Gradient Learning. The first two cases, RBF networks and MLPs, are regression problem, where the models are required to imitate the function specified by the teacher:

$$y = f_0(\mathbf{x}). \tag{2.8}$$

Instead of the analytical form of $f_0(\mathbf{x})$, input-output examples $(y_1, \mathbf{x}_1), \ldots, (y_M, \mathbf{x}_M)$ are given, where y_i are the noisy versions of the true outputs,

$$y_i = f_0(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, 2, \ldots, M, \tag{2.9}$$

and ε_i denotes the additive noise. The model parameter $\boldsymbol{\theta}$ is adjusted to fit the training examples. The distributions of training input $\mathbf{x}$ are assumed

to be uncorrelated with noise ε, which the latter is subject to zero mean gaussian distribution.

Generally when the parameter of the student model is $\boldsymbol{\theta}$, $f(\mathbf{x}, \boldsymbol{\theta})$ is different from $y = f_0(\mathbf{x}) + \varepsilon$. Then we define the instantaneous loss function as

$$l(y, \mathbf{x}, \boldsymbol{\theta}) = \frac{1}{2}(y - f(\mathbf{x}, \boldsymbol{\theta}))^2. \tag{2.10}$$

Although there are dozens of methods for minimizing the above loss function, this letter concentrates on the online learning scheme, in which parameter $\boldsymbol{\theta}$ is modified by the stochastic gradient descent algorithm. The learning algorithm is thus

$$\boldsymbol{\theta}(t+1) = \boldsymbol{\theta}(t) - \eta \frac{\partial l(y_t, \mathbf{x}_t, \boldsymbol{\theta}_t)}{\partial \boldsymbol{\theta}}, \tag{2.11}$$

where η is a learning constant. This is a stochastic difference equation. Its behavior is approximated by the continuous time equation averaged over all possible inputs, outputs, and noises. This average is justified by the stochastic approximation when source signals $(y_t, \mathbf{x}_t)$ are ergodic (Amari, 1967, 1977). Thus, we investigate the following averaged learning equation:

$$\dot{\boldsymbol{\theta}}(t) = -\eta \left\langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle. \tag{2.12}$$

In equation 2.12, $\langle \cdot \rangle$ denotes the expectation with respect to the input $\mathbf{x}$ and the corresponding teacher's signal y,

$$\left\langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\rangle = \int \frac{\partial l(y, \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} p_0(y, \mathbf{x})\, dy\, d\mathbf{x}, \tag{2.13}$$

where $p_0(y, \mathbf{x})$ is the joint probability of $(y, \mathbf{x})$ of the teacher's signal,

$$p_0(y, \mathbf{x}) = p_0(\mathbf{x}) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - f_0(\mathbf{x}))^2\right), \tag{2.14}$$

and $p_0(\mathbf{x})$ is the probability density of the training input $\mathbf{x}$.

2.3 Singularity Regions. Since we are interested in the learning dynamics near the singularity, we should first find out where the singular regions are in the parameter space of the model. We focus on the regression models of cases 1 and 2 in section 2.1, but the structure is the same in case 3.

For a wide range of neural models, Amari et al. (2006) discussed the singular regions from the point of view of information geometry (Amari

& Nagaoka, 2000). They showed that in the parameter space of the neural models, there are singular regions where the Fisher information matrix degenerates and some of the parameters are not identifiable. When the model parameter is located in one of these regions, the conventional Cramér-Rao paradigm does not hold, and odd behaviors happen. Next, following the method in that article, we study the singular regions in our hierarchical neural models.

The model output function $f(\mathbf{x}, \boldsymbol{\theta})$ includes the term $w_i\phi(\mathbf{x}, \mathbf{J}_i)$, which vanishes identically when $w_i = 0$. We denote this region by $\mathcal{R}(i) = \{\boldsymbol{\theta} \mid w_i = 0\}$, which is a subspace in the parameter space. Whatever value $\mathbf{J}_i$ takes, $f(\mathbf{x}, \boldsymbol{\theta})$ is the same in $\mathcal{R}(i)$. That is, $\mathbf{J}_i$ does not have identifiability. The Fisher information degenerates in $\mathcal{R}(i)$. In this region, the ith unit does not play any role, so it is eliminable. Hence $\mathcal{R}(i)$ is called the elimination singularity in this letter. In the case of MLPs where activation function 2.3 or 2.4 is odd and satisfies $\phi(\mathbf{x}, \mathbf{0}) = 0$, the relation $\mathbf{J}_i = \mathbf{0}$ also indicates the vanishing of unit i, so the elimination singularity in MLPs is given by $w_i\mathbf{J}_i = \mathbf{0}$.

When $\mathbf{J}_i = \mathbf{J}_j$ holds, the model output function $f(\mathbf{x}, \boldsymbol{\theta})$ includes terms that reduce to

$$w_i\phi(\mathbf{x}, \mathbf{J}_i) + w_j\phi(\mathbf{x}, \mathbf{J}_j) = (w_i + w_j)\phi(\mathbf{x}, \mathbf{J}_i). \tag{2.15}$$

In this case, when $w_i + w_j = c$ is satisfied by a constant c, the output function is the same whatever value each of w_i and w_j takes. Hence, w_i and w_j lose identifiability. We denote this region by $\mathcal{R}(i, j)$. In this case, since the two hidden units i and j are exactly the same and overlap, $\mathcal{R}(i, j)$ is called the overlap singularity in this letter. In the case of RBF networks, units i and j overlap completely, so that they behave as one unit. In the case of MLPs with odd activation functions, the two units are also regarded as being the same when $\mathbf{J}_i = -\mathbf{J}_j$, because of $\phi(\mathbf{x}, -\mathbf{J}_i) = -\phi(\mathbf{x}, \mathbf{J}_j)$. Hence, $\mathcal{R}(i, j)$ also includes this region. It is immediate to show that the Fisher information degenerates in this region.

The above two (and their intersections) are the only singular regions in RBF networks and gaussian mixture systems where the Fisher information matrix degenerates (Fukumizu, 1996).

In summary, there are two types of singular regions $\mathcal{R}(i, j)$ and $\mathcal{R}(i)$, which are given by

$$\mathcal{R}(i, j) = \{\boldsymbol{\theta} \mid \mathbf{J}_i = \mathbf{J}_j\} \quad \text{or} \quad \{\boldsymbol{\theta} \mid \mathbf{J}_i = \pm\mathbf{J}_j\} \tag{2.16}$$

and

$$\mathcal{R}(i) = \{\boldsymbol{\theta} \mid w_i = 0\} \quad \text{or} \quad \{\boldsymbol{\theta} \mid w_i\mathbf{J}_i = \mathbf{0}\}, \tag{2.17}$$

where the first part (before "or") is for RBF networks and the latter part is for MLPs. This letter mainly deals with the dynamics of online learning near the overlap singularity where any two of the hidden units coincide. Without loss of generality, we discuss the learning dynamics in the neighborhood of $\mathcal{R}(a, b)$ for two hidden units a and b, where $\mathbf{J}_a \approx \mathbf{J}_b$, that is, the two hidden units a and b almost overlap. Thus, the student model can be rewritten as

$$f(\mathbf{x}, \boldsymbol{\theta}) = g(\mathbf{x}, \mathbf{s}) + w_a \phi(\mathbf{x}, \mathbf{J}_a) + w_b \phi(\mathbf{x}, \mathbf{J}_b), \tag{2.18}$$

where $\mathbf{s} = [\mathbf{J}_1, w_1, \ldots, \mathbf{J}_i, w_i, \ldots, \mathbf{J}_k, w_k]$, $i \neq a, i \neq b$, represents the parameters of all the units except a and b, and $g(\mathbf{x}, \mathbf{s}) = \sum_{i=1, i\neq a, i\neq b}^{k} w_i \phi(\mathbf{x}, \mathbf{J}_i)$. Note that $\mathbf{s} \in R^{(m+1)(k-2)}$. In the following, we denote the overlap singularity as $\mathcal{R}_1 = \mathcal{R}(a, b) = \{\boldsymbol{\theta} \mid \mathbf{J}_a = \mathbf{J}_b\}$. The related elimination singularity $\mathcal{R}_2 = \mathcal{R}(a) \bigcup \mathcal{R}(b) = \{\boldsymbol{\theta} \mid w_a w_b = 0\}$ intersects $\mathcal{R}_1$, and we also take this into account.

2.4 Coordinate Transformation. We focus on the dynamical behavior of online learning around the singular region $\mathcal{R} = \mathcal{R}_1 \bigcup \mathcal{R}_2$, where the output function is equivalently represented by using only $k - 1$ hidden units:

$$f(\mathbf{x}, \boldsymbol{\theta}) = g(\mathbf{x}, \mathbf{s}) + w\phi(\mathbf{x}, \mathbf{v}). \tag{2.19}$$

In the above equation, if $\boldsymbol{\theta}$ is on the overlap singularity $\mathcal{R}_1$, then $\mathbf{v} = \mathbf{J}_a = \mathbf{J}_b$ and $w = w_a + w_b$; if $\boldsymbol{\theta}$ is on the elimination singularity $\mathcal{R}_2$ where $w_a = 0$ ($w_b = 0$), then $\mathbf{v} = \mathbf{J}_b$ ($\mathbf{v} = \mathbf{J}_a$) and $w = w_b$ ($w = w_a$).

Let us consider the following coordinates in the neighborhood of $\mathcal{R}_1$, which was introduced and used by Fukumizu and Amari (2000) and by Amari et al. (2006):

$$\mathbf{u} = \mathbf{J}_b - \mathbf{J}_a, \tag{2.20}$$

$$\mathbf{v} = \frac{w_a \mathbf{J}_a + w_b \mathbf{J}_b}{w_a + w_b}, \tag{2.21}$$

$$w = w_a + w_b, \tag{2.22}$$

$$z = \frac{w_a - w_b}{w_a + w_b}, \tag{2.23}$$

and $\mathbf{s}$ is the same as before.

The equations to transform back to the original coordinates are

$$\mathbf{J}_a = \mathbf{v} + \frac{1}{2}(z - 1)\mathbf{u}, \tag{2.24}$$

$$\mathbf{J}_b = \mathbf{v} + \frac{1}{2}(z + 1)\mathbf{u}, \tag{2.25}$$

$$w_a = \frac{1}{2}w(1+z), \tag{2.26}$$

$$w_b = \frac{1}{2}w(1-z). \tag{2.27}$$

Then the model parameter $\boldsymbol{\theta} = [\mathbf{s}, \mathbf{J}_a, w_a, \mathbf{J}_b, w_b]$ becomes $\boldsymbol{\xi} = [\mathbf{s}, \mathbf{v}, w, \mathbf{u}, z]$. Correspondingly, the singular regions $\mathcal{R}_1$ and $\mathcal{R}_2$ are now represented by $\{\boldsymbol{\xi} \mid \mathbf{u} = \mathbf{0}\}$ and $\{\boldsymbol{\xi} \mid z = \pm 1\}$ in the new coordinate system.

Now the student model can be rewritten as

$$f(\mathbf{x}, \boldsymbol{\xi}) = g(\mathbf{x}, \mathbf{s}) + \frac{1}{2}w(1+z)\phi\left(\mathbf{x}, \mathbf{v} + \frac{1}{2}(z-1)\mathbf{u}\right) + \frac{1}{2}w(1-z)\phi\left(\mathbf{x}, \mathbf{v} + \frac{1}{2}(z+1)\mathbf{u}\right). \tag{2.28}$$

Since the activation function ϕ is infinitely differentiable, we can perform the Taylor expansion around $\mathbf{u} = \mathbf{0}$ and get

$$f(\mathbf{x}, \boldsymbol{\xi}) = g(\mathbf{x}, \mathbf{s}) + w\phi(\mathbf{x}, \mathbf{v}) + \frac{1}{8}w(1-z^2)\mathbf{u}^T \frac{\partial \phi^2(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^T}\mathbf{u} + \frac{1}{24}wz(1-z^2)D(\mathbf{x}, \mathbf{v}, \mathbf{u}) + O(\mathbf{u}^4), \tag{2.29}$$

where $D(\mathbf{x}, \mathbf{v}, \mathbf{u}) = \sum_{i,j,k} \frac{\partial \phi^3(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}_i \partial \mathbf{v}_j \partial \mathbf{v}_k}\mathbf{u}_i\mathbf{u}_j\mathbf{u}_k$, $\mathbf{v}_i$ and $\mathbf{u}_i$ are the ith elements of $\mathbf{v}$ and $\mathbf{u}$, respectively. Note that $D(\mathbf{x}, \mathbf{v}, \mathbf{u})$ is of order $O(\mathbf{u}^3)$.

Note that if $\mathbf{u} = \mathbf{0}$, equation 2.29 reduces to the regular model, equation 2.19, with $k-1$ hidden units. In this sense, $\mathbf{s}$, $\mathbf{v}$, and w are parameters to specify the regular model, while parameters $\mathbf{u}$ and z indicate the deviations from the singular regions.

2.5 Equations of Learning. Next we derive the equations of learning in terms of the new parameter $\boldsymbol{\xi} = [\mathbf{s}, \mathbf{v}, w, \mathbf{u}, z]$. The learning equation, 2.12, is rewritten in terms of $\boldsymbol{\xi}$ as

$$\dot{\boldsymbol{\xi}} = -\eta \mathbf{T}\mathbf{T}^T \left\langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \right\rangle, \tag{2.30}$$

where $\mathbf{T} = \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}^T}$ is the Jacobian of the coordinate transformation (see the appendix). Note that this is different from $\dot{\boldsymbol{\xi}} = -\eta \langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \rangle$.

According to equation 2.29, the gradients of $f(\mathbf{x}, \boldsymbol{\xi})$ with respect to $\mathbf{s}$, $\mathbf{v}$, w, $\mathbf{u}$, and z are

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\xi})}{\partial \mathbf{s}} = \frac{\partial g(\mathbf{x}, \mathbf{s})}{\partial \mathbf{s}}, \tag{2.31}$$

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\xi})}{\partial \mathbf{v}} = w\frac{\partial \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}} + \frac{1}{8}w(1-z^2)\mathbf{q}(\mathbf{x}, \mathbf{v}, \mathbf{u}) + O(\mathbf{u}^3), \tag{2.32}$$

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\xi})}{\partial w} = \phi(\mathbf{x}, \mathbf{v}) + \frac{1}{8}(1-z^2)\mathbf{u}^T\frac{\partial \phi^2(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}\partial \mathbf{v}^T}\mathbf{u} + O(\mathbf{u}^3), \tag{2.33}$$

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\xi})}{\partial \mathbf{u}} = \frac{1}{4}w(1-z^2)\frac{\partial \phi^2(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}\partial \mathbf{v}^T}\mathbf{u} + \frac{1}{24}wz(1-z^2)\frac{\partial D(\mathbf{x}, \mathbf{v}, \mathbf{u})}{\partial \mathbf{u}} + O(\mathbf{u}^3), \tag{2.34}$$

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\xi})}{\partial z} = -\frac{1}{4}wz\mathbf{u}^T\frac{\partial \phi^2(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}\partial \mathbf{v}^T}\mathbf{u} + O(\mathbf{u}^3), \tag{2.35}$$

where $\mathbf{q}(\mathbf{x}, \mathbf{v}, \mathbf{u}) = \frac{\partial}{\partial \mathbf{v}}(\mathbf{u}^T\frac{\partial \phi^2(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}\partial \mathbf{v}^T}\mathbf{u})$ and $\frac{\partial D(\mathbf{x}, \mathbf{v}, \mathbf{u})}{\partial \mathbf{u}}$ are both vectors of order $O(\mathbf{u}^2)$. Then we get the negative gradient of the averaged loss function $\langle l(y, \mathbf{x}, \boldsymbol{\xi})\rangle$ with respect to the new parameters:

$$l_{\mathbf{s}}(\boldsymbol{\xi}) = \left\langle e(y, \mathbf{x}, \boldsymbol{\xi})\frac{\partial g(\mathbf{x}, \mathbf{s})}{\partial \mathbf{s}}\right\rangle, \tag{2.36}$$

$$l_{\mathbf{v}}(\boldsymbol{\xi}) = w\left\langle e(y, \mathbf{x}, \boldsymbol{\xi})\frac{\partial \phi(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}}\right\rangle + \frac{1}{8}w(1-z^2)\mathbf{Q}(\mathbf{v}, \mathbf{u}) + O(\mathbf{u}^3), \tag{2.37}$$

$$l_w(\boldsymbol{\xi}) = \langle e(y, \mathbf{x}, \boldsymbol{\xi})\phi(\mathbf{x}, \mathbf{v})\rangle + \frac{1}{8}(1-z^2)\left\langle e(y, \mathbf{x}, \boldsymbol{\xi})\mathbf{u}^T\frac{\partial \phi^2(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}\partial \mathbf{v}^T}\mathbf{u}\right\rangle + O(\mathbf{u}^3), \tag{2.38}$$

$$l_{\mathbf{u}}(\boldsymbol{\xi}) = \frac{1}{4}w(1-z^2)\left\langle e(y, \mathbf{x}, \boldsymbol{\xi})\frac{\partial \phi^2(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}\partial \mathbf{v}^T}\mathbf{u}\right\rangle + \frac{1}{24}wz(1-z^2)\left\langle e(y, \mathbf{x}, \boldsymbol{\xi})\frac{\partial D(\mathbf{x}, \mathbf{v}, \mathbf{u})}{\partial \mathbf{u}}\right\rangle + O(\mathbf{u}^3), \tag{2.39}$$

$$l_z(\boldsymbol{\xi}) = -\frac{1}{4}wz\left\langle e(y, \mathbf{x}, \boldsymbol{\xi})\mathbf{u}^T\frac{\partial \phi^2(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v}\partial \mathbf{v}^T}\mathbf{u}\right\rangle + O(\mathbf{u}^3), \tag{2.40}$$

where $e(y, \mathbf{x}, \boldsymbol{\xi}) = f_0(\mathbf{x}) - f(\mathbf{x}, \boldsymbol{\xi}) + \varepsilon$ is the error between the model and the teacher. In equation 2.37, $\mathbf{Q}(\mathbf{v}, \mathbf{u}) = \langle e(y, \mathbf{x}, \boldsymbol{\xi})\mathbf{q}(\mathbf{x}, \mathbf{v}, \mathbf{u})\rangle$, and is still of order $O(\mathbf{u}^2)$. Note that $l_{\mathbf{u}}$ is of order $O(\mathbf{u})$ and l_z is of order $O(\mathbf{u}^2)$, whereas $l_{\mathbf{s}}$, $l_{\mathbf{v}}$, and l_w are all of order $O(1)$.

Consequently, by using some mathematical operations, we obtain the equations of learning in the new coordinate system as

$$\dot{\mathbf{s}} = l_{\mathbf{s}}, \tag{2.41}$$

$$\dot{\mathbf{v}} = \frac{z^2+1}{2} l_{\mathbf{v}} + \frac{z^2+1}{2w^2} \mathbf{u}\mathbf{u}^T l_{\mathbf{v}} + \frac{z}{w} \mathbf{u} l_w - z l_{\mathbf{u}} - \frac{z^2+1}{w^2} \mathbf{u} l_z, \tag{2.42}$$

$$\dot{w} = \frac{z}{w} \mathbf{u}^T l_{\mathbf{v}} + 2 l_w - \frac{2z}{w} l_z, \tag{2.43}$$

$$\dot{\mathbf{u}} = -z l_{\mathbf{v}} + 2 l_{\mathbf{u}}, \tag{2.44}$$

$$\dot{z} = -\frac{z^2+1}{w^2} \mathbf{u}^T l_{\mathbf{v}} - \frac{2z}{w} l_w + \frac{2(z^2+1)}{w^2} l_z. \tag{2.45}$$

Note that we have omitted the learning rate η in the above equations.

2.6 Critical Line $\mathcal{R}_1^*$ in the Singular Region. In the singular region $\mathcal{R}_1$ where $\mathbf{u} = \mathbf{0}$, for fixed parameters $\mathbf{s}$, $\mathbf{v}$, and w, the output function is $f(\mathbf{x}, \mathbf{s}, \mathbf{v}, w, \mathbf{0}, z)$, but it does not depend on z. So it reduces to the output function with $k-1$ hidden units whose parameters are specified by $(\mathbf{s}, \mathbf{v}, w)$. Let us consider the line $\mathcal{R}_1(\mathbf{s}, \mathbf{v}, w)$ in the parameter space where $(\mathbf{s}, \mathbf{v}, w)$ are fixed, $\mathbf{u} = \mathbf{0}$, and z is arbitrary. All output functions are the same on this line. We can add the region $\mathcal{R}_2(\mathbf{s}, \mathbf{v}, w)$ specified by $z = \pm 1$ with arbitrary $\mathbf{u}$ to this line, still keeping the same output function. Note that $\mathcal{R}_2(\mathbf{s}, \mathbf{v}, w)$ is a pair of n-dimensional subspaces, where $z = \pm 1$ and $\mathbf{u}$ is arbitrary. So the set $\mathcal{R} = \mathcal{R}_1 \bigcup \mathcal{R}_2$ consists of three regions with fixed $(\mathbf{s}, \mathbf{v}, w)$.

Given the teacher function $y = f_0(\mathbf{x})$, assume $(\mathbf{s}^*, \mathbf{v}^*, w^*)$ is its best approximation by the model with $k-1$ units. Let $\mathcal{R}_1^*$ and $\mathcal{R}_2^*$ be the singular regions specified by $(\mathbf{s}^*, \mathbf{v}^*, w^*)$, and put $\mathcal{R}^* = \mathcal{R}_1^* \bigcup \mathcal{R}_2^*$. If $\boldsymbol{\xi}^* = (\mathbf{s}^*, \mathbf{v}^*, w^*, \mathbf{0}, z)$ is a point on the line $\mathcal{R}_1^* = (\mathbf{s}^*, \mathbf{v}^*, w^*)$, then we have

$$\left\langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial \mathbf{s}} \right\rangle \Big|_{\xi=\xi^*} = l_{\mathbf{s}}(\boldsymbol{\xi}^*) = \mathbf{0}, \tag{2.46}$$

$$\left\langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial \mathbf{v}} \right\rangle \Big|_{\xi=\xi^*} = l_{\mathbf{v}}(\boldsymbol{\xi}^*) = \mathbf{0}, \tag{2.47}$$

$$\left\langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial w} \right\rangle \Big|_{\xi=\xi^*} = l_w(\boldsymbol{\xi}^*) = 0. \tag{2.48}$$

Here, since at $\boldsymbol{\xi}^*$ the loss function $l(y, \mathbf{x}, \boldsymbol{\xi})$ does not depend on z, we have

$$\left\langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial z} \right\rangle \Big|_{\xi=\xi^*} = l_z(\boldsymbol{\xi}^*) = 0. \tag{2.49}$$

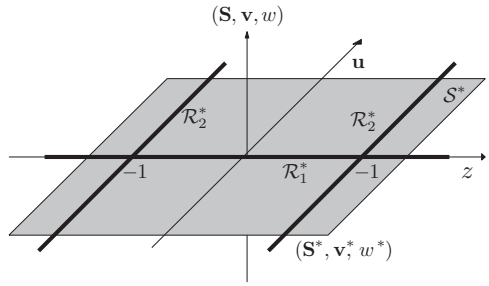


Figure 1: $\mathcal{R}^* = \mathcal{R}_1^* \bigcup \mathcal{R}_2^*$ in the parameter space.

Moreover, the output function does not change its value when we replace $(\mathbf{u}, z)$ by $(-\mathbf{u}, -z)$ because this corresponds to the replacement of unit b by unit a. Hence,

$$\left\langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial \mathbf{u}} \right\rangle \Big|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} = l_{\mathbf{u}}(\boldsymbol{\xi}^*) = \mathbf{0} \tag{2.50}$$

also holds at $\boldsymbol{\xi}^*$.

The above shows that $\mathcal{R}_1^*$ consists of critical points at which the equations of learning in equations 2.41 to 2.45 are all zero. In fact, if we let $l_{\mathbf{s}} = \mathbf{0}, l_{\mathbf{v}} = \mathbf{0}$, and $l_w = 0$ in equations 2.41 to 2.45, then it is quite clear that $\dot{\boldsymbol{\xi}} \mid_{\mathbf{u}=\mathbf{0}} = \mathbf{0}$, and vice versa. So $l_{\mathbf{s}} = \mathbf{0}, l_{\mathbf{v}} = \mathbf{0}$, and $l_w = 0$ is a necessary and sufficient condition of $\mathbf{u} = \mathbf{0}$ being a critical line.

However, although $(\mathbf{s}^*, \mathbf{v}^*, w^*)$ is the optimal approximation of teacher function $y = f_0(\mathbf{x})$, $\mathcal{R}_1^*$ is not necessarily a stable critical line because the Hessian submatrix,

$$\left\langle \frac{\partial^2 l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial \mathbf{u} \partial \mathbf{u}^T} \right\rangle \Big|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*}, \tag{2.51}$$

might not be seminegative definite in general.

Interestingly, $\mathcal{R}_2^*$ is not critical in most cases, and we will show this in section 3.3.

Figure 1 illustrates the shape of $\mathcal{R}_1^*$ (the thick black line on the z-axis) and $\mathcal{R}_2^*$ (the two thick lines parallel to the $\mathbf{u}$-axis, which are n-dimensional) in the parameter space. The gray plane $\mathcal{S}^*$, which includes $\mathcal{R}_1^*$ and $\mathcal{R}_2^*$, is the region determined by $(\mathbf{s}, \mathbf{v}, w) = (\mathbf{s}^*, \mathbf{v}^*, w^*)$, that is, $\mathcal{S}^* = \{(\mathbf{s}, \mathbf{v}, w, \mathbf{u}, z) \mid \mathbf{s} = \mathbf{s}^*, \mathbf{v} = \mathbf{v}^*, w = w^*\}$.

2.7 Stability Analysis. The dynamics of learning near the singularity depends crucially on the stability of the singularity. The stability analysis on

the overlap singularity can be performed by analyzing the Hessian matrix at $\mathcal{R}_1^*$, that is, the matrix

$$\mathbf{F}\Big|_{\xi=\xi^*} = \left\langle \frac{\partial^2 l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi} \partial \boldsymbol{\xi}^T} \right\rangle \Big|_{\xi=\xi^*} \tag{2.52}$$

evaluated at $\boldsymbol{\xi}^* = (\mathbf{s}^*, \mathbf{v}^*, w^*, \mathbf{0}, z)$.

Fukumizu and Amari (2000) have calculated the above Hessian for MLPs and shown that the stability of $\mathcal{R}_1^*$ depends solely on the signature of the Hessian $\langle \frac{\partial^2 l(y,\mathbf{x},\boldsymbol{\xi})}{\partial \mathbf{u} \partial \mathbf{u}^T} \rangle \mid_{\xi=\xi^*}$, which is calculated as

$$\left\langle \frac{\partial^2 l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial \mathbf{u} \partial \mathbf{u}^T} \right\rangle \Big|_{\xi=\xi^*} = (1 - z^2)\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*), \tag{2.53}$$

where

$$\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*) = \frac{1}{4} w^* \left\langle e(y, \mathbf{x}, \boldsymbol{\xi}) \frac{\partial \phi^2(\mathbf{x}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^T} \right\rangle \Big|_{\xi=\xi^*}. \tag{2.54}$$

This results still holds in our general case. For the special case where the teacher is on the singularity $\mathcal{R}_1^*$, we have $f(\mathbf{x}, \boldsymbol{\xi}^*) = f_0(\mathbf{x})$ and $\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*) = \mathbf{0}$. In this case, the whole line of $\mathcal{R}_1^*$ is stable. That is, it is a line attractor. For the general case where the teacher cannot be emulated exactly by the model, the stability of $\mathcal{R}_1^*$ is determined by $(1 - z^2)\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)$.

We summarize this in the following theorem:

Theorem 1. *When the teacher is on the singularity $\mathcal{R}_1^*$, the whole critical line of $\mathcal{R}_1^*$ is stable. When $\mathbf{H}$ has both positive and negative eigenvalues, all points on the critical line $\mathcal{R}_1^*$ are saddles. Otherwise the line $\mathcal{R}_1^*$ is divided into stable and unstable parts. When $\mathbf{H}$ is negative definite, the part $z^2 < 1$ is stable and attractive, whereas the part $z^2 > 1$ is repulsive. When $\mathbf{H}$ is positive definite, the part $z^2 > 1$ is stable and attractive, whereas the part $z^2 < 1$ is repulsive.*

3 Dynamics of Learning

3.1 Trajectories of Learning Near a Singularity. When the matrix $\mathbf{H}$ in equation 2.54 is positive or negative definite, part of $\mathcal{R}_1^*$ is a stable attractor. Here, we analyze the trajectories of learning in the neighborhood of $\mathcal{R}_1^*$. Such trajectories were obtained in the special cases of gaussian mixtures (Amari et al., 2006), multilayer perceptrons (Cousseau et al., in press), and RBF networks (Wei & Amari, 2006) when the teacher is on the singularity. However, in Amari et al. (2006) and Cousseau et al. (in press), the gradient flows were calculated in terms of the new coordinates and are different from the original gradients. Here, we give ubiquitous trajectories in terms

of the original gradients by using the new coordinate system in general cases, where the teacher may be located at any place.

Let us consider the subspace $\mathcal{S}^*$ determined from $\mathbf{s}^*$, $\mathbf{v}^*$, and w^*, which includes $\mathcal{R}^* = \mathcal{R}_1^* \bigcup \mathcal{R}_2^*$. It is shown by the gray plane in Figure 1. It includes two free parameters $\mathbf{u}$ and z, and the other parameters $\mathbf{s}$, $\mathbf{v}$, and w are determined from equations 2.46 to 2.48:

$$\mathbf{s} = \mathbf{s}^*, \ \mathbf{v} = \mathbf{v}^*, \ w = w^*. \tag{3.1}$$

Thus the dynamics of learning near the critical line $\mathcal{R}_1^*$ in $\mathcal{S}^*$ is governed by higher-order terms of $\mathbf{u}$ in $l_{\mathbf{v}}$, l_w, $l_{\mathbf{u}}$, and l_z. More concretely, if we put $\tilde{\boldsymbol{\xi}} = (\mathbf{s}^*, \mathbf{v}^*, w^*, \mathbf{u}, z)$, then equations 2.37 to 2.40 change to

$$l_{\mathbf{v}}(\tilde{\boldsymbol{\xi}}) = \frac{1}{8} w^* (1 - z^2) \mathbf{Q}(\mathbf{v}^*, \mathbf{u}) + O(\mathbf{u}^3), \tag{3.2}$$

$$l_w(\tilde{\boldsymbol{\xi}}) = \frac{1}{2} \frac{1 - z^2}{w^*} \mathbf{u}^T \mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*) \mathbf{u} + O(\mathbf{u}^3), \tag{3.3}$$

$$\begin{aligned} l_{\mathbf{u}}(\tilde{\boldsymbol{\xi}}) &= (1 - z^2) \mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*) \mathbf{u} \\ &\quad + \frac{1}{24} w^* z (1 - z^2) \left\langle e(y, \mathbf{x}, \boldsymbol{\xi}) \frac{\partial D(\mathbf{x}, \mathbf{v}, \mathbf{u})}{\partial \mathbf{u}} \right\rangle \Bigg|_{\boldsymbol{\xi} = \tilde{\boldsymbol{\xi}}} + O(\mathbf{u}^3), \end{aligned} \tag{3.4}$$

$$l_z(\tilde{\boldsymbol{\xi}}) = -z \mathbf{u}^T \mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*) \mathbf{u} + O(\mathbf{u}^3), \tag{3.5}$$

where $\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)$ is as in equation 2.54. Now $l_{\mathbf{u}}(\tilde{\boldsymbol{\xi}})$ is of order $O(\mathbf{u})$, and $l_{\mathbf{v}}(\tilde{\boldsymbol{\xi}})$, $l_w(\tilde{\boldsymbol{\xi}})$, and $l_z(\tilde{\boldsymbol{\xi}})$ are all of order $O(\mathbf{u}^2)$. Neglecting higher-order terms in the above equations and focusing changes only in $\mathbf{u}$ and z, the dynamics near $\mathcal{R}_1^*$ in equations 2.44 and 2.45 can be rewritten as

$$\dot{\mathbf{u}} = 2(1 - z^2) \mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*) \mathbf{u}, \tag{3.6}$$

$$\dot{z} = -\frac{z(1 - z^2)}{w^{*2}} \mathbf{u}^T \mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*) \mathbf{u} - \frac{2z(z^2 + 1)}{w^{*2}} \mathbf{u}^T \mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*) \mathbf{u}. \tag{3.7}$$

In the above equations, the right side of $\dot{\mathbf{u}}$ corresponds to the term of $l_{\mathbf{u}}$ in equation 2.44, while the two terms on the right side of $\dot{z}$ correspond to those of l_w and l_z in equation 2.45, respectively. By putting

$$h(\mathbf{u}) = \frac{1}{2} \mathbf{u}^T \mathbf{u}, \tag{3.8}$$

we obtain the following equation of the closed form from equations 3.6 and 3.7:

$$\dot{h} = \mathbf{u}^T \dot{\mathbf{u}} = \frac{2w^{*2}(z^2-1)}{z(z^2+3)}\dot{z}. \tag{3.9}$$

This is integrable, leading us to the next theorem:

Theorem 2. *The trajectories of the averaged learning equations are given by*

$$h(\boldsymbol{u}) \stackrel{def}{=} \frac{1}{2}\boldsymbol{u}^T\boldsymbol{u} = \frac{2w^{*2}}{3}\log\frac{(z^2+3)^2}{|z|} + C \tag{3.10}$$

in the neighborhood of $\mathcal{R}_1^*$, where C is a constant depending on the initial model parameter $(\boldsymbol{u}^{(0)}, z^{(0)})$.

The two terms in the right side of equation 3.7 are both of order $O(\mathbf{u}^2)$; they are comparable to each other in general. However, when we discuss the dynamics near the elimination singularity $\mathcal{R}_2^*$, where $z^2 \approx 1$, the first term on the right side of equation 3.7 can also be neglected. That is, in the neighborhood of $\mathcal{R}_1^* \bigcap \mathcal{R}_2^*$, the l_w term in equation 2.45 can be left out too. Thus, we get

$$\dot{h} = \frac{w^{*2}(z^2-1)}{z(z^2+1)}\dot{z}, \tag{3.11}$$

and this leads to a simpler form of the learning trajectories near $\mathcal{R}_1^* \bigcap \mathcal{R}_2^*$.

Corollary 1. *The trajectories of the averaged learning equations are given by*

$$h(\boldsymbol{u}) = w^{*2} log\left(|z| + \frac{1}{|z|}\right) + C \tag{3.12}$$

in the neighborhood of $\mathcal{R}_1^* \bigcap \mathcal{R}_2^*$.

Remark 1. The trajectories determined from equations 2.41 to 2.45 are not closed in $\mathcal{S}^*$, where $\mathbf{s}$, $\mathbf{v}$, and w are determined by equation 3.10. Hence, equations 3.10 and 3.12 are the projections of the true trajectories to $\mathcal{S}^*$. However, when they deviate from this submanifold, they soon return to $\mathcal{S}^*$, provided the corresponding part of $\mathcal{R}_1^*$ is stable and attractive, because the dynamical equations of $\mathbf{s}$, $\mathbf{v}$, and w force them to the submanifold except for the small-order term of $\mathbf{u}$.

Remark 2. The trajectories in equations 3.10 and 3.12 are obtained from the averaged learning equations, so they reflect the average learning behaviors.

The actual trajectories of learning, however, are not smooth and fluctuate around equations 3.10 and 3.12. Such fluctuations are due to the fact that the training examples are given one by one in online learning and are contaminated by noise.

3.2 Dynamic Vector Fields: Redundant Case. Here we show the general results of dynamical behaviors around a singularity in hierarchical systems. In this section, we study the dynamical behavior of learning when the teacher function $f_0(\mathbf{x})$ is also included in the student networks and the number of hidden units in the student model is larger than that of the teacher. This implies that the teacher parameter might be on the singularity of $\mathbf{u} = \mathbf{0}$ or $z = \pm 1$, such that the teacher can be realized by the student model exactly. In this case, we have

$$f_0(\mathbf{x}) = g(\mathbf{x}, \mathbf{s}_0) + w_0\phi(\mathbf{x}, \mathbf{v}_0). \tag{3.13}$$

Then from equation 2.19, the teacher function is realized by any function in $\mathcal{R}^* = \mathcal{R}_1^* \bigcup \mathcal{R}_2^*$, and $\langle l(y, \mathbf{x}, \boldsymbol{\xi})\rangle = 0$ on $\mathcal{R}^*$. So the region $\mathcal{R}^*$ itself is stable, which means that the critical lines $\mathbf{u} = \mathbf{0}$ (i.e., $\mathcal{R}_1^*$) and $z = \pm 1$ (i.e., $\mathcal{R}_2^*$) are both attractive, and any trajectories approaching these lines finally stop at their intersection points with $\mathcal{R}^*$.

The proof of stability is trivial. Because the noise ε in equation 2.9 is assumed to be subject to zero mean gaussian distribution uncorrelated with the input $\mathbf{x}$, the averaged loss function $\langle l(y, \mathbf{x}, \boldsymbol{\xi})\rangle$ takes the minimum value, which is due to the random noise on the singularity $\mathcal{R}^*$ and is larger otherwise. Therefore, $\langle l(y, \mathbf{x}, \boldsymbol{\xi})\rangle$ is a Lyapunov function, and thus the singularity $\mathcal{R}^*$ is stable.

Next we discuss the dynamic vector field near the singularity $\mathcal{R}_1^*$ by fixing $\mathbf{s}$, $\mathbf{v}$, w at their optimal values. In this subspace, the trajectories are written in terms of $h(\mathbf{u})$ and z:

$$h(\mathbf{u}) = \frac{2w_0^2}{3} \log \frac{(z^2+3)^2}{|z|} + C. \tag{3.14}$$

The final trajectories of the dynamic vector field of (h, z) are shown in Figure 2. The most interesting result is that near the singularity, the trajectories converge to $\mathcal{R}^*$, depending on the initial value of (h, z), that is, $(h^{(0)}, z^{(0)})$. The two destinations are $\mathcal{R}_2^*$ ($z = \pm 1$, h is arbitrary) or $\mathcal{R}_1^*$ ($h = 0$, z is arbitrary). Let us examine these two cases in detail:

- Converging to $\mathcal{R}_1^*$ ($h = 0$). In this case, $\mathbf{u} = \mathbf{0}$, which means that the two student units a and b coincide, reaching the overlap singularity

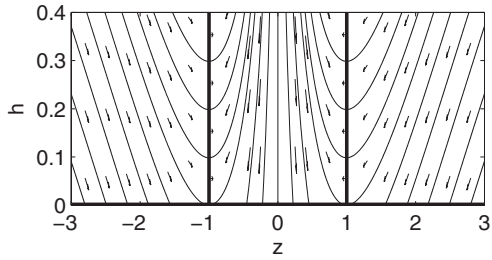


Figure 2: Dynamic vector fields in redundant case.

$\mathcal{R}_1^*$. Since z is arbitrary, units a and b are both active in the student network. They cooperate to perform exact learning of the teacher unit parameterized by $(\mathbf{v}_0, w_0)$, since $\mathbf{J}_a = \mathbf{J}_b = \mathbf{v}_0$, $w_a + w_b = w_0$. In this case, the two hidden units overlap completely.

- Converging to $\mathcal{R}_2^*$ ($z = \pm 1$). Here, we discuss only the situation of $z = +1$ (the discussion of $z = -1$ is the same). According to the coordinate transformation in section 2.4, we have $w_a = w_0$, $\mathbf{J}_a = \mathbf{v}_0$, while $w_b = 0$, $\mathbf{J}_b = \mathbf{v}_0 + \mathbf{u}$. This means that the student model has reached the elimination singularity $\mathcal{R}_2^*$, and only unit a imitates the teacher unit $(\mathbf{v}_0, w_0)$. As for unit b, because its output weight is 0, it is eliminated completely.

As can be seen from Figure 2, the basin of attraction of $\mathcal{R}_1^*$ is limited in the regions that $|z|$ is small or $|z|$ is large. So in most cases, the model parameter reaches $\mathcal{R}_2^*$. Even when the state reaches $\mathcal{R}_1^*$, it may move to $\mathcal{R}_2^*$ by random fluctuation. This confirms the results for MLPs (Cousseau et al., in press), in which the trajectories of natural gradient learning were also studied in detail.

3.3 Dynamic Vector Fields: General Case. Now we consider the general case where the teacher function $f_0(\mathbf{x})$ cannot be realized by the student model exactly. This implies that $\langle l(y, \mathbf{x}, \boldsymbol{\xi}) \rangle \neq 0$ on $\mathcal{R}_1^*$ even without additive noise. In spite of this, it is interesting to see that the trajectories of learning in the neighborhood of $\mathcal{R}_1^*$ are exactly the same as in the redundant case. However, the directions of the flow are quite complex because the stability of $\mathcal{R}_1^*$ and $\mathcal{R}_2^*$ is completely different from the previous one. On one hand, except for the intersection points with $\mathcal{R}_1^*$, there are no critical points on $\mathcal{R}_2^*$. This is obvious from equation 3.7. So even though the trajectories intersect $\mathcal{R}_2^*$, they do not stop at any intersection point unless $\mathbf{u} = \mathbf{0}$. As a result, the directions of trajectories depend on only the stability of $\mathcal{R}_1^*$, which is determined from the signature of the matrix $\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)$ in equation 2.54. There are three cases:

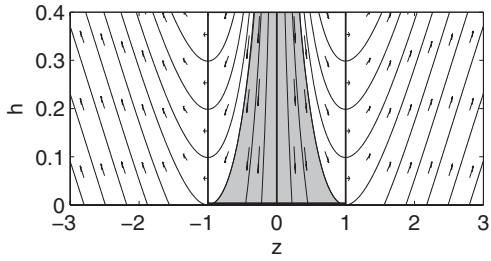


Figure 3: Stable region is $z^2 < 1$ on $\mathcal{R}_1^*$.

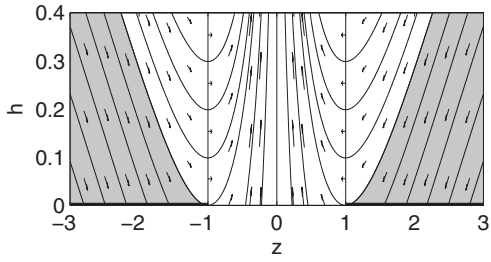


Figure 4: Stable region is $z^2 > 1$ on $\mathcal{R}_1^*$.

1. **H includes both positive and negative eigenvalues.** In this case $\mathcal{R}_1^*$ is unstable, and all trajectories leave $\mathcal{S}^*$, going quickly outside $\mathcal{S}^*$. Therefore, the trajectories in $\mathcal{S}^*$ have little meaning.
2. **H is negative-definite.** The dynamic vector field is shown in Figure 3. In this case, the segment of $z^2 < 1$ on $\mathcal{R}_1^*$ is stable (the thick black line), while the other parts of $\mathcal{R}_1^*$ are unstable. The basin of attraction is shown in the gray area. When the trajectories do not reach the attractor, they enter the region of $z^2 > 1$, where $w_a w_b < 0$, and leave the neighborhood of $\mathbf{u} = \mathbf{0}$.
3. **H is positive-definite.** The dynamic vector field is shown in Figure 4. In this case, the two segments of $z^2 > 1$ on $\mathcal{R}_1^*$ are stable (the two thick black line segments), while the part of $z^2 < 1$ is unstable. The gray basin of attraction in Figure 4 shows that unless the trajectories reach the stable part, they first approach the region $z^2 < 1$, where $w_a w_b > 0$, and then leave the singularity.

In the dynamic vector fields of Figures 2, 3, and 4, whether a trajectory intersects the critical line $\mathbf{u} = \mathbf{0}$, depends on the constant C in equation 3.10 or 3.12. When

$$C = C_0 = -\frac{8w^{*2}}{3} \log 2, \tag{3.15}$$

the trajectory is tangential to $\mathcal{R}_1^*$. Otherwise, if $C < C_0$, the trajectory has two intersection points with $\mathcal{R}_1^*$; if $C > C_0$, the trajectory is totally above $\mathcal{R}_1^*$ and certainly passes the elimination singularity $\mathcal{R}_2^*$, where $w_a w_b = 0$. In this case, C represents the distance between the trajectory and overlap singularity $\mathbf{u} = \mathbf{0}$. Note that if we use equation 3.12, then $C_0 = -w^{*2} \log 2$.

An interesting result from Figures 3 and 4 is that near the singularity $\mathcal{R}_1^*$, the destination of a flow also depends on the initial model parameter $(h^{(0)}, z^{(0)})$, which is equivalent to $(\mathbf{u}^{(0)}, z^{(0)})$. When $(h^{(0)}, z^{(0)})$ is located in the stable region (gray area), $(\mathbf{u}, \mathbf{z})$ converges to $\mathcal{R}_1^*$, with its destination being either $z^2 > 1$ or $z^2 < 1$; otherwise, $(\mathbf{u}, \mathbf{z})$ passes the elimination singularity $\mathcal{R}_2^*$, changes the sign of w_a or w_b, and then goes to other singularities or local minima.

Figures 3 and 4 are the averaged dynamic vector fields with small $\|\mathbf{u}\|$. The dynamic vector fields, however, are different when $\|\mathbf{u}\|$ is large. Such trajectories for RBF network learning will be discussed in a separate article (Wei & Amari, in press).

4 Plateau Phenomena Near the Singularity

The plateau phenomenon is ubiquitous in the learning process of various hierarchical neural models (Amari et al., 2006). It has been believed that permutation symmetry gives rise to plateaus. Here, we show that there are two kinds of plateaus: the on-singularity plateau and the near-singularity plateau.

4.1 On-Singularity Plateau. According to the previous discussion, when $\mathcal{R}_1^*$ is partially stable and the initial state of the model parameter belongs to its basin of attraction, the averaged dynamics of the model parameter is attracted to its stable part and stays on it (see Figures 3 and 4). However, since our learning paradigm is online, training examples are received one by one, and the model parameters are adapted each time when a new example is given. Therefore, the actual change in parameters includes fluctuations due to the random sampling of $\mathbf{x}$ and additive noise. Such fluctuations exist after the model parameters reach the stable part of $\mathcal{R}_1^*$, even when the teacher signals do not include noise. When the model parameter leaves $\mathcal{R}_1^*$, it returns to the critical line because of the partial stableness of the line. However, since all the points on $\mathcal{R}_1^*$ correspond to the same model function, which does not depend on z, such fluctuations result in a random walk process of the model parameter on $\mathcal{R}_1^*$. That is, each time after the model parameter is adjusted, it moves randomly to a different point on the stable part of $\mathcal{R}_1^*$.

Unfortunately, $\mathcal{R}_1^*$ is only partially stable. This means that once the parameter reaches the unstable part of $\mathcal{R}_1^*$ by passing through $\mathcal{R}_2^*$, that is, $|z| = 1$, it begins to move away from the singularity since $\mathcal{R}_1^*$ becomes repulsive. As a result, an arbitrary small noise in the teacher signal may kick

the parameters away from the singularity, which means that the singularity $\mathcal{R}_1^*$ is attractive but eventually unstable. Although it has a positive measure of basin of attraction, the state finally leaves it from its unstable part. In this sense the singularity is completely different from the saddle point and is a Milnor-like attractor (Milnor, 1985).

Unlike the ordinary saddle point, the basin of attraction of the singularity has a positive measure. Hence, starting from the basin of attraction, the state is first attracted to the stable interval of $\mathcal{R}_1^*$; then it undergoes a long period of random walk on $\mathcal{R}_1^*$; finally it leaves $\mathcal{R}_1^*$. During the period of random walk, because the model functions on $\mathcal{R}_1^*$ are the same, the averaged loss function (i.e., the generalization error) also remains almost unchanged. This is the mechanism of the on-singularity plateau phenomenon in online learning.

The random walk on the critical line is one-dimensional. According to the random walk theory (see, e.g., Feller, 1971), for any random walk in one dimension, every random trajectory starting at a point on the line will cross the boundary almost surely. As a result, once the random walk begins, the points of $z = +1$ or $z = -1$ on $\mathcal{R}_1^*$ will be crossed almost surely. However, it might take a long time for the system to escape the singularity.

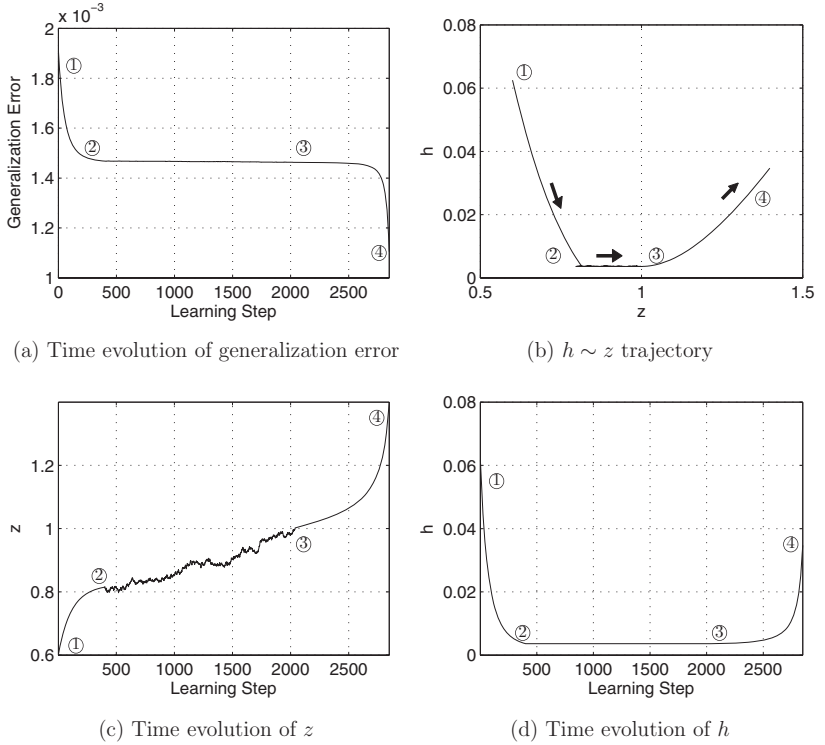
Next we give some experimental results. Here we give just a brief introduction. Detailed discussion about the dynamics of learning in RBF networks will be shown in Wei and Amari (in press). For RBF networks and MLPs that use a modified error function in equation 2.4, if we assume that the teacher function is also generated by the student network and the training input $\mathbf{x}$ is subject to gaussian distribution with zero mean and unit variance, then both the averaged learning equation $\dot{\boldsymbol{\theta}}(t)$ in equation 2.12 and the matrix $\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)$ in equation 2.54 can be integrated explicitly. The equation $\dot{\boldsymbol{\theta}}(t)$ can be solved numerically to obtain the time evolution of model parameter $\boldsymbol{\theta}$. Then we can obtain the best approximation parameter $(\mathbf{s}^*, \mathbf{v}^*, w^*)$ of the student model with $k-1$ hidden units by simply letting the initial parameters of two of the hidden units be identical when we solve $\dot{\boldsymbol{\theta}}(t)$. If we have obtained $(\mathbf{s}^*, \mathbf{v}^*, w^*)$ and $\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)$, then $\dot{\mathbf{u}}$ and $\dot{z}$ in equations 3.6 and 3.7 can also be solved numerically to get their time evolutions within the subspace $\mathcal{S}^*$. It is straightforward to plot the time evolution of h and the corresponding $h \sim z$ trajectory from the time evolutions of $\mathbf{u}$ and z.

To observe the plateau phenomenon, it is also very important to know the time evolution of the generalization error, which is measured by the averaged loss function,

$$E(\boldsymbol{\xi}) = \langle l(y, \mathbf{x}, \boldsymbol{\xi}) \rangle = \frac{1}{2} \langle (f_0(\mathbf{x}) - f(\mathbf{x}, \boldsymbol{\xi}))^2 \rangle. \tag{4.1}$$

Near the overlap singularity $\mathcal{R}_1^*$, $E(\boldsymbol{\xi})$ can be simplified by using equations 3.1 and 2.29,

$$E(\boldsymbol{\xi}) = E_0(\mathbf{s}^*, \mathbf{v}^*, w^*) - \frac{1}{2}(1 - z^2)\mathbf{u}^T \mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)\mathbf{u} + O(\mathbf{u}^4), \tag{4.2}$$



(a) Time evolution of generalization error

(b) $h \sim z$ trajectory

(c) Time evolution of z

(d) Time evolution of h

Figure 5: On-singularity plateau observed by the numeric method.

where $\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)$ is as in equation 2.54 and

$$E_0(\mathbf{s}^*, \mathbf{v}^*, w^*) = \frac{1}{2}\langle (f_0(\mathbf{x}) - g(\mathbf{x}, \mathbf{s}^*) - w^*\phi(\mathbf{x}, \mathbf{v}^*))^2 \rangle. \tag{4.3}$$

Since the time evolutions of $\mathbf{u}$ and z have been obtained numerically, we can also plot the evolution of generalization error according to equation 4.2.

Figure 5 shows the on-singularity plateau observed in RBF networks by using the above numeric method. We use two hidden units in both the teacher and student networks, but the results are the same for large networks (Wei & Amari, in press). The teacher network parameters are $\boldsymbol{\mu}_1^{(t)} = [0.2, 0.3]$, $w_1^{(t)} = -0.3$, $\boldsymbol{\mu}_2^{(t)} = [-0.2, -0.3]$, $w_2^{(t)} = 0.9$, and $\sigma_1^{(t)} = \sigma_2^{(t)} = 0.5$. Using the numeric method, we can find that one of the best approximations is $\mathbf{v}^* = [-0.2837, -0.4255]$, $w^* = 0.7510$. Then from the definiteness of $\mathbf{H}(\mathbf{v}^*, w^*)$, we can check that the $z^2 < 1$ part of $\mathcal{R}_1^*$ is attractive. Given $\mathbf{v}^*$, w^*, and an initial state, now we can solve $\dot{\mathbf{u}}$ and

$\dot{z}$ in equations 3.6 and 3.7 numerically. We choose the following initial student parameters: $\boldsymbol{\mu}_1^{(0)} = [-0.50, -0.85]$, $w_1^{(0)} = 0.4$, $\boldsymbol{\mu}_2^{(0)} = [-0.25, -0.60]$, $w_2^{(0)} = 0.1$. The widths of the two student units are fixed at $\sigma = 0.5$.

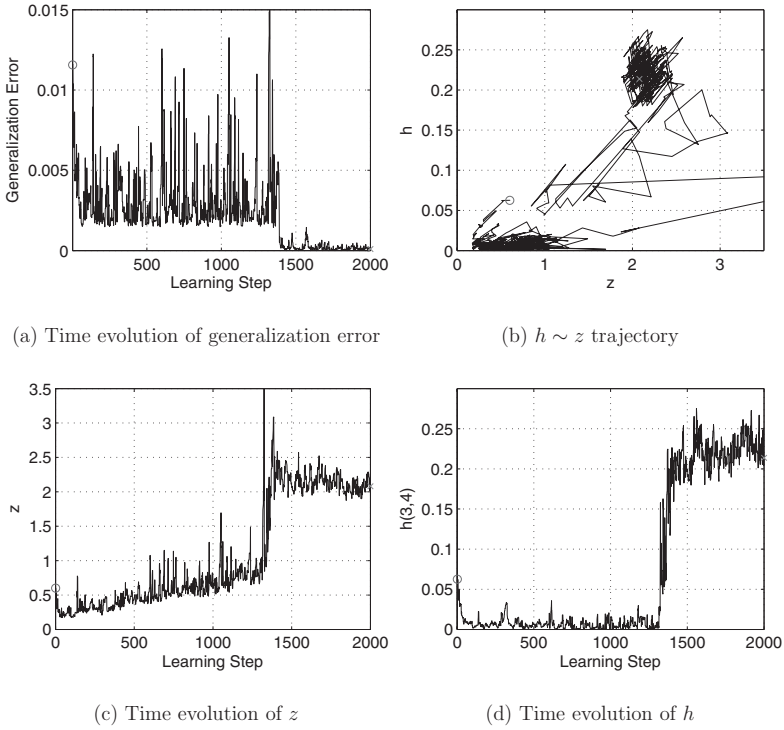
With the above configuration, the model parameter converges to a point on the stable part of $\mathcal{R}_1^*$. Because equations 3.6 and 3.7 reflect only the averaged dynamics of online learning, we use a stationary gaussian stochastic process with zero mean and small covariance instead of the subsequent random walk. Figure 5 shows the results of the test. Figure 5a is the time evolution of the generalization error, Figure 5b is the corresponding $h \sim z$ trajectory of learning, and Figures 5c and 5d are the time evolutions of z and h. In all figures, segment 1 to 2 represents the period of model parameter converging to the singularity, segment 2 to 3 is the period of random walk, and segment 3 to 4 is the period of model parameter leaving the singularity. Note that in real online learning, segments 1 to 2 and 3 to 4 might include fluctuations around them. We will show this later. It is clear that during the period of random walk, the on-singularity plateau occurs.

In Figure 5b the trajectories 1 to 2 and 3 to 4 correspond to those moving to and leaving from the overlap singularity in Figure 3. However, the moving speeds of h and z during these two periods might be different. According to equations 3.6 and 3.7, if $(\mathbf{u}, z)$ is close to the elimination singularity $\mathcal{R}_2^*$ where $1 - z^2 \approx 0$, then z changes faster than $\mathbf{u}$ because $\dot{\mathbf{u}} \approx 0$; if $(\mathbf{u}, z)$ is far away from $\mathcal{R}_2^*$, then $\mathbf{u}$ changes much faster than z because $\dot{\mathbf{u}}$ is of order $O(\mathbf{u})$ and $\dot{z}$ is of order $O(\mathbf{u}^2)$.

The above discussion also indicates that during the period of z drifting toward the elimination singularity $\mathcal{R}_2^*$, h may have a small but nonzero value. The closer z is to $\mathcal{R}_2^*$, the larger h may be. This phenomenon is shown by the nonzero h value in Figures 5b and 5d.

Trajectories in Figure 5 are plotted based on equations 3.6, 3.7, and 4.2. They are the projections of the true trajectories to $\mathcal{S}^*$, by fixing $\mathbf{s}$, $\mathbf{v}$, and w at their optimal values $\mathbf{s}^*$, $\mathbf{v}^*$, and w^*. Next we compare them with the real trajectories in online learning by simulation. The teacher parameters and the initial student model parameters in simulation are exactly the same as in the numeric test in Figure 5. Training examples $(y_t, \mathbf{x}_t)$ are obtained one by one, with additive noise subject to gaussian distribution with zero mean and variance 0.02. Each time we get an example, the model parameter is trained according to equation 2.11 with a learning rate $\eta = 0.05$. Training on the current example is stopped until the algorithm reaches a maximum times of 200, or $|y_t - f(\mathbf{x}_t, \boldsymbol{\theta}_t)| < 10^{-3}$. Then the algorithm waits for the next example.

Figure 6 shows the real online learning dynamics. In all figures $\circ$ and $\times$ represent the initial and final states, respectively. We can see that the trajectories in Figure 6 are very similar to those in Figure 5, except that the online trajectories include fluctuations due to the additive noise and the random sampling of $\mathbf{x}_t$. It is quite clear that the random walk on $\mathcal{R}_1^*$ results in the on-singularity plateau.



(a) Time evolution of generalization error

(b) $h \sim z$ trajectory

(c) Time evolution of z

(d) Time evolution of h

Figure 6: On-singularity plateau in real online learning.

During the period of random walk, although the generalization error keeps almost unchanged, the eigenvalues of the Hessian $\langle \frac{\partial^2 l(y,\mathbf{x},\boldsymbol{\xi})}{\partial \mathbf{u} \partial \mathbf{u}^T} \rangle|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*}$ on $\mathcal{R}_1^*$ are always fluctuating along with the drift of z. This is an important feature of the random walk, which can be seen clearly from equation 2.53. When z walks to the unstable part of $\mathcal{R}_1^*$, the eigenvalues of the Hessian flip their signs. As a result, the system becomes unstable because of the positive eigenvalues, and the system escapes the singularity.

It should also be pointed out that in online learning, the overlap singularities look like local minima, although they are not. A local minimum is always stable, but an overlap singularity is unstable. If we have enough patience, the algorithm will finally escape it. Note that the fluctuations due to the noise and the random sampling of input play an important role in such an on-singularity plateau.

4.2 Near-Singularity Plateau. Even when the initial state of the model parameter does not belong to the basin of attraction of $\mathcal{R}_1^*$, the plateau-like phenomenon appears. This is called the near-singularity plateau, which

occurs when trajectories are close to the overlap singularity $\mathcal{R}_1^*$. Such trajectories are called near-singularity trajectories. Although they have no intersection with $\mathcal{R}_1^*$ (totally above $\mathcal{R}_1^*$), they cross the elimination singularity $\mathcal{R}_2^*$.

Now we elucidate the mechanism of the near-singularity plateau. Near the overlap singularity, the rate of change of $E(\boldsymbol{\xi})$ in equation 4.2 is

$$\dot{E}(\boldsymbol{\xi}) = z\mathbf{u}^T\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)\mathbf{u}\dot{z} - (1 - z^2)\mathbf{u}^T\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)\dot{\mathbf{u}} + O(\mathbf{u}^4). \quad (4.4)$$

We know that $\dot{z}$ is of order $O(\mathbf{u}^2)$ and $\dot{\mathbf{u}}$ is of order $O(\mathbf{u})$. So in equation 4.4, the first term of $\dot{E}(\boldsymbol{\xi})$ is of order $O(\mathbf{u}^4)$, and the second term is at most of order $O(\mathbf{u}^2)$. When the model parameter is very close to the elimination singularity $\mathcal{R}_2^*$, where $z \approx \pm 1$, then the first term dominates in $\dot{E}(\boldsymbol{\xi})$, and $\dot{E}(\boldsymbol{\xi})$ is of order $O(\mathbf{u}^4)$. This tells us that if $\mathbf{u}$ is small, $E(\boldsymbol{\xi})$ remains almost unchanged when the model parameter crosses the lines $z = \pm 1$. As a result, a plateau-like phenomenon appears near the elimination singularity.

We can see this more clearly by plotting the error evolution when the model parameter passes a near-singularity trajectory. Since $\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)$ is a real symmetric matrix, its Rayleigh quotient $\kappa(\mathbf{u})$, defined as

$$\kappa(\mathbf{u}) = \frac{\mathbf{u}^T\mathbf{H}(\mathbf{s}^*, \mathbf{v}^*, w^*)\mathbf{u}}{\mathbf{u}^T\mathbf{u}}, \quad (4.5)$$

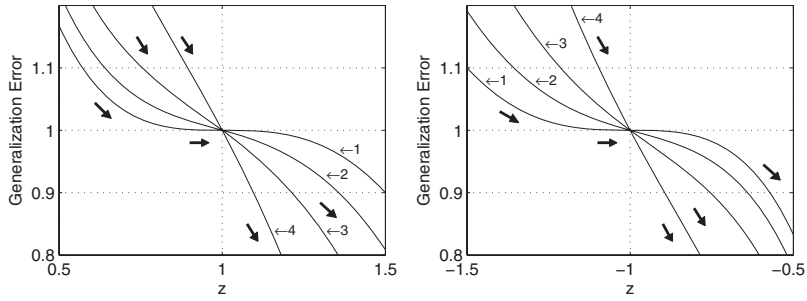
satisfies

$$\lambda_{\min} \leq \kappa(\mathbf{u}) \leq \lambda_{\max}, \quad (4.6)$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues of $\mathbf{H}$. Since the dynamics is discussed around $\mathcal{R}_1^* \bigcap \mathcal{R}_2^*$, we can use equation 3.12 and get

$$E(\boldsymbol{\xi}) = E_0(\mathbf{s}^*, \mathbf{v}^*, w^*) - \kappa(\mathbf{u})(1 - z^2)\left(w^{*2}\log\left(|z| + \frac{1}{|z|}\right) + C\right). \quad (4.7)$$

Note that if $\mathbf{H} < \mathbf{0}$ ($\mathbf{H}$ is negative-definite), then $\kappa(\mathbf{u}) < 0$, and the directions of near-singularity trajectories (see Figure 3) are from the $z^2 < 1$ region to the $z^2 > 1$ region; if $\mathbf{H} > \mathbf{0}$ ($\mathbf{H}$ is positive-definite), then $\kappa(\mathbf{u}) > 0$, and the directions of near-singularity trajectories (see Figure 4) are from the $z^2 > 1$ region to the $z^2 < 1$ region. According to equation 4.7, Figures 7a and 7b plot the error evolutions ($E(\boldsymbol{\xi})$ versus z) when the model parameter moves along near-singularity trajectories in Figures 3 and 4. In Figure 7, $\mathbf{u}$ is assumed to be one-dimensional, and the constant $E_0(\mathbf{s}^*, \mathbf{v}^*, w^*)$ is assumed to be 1. It is obvious that there is a plateau near $z = \pm 1$ when $C = C_0$ (curve



(a) $\mathbf{H} < \mathbf{0}$ (b) $\mathbf{H} > \mathbf{0}$

Figure 7: $E(\boldsymbol{\xi}) \sim z$ curves near $z = \pm 1$.

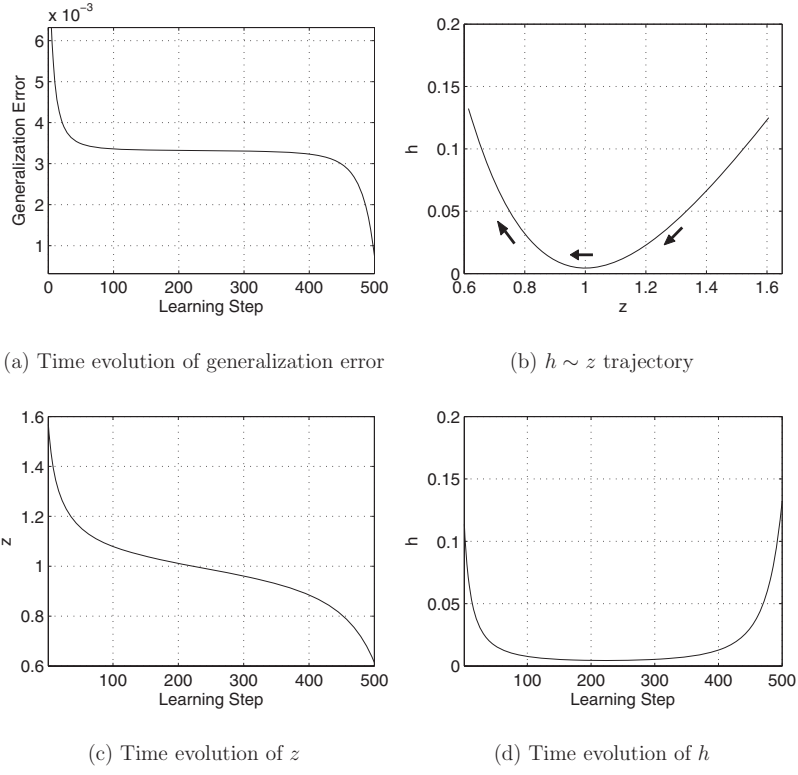
1). But if we increase C, the plateau vanishes gradually (curves 2, 3, and 4). Note that such near-singularity trajectories exist only when $C > C_0$. If $C < C_0$, then the trajectories intersect the singularity line $\mathbf{u} = \mathbf{0}$, and the on-singularity plateau occurs.

Figure 7 indicates that trajectories above the line $\mathcal{R}_1^*$ give rise to plateaus when they pass the elimination singularity. The closer the trajectory is to $\mathcal{R}_1^*$, the more serious the plateau is. In this sense, the so-called quasi-plateaus (Park et al., 2005) are also easily observed (see curves 2 and 3 in Figure 7).

Figure 8 shows the near-singularity plateau observed in RBF networks when the model parameter crosses the elimination singularity $\mathcal{R}_2^*$. The numeric method of obtaining the figure is the same as depicted in section 4.1. However, the teacher and the initial student network parameters are different. The teacher network parameters are $\boldsymbol{\mu}_1^{(t)} = [0.4, 0.3]$, $w_1^{(t)} = 0.4$, $\boldsymbol{\mu}_2^{(t)} = [-0.4, -0.3]$, $w_2^{(t)} = 0.9$, and $\sigma_1^{(t)} = \sigma_2^{(t)} = 0.5$. Using the numeric method, we can solve that the best approximation is $\mathbf{v}^* = [-0.2323, -0.1742]$, $w^* = 1.0555$, and the $z^2 > 1$ part of $\mathcal{R}_1^*$ is attractive. We choose the following initial student parameters so that the averaged trajectory does not reach $\mathcal{R}_1^*$ but passes through the elimination singularity $z = 1$: $\boldsymbol{\mu}_1^{(0)} = [0, -0.3]$, $w_1^{(0)} = 1.3$, $\boldsymbol{\mu}_2^{(0)} = [0.4, 0]$, $w_2^{(0)} = -0.3$. The widths of the two student units are also fixed at $\sigma = 0.5$.

Figure 8a shows the time evolutions of generalization error, Figure 8b shows the corresponding near-singularity trajectories, and Figures 8c and 8d show the time evolutions of z and h. We can see that when the model parameter passes the line $z = 1$, where $w_b = 0$, a plateau-like phenomenon occurs, although $h \neq 0$.

Figure 9 shows the near-singularity plateau observed in real online learning simulation. In simulation, the teacher network parameters and the initial student parameters are also the same as the numeric test in Figure 8. Note that in Figure 9, the slow learning speed around $z = 1$ is shown as a period



(a) Time evolution of generalization error

(b) $h \sim z$ trajectory

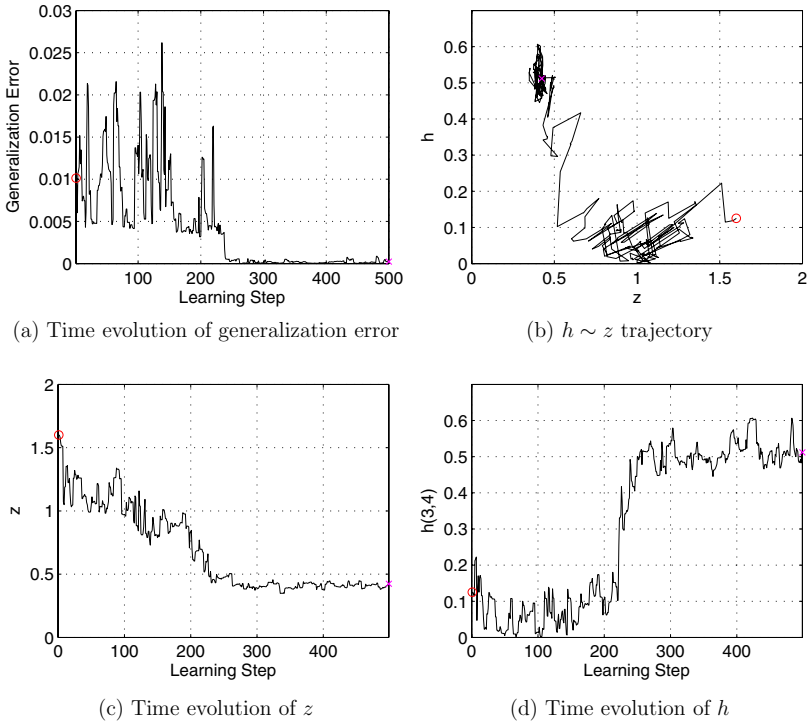
(c) Time evolution of z

(d) Time evolution of h

Figure 8: Near-singularity plateau observed by the numeric method.

of random walk in the neighborhood of $z = 1$. It is during this period that the near-singularity plateau happens.

From the dynamic vector fields in Figures 3 and 4, which are projections of the true learning trajectories on the $\mathcal{S}^*$ plane (see Figure 1), we can imagine the shape of the error surface of the averaged cost function around the overlap singularity $\mathbf{u} = \mathbf{0}$. Here let us consider only the $z > 0$ part of the overlap singularity whose stable part is $z > 1$ (discussion of other cases are similar). $\mathbf{u} = \mathbf{0}$ and $z = 1$ are both contour lines with the same $E(\boldsymbol{\xi})$. Around the line of $\mathbf{u} = \mathbf{0}$, the error surface in the neighborhood of $z > 1$ is above the line, while the error surface in the neighborhood of $0 < z < 1$ is under the line. Moreover, this error surface becomes flat around $z = 1$. The closer the z is to $z = 1$, the flatter the surface is. Obviously the learning trajectories around $\mathbf{u} = \mathbf{0}$ depend on the initial states. If the model parameter is initialized in the neighborhood where $z \gg 1$, then it first moves toward $\mathbf{u} = \mathbf{0}$, then drifts to $z = 1$ along the critical line, with some fluctuations (so $h \neq 0$). This corresponds to the on-singularity plateau. But if the initial parameter is close to the intersection point of $\mathbf{u} = \mathbf{0}$ and $z = 1$, then the model parameter



(a) Time evolution of generalization error

(b) $h \sim z$ trajectory

(c) Time evolution of z

(d) Time evolution of h

Figure 9: Near-singularity plateau in real online learning.

just passes the line of $z = 1$ and leaves the overlap singularity. Since the error surface in this area is flat, a near-singularity plateau phenomenon occurs, and $h \neq 0$ also holds since the near-singularity trajectory never intersects the line of $\mathbf{u} = \mathbf{0}$. So from the point of view of error surface, these two kinds of plateaus have different mechanisms.

In the case of the on-singularity plateau, $\dot{E}(\boldsymbol{\xi}) = 0$ on average when the model parameter walks randomly on $\mathcal{R}_1^*$. However, during the period of the near-singularity plateau, $\dot{E}(\boldsymbol{\xi}) < 0$ always holds, although its value is very small. This is an important difference between the two kinds of plateaus, which implies that it is relatively easy for the learning algorithm to leave the near-singularity plateau. This also explains why the time period of near-singularity plateau in Figures 8 and 9 is much shorter than that of the on-singularity plateau in Figures 5 and 6.

5 Conclusion

This letter investigates the dynamics of learning near singularities in layered networks by discussing the stability, the trajectories of learning, and

the plateau phenomena in a unified framework. We show that various hierarchical models share common trajectories of learning near an overlap singularity. They are represented by a very simple form, $\frac{1}{2}\mathbf{u}^T\mathbf{u} = \frac{2w^{*2}}{3}\log\frac{(z^2+3)^2}{|z|} + C$, provided that irrelevant parameters $(\mathbf{s}^*, \mathbf{v}^*, w^*)$ are the best approximation to the teacher by the model with $k-1$ units. The result applies to many other hierarchical models trained by the gradient descent method as long as their activation function is smooth enough so that Taylor expansion can be used (e.g., the reformulated RBF networks by Karayiannis & Randolph-Gips, 2003).

The universal mechanism of plateau has been elucidated. The dynamical analysis shows that the overlap singularity gives rise to a Milnor-like attractor driven by noise. If a trajectory converges to the stable part of the overlap singularity, then the subsequent random walk gives rise to the on-singularity plateau. Even when the trajectory does not converge to the overlap singularity, near-singularity plateaus appear.

This letter mainly discussed the learning dynamics near the overlap singularity and the elimination singularity close to them. The dynamics of learning near the elimination singularities far away from the overlap singularities still remains unknown. Determining the influences of such singularities on the dynamics of learning is one of our topics for further research.

In addition to giving rise to plateaus, the overlap singularity may also cause other problems. For example, a lot of Milnor-like attractors are caused by them. Considering that so many overlap singularities exist because of the permutation symmetry, we can imagine that there are also many black hole–like attractors in the parameter space. Once the model parameter is attracted to such a singularity, it is very hard for it to get away. This explains why there are so many local minima in backpropagation learning for hierarchical models, including MLPs and RBF networks. How to escape such singularities in learning is another interesting topic. The natural gradient method (Amari, 1998; Park, Amari, & Fukumizu, 2000) gives one solution to it. It is interesting to investigate it further in the present context (see Cousseau et al., in press).

Appendix: Learning Equations in the New Coordinate System

According to the coordinate transformation, the relation between the learning equations in the new and original coordinate systems is represented by

$$\dot{\boldsymbol{\xi}} = \mathbf{T}\dot{\boldsymbol{\theta}}, \tag{A.1}$$

where $\mathbf{T}$ is the Jacobian matrix:

$$\mathbf{T} = \frac{d\boldsymbol{\xi}}{d\boldsymbol{\theta}^T}. \tag{A.2}$$

On the other hand, the gradients of f in terms of the two coordinates are related as

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{T}^T \frac{\partial f(\mathbf{x}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}}. \tag{A.3}$$

Arranging the above equations, we obtain

$$\dot{\boldsymbol{\xi}} = -\eta \mathbf{T}\mathbf{T}^T \left\langle \frac{\partial l(y, \mathbf{x}, \boldsymbol{\xi})}{\partial \boldsymbol{\xi}} \right\rangle \tag{A.4}$$

Now it is easy to obtain the learning equations 2.41 to 2.45 in the new coordinate system.

References

Amari, S. (1967). Theory of adaptive pattern classifiers. *IEEE Trans. Electronic Computers, EC-16*(3), 299–307.

Amari, S. (1977). Neural theory of association and concept-formation. *Biological Cybernetics, 26*, 175–185.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation, 10*, 251–276.

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry*. New York: Oxford University Press.

Amari, S., & Nakahara, H. (2005). Difficulty of singularity in population coding. *Neural Computation, 17*, 839–858.

Amari, S., & Ozeki, T. (2001). Differential and algebraic geometry of multilayer perceptrons. *IEICE Trans., E84-A*, 31–38.

Amari, S., Park, H., & Ozeki, T. (2002). Geometrical singularities in the neuromanifold of multilayer perceptrons. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14* (pp. 343–350). Cambridge, MA: MIT Press.

Amari, S., Park, H., & Ozeki, T. (2006). Singularities affect dynamics of learning in neuromanifolds. *Neural Computation, 18*, 1007–1065.

Biehl, M., & Caticha, N. (2002). Statistical mechanics of on line learning and generalization. In M. A. Arbib (Ed.), *Handbook of brain theory and neural networks*. Cambridge, MA: MIT Press.

Biehl, M., Riegler, P., & Wöhler, C. (1996). Transient dynamics of online learning in two-layered neural networks. *J. Phys. A: Math. Gen., 29*, 4769–4780.

Biehl, M., & Schwarze, H. (1995). Learning by online gradient descent. *J. Phys. A: Math. Gen., 28*, 643–656.

Chen, A., Lu, H., & Hecht-Nielsen, R. (1993). On the geometry of feedforward neural network error surfaces. *Neural Computation, 5*, 910–927.

Cousseau, F., Ozeki, T., & Amari, S. (in press). Dynamics of learning in multilayer perceptrons near the singularity. *IEEE Trans. Neural Networks*.

Dacunha-Castelle, D., & Gassiat, E. (1997). Testing in locally conic models, and application to mixture models. *Probability and Statistics, 1*, 285–317.

Feller, W. (1971). *An introduction to probability theory and its applications* (Vol. 2), New York: Wiley.

Freeman, J., & Saad, D. (1997a). Online learning in radial basis function networks. *Neural Computation, 9*, 1601–1622.

Freeman, J., & Saad, D. (1997b). Dynamics of online learning in radial basis function networks. *Physical Review E, 56*, 907–916.

Fukumizu, K. (1996). A regularity condition of information matrix of a multilayer perceptron network. *Neural Networks, 9*, 871–879.

Fukumizu, K. (2003). Likelihood ratio of unidentifiable models and multilayer neural networks. *Annals of Statistics, 31*, 833–851.

Fukumizu, K., & Amari, S. (2000). Local minima and plateaus in hierarchical structure of multilayer perceptrons. *Neural Networks, 13*, 317–327.

Hagiwara, K. (2002). On the problem in model selection of neural network regression in overrealizable scenario. *Neural Computation, 14*, 1979–2002.

Hartigan, J. (1985). A failure of likelihood asymptotics for normal mixtures. In L. M. Lecam & R. A. Olshen (Eds.), *Proc. Berkeley Conf. in Honor of J. Neyman and J. Kiefer* (Vol. 2, pp. 807–810). Belmont, CA: Wadsworth.

Heskes, T., & Kappen, B. (1991). Learning process in neural networks. *Physical Review, A44*, 2718–2762.

Huh, N., Oh, J., & Kang, K. (2000). Online learning of a mixture-of-experts neural network. *J. Phys. A: Math. Gen., 33*, 8663–8672.

Inoue, M., Park, H., & Okada, M. (2003). online learning theory of soft committee machines with correlated hidden units: Steepest gradient descent and natural gradient descent. *J. Phys. Soc. Japan, 72*, 805–810.

Inoue, M., Park, H., & Okada, M. (2004). Dynamics of the adaptive natural gradient descent method for soft committee machines. *Physical Review E, 69*(5), 056120.

Karayiannis, N., & Randolph-Gips, M. (2003). On the construction and training of reformulated radial basis function neural networks. *IEEE Trans. Neural Networks, 14*, 835–846.

Kurková, V., & Kainen, P. (1994). Functionally equivalent feedforward neural networks. *Neural Computation, 6*, 543–558.

Milnor, J. (1985). On the concept of attractor. *Comm. in Math. Phys. 99*, 177–195.

Park, H., Amari, S., & Fukumizu, K. (2000). Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks, 13*, 755–764.

Park, H., Inoue, M., & Okada, M. (2003). Online learning dynamics of multilayer perceptrons with unidentifiable parameters. *J. Phys. A: Math. Gen., 36*, 11753–11764.

Park, H., Inoue, M., & Okada, M. (2005). Slow dynamics due to singularities of hierarchical learning machines. *Progress of Theoretical Physics Supplement, 157*, 275–279.

Riegler, P., & Biehl, M. (1995). Online backpropagation in two-layered neural networks. *J. Phys. A: Math.Gen., 28*, L507–L513.

Saad, D., & Solla, A. (1995a). On-line learning in soft committee machines. *Physical Review E, 52*, 4225–4243.

Saad, D., & Solla, A. (1995b). Exact solution for online learning in multilayer neural networks. *Physical Review Letters, 74*, 4337–4340.

Sussmann, H. (1992). Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks, 5*, 589–593.

Watanabe, S. (2001a). Algebraic analysis for non-identifiable learning machines. *Neural Computation, 13*, 899–933.

Watanabe, S. (2001b). Algebraic geometrical methods for hierarchical learning machines. *Neural Networks, 14*, 1409–1060.

Watanabe, S., & Amari, S. (2003). Learning coefficients of layered models when the true distribution mismatches the singularities. *Neural Computation, 15*, 1013–1033.

Wei, H., & Amari, S. (2006). Online learning dynamics of radial basis function neural networks near the singularity. In *Proc. IJCNN2006* (pp. 4770–4776). Piscataway, NJ: IEEE Press.

Wei, H., & Amari, S. (in press). Dynamics of learning near singularities in radial basis function networks. *Neural Networks*.

Received December 11, 2006; accepted May 24, 2007.