Integral representation of shallow neural network that attains the global minimum

Sho Sonoda, Isao Ishikawa, Masahiro Ikeda, Kei Hagihara
RIKEN AIP, Tokyo, Japan
{sho.sonoda, isao.ishikawa, masahiro.ikeda, kei.hagihara}@riken.jp

Yoshihiro Sawano
Tokyo Metropolitan University, Tokyo, Japan
ysawano@tmu.ac.jp

Takuo Matsubara, Noboru Murata
Waseda University, Tokyo, Japan
takuo.matsubara@suou.waseda.jp
noboru.murata@eb.waseda.ac.jp

October 4, 2018

Abstract

We consider the supervised learning problem with shallow neural networks. According to our unpublished experiments conducted several years prior to this study, we had noticed an interesting similarity between the distribution of hidden parameters after backpropagation (BP) training, and the ridgelet spectrum of the same dataset. Therefore, we conjectured that the distribution is expressed as a version of ridgelet transform, but it was not proven until this study. One difficulty is that both the local minimizers and the ridgelet transforms have an infinite number of varieties, and no relations are known between them. By using the integral representation, we reformulate the BP training as a strong-convex optimization problem and find a global minimizer. Finally, by developing ridgelet analysis on a reproducing kernel Hilbert space (RKHS), we write the minimizer explicitly and succeed to prove the conjecture. The modified ridgelet transform has an explicit expression that can be computed by numerical integration, which suggests that we can obtain the global minimizer of BP, without BP.

1 INTRODUCTION

Training a neural network is conducted by backpropagation (BP), which results in a high-dimensional and non-convex optimization problem. Despite the complexity of the loss surface, deep learning has achieved great success in a wide range of applications such as image recognition (Redmon et al., 2016), speech synthesis (van den Oord et al., 2016), and game playing (Silver et al., 2017). Investigations on the geometry of loss surfaces, the distribution of local minima, and the behavior of stochastic gradient descent are recent trends in theoretical research on deep learning. (See § 1.5
Figure 1: Our motivating results: (a) scatter plot of BP trained parameters \( \{(a_i, b_i)\}_{i=1}^{np} \), and (b) spectrum of the ridgelet transform \( R[f](a, b) \).

For example.) In this paper, by reformulating the BP training of the shallow neural network in the integral representation, we prove that there uniquely exists a global minimizer, and obtain its explicit expression, which can be computed by numerical integration.

Figure 1 motivated us to study this problem. Both pictures are obtained from the same dataset \( \{(x_i, y_i)\}_{i=1}^{200} \), where \( x_i \in [-1, 1] \) and \( y_i = \sin 2\pi x_i \). Figure 1a shows the BP trained hidden parameters \( \{(a_j, b_j)\}_{j=1}^{np} \) optimized by LBFGS, obtained from \( n = 1,000 \) networks with \( p = 20 \) ReLU units. Figure 1b shows a numerically integrated spectrum of the ridgelet transform \( R[f](a, b) \) of \( f(x) = \sin 2\pi x, (x \in [-1, 1]) \). (See §2 for the definition of \( R \).) Even though two figures are obtained from different procedures—numerical optimization and numerical integration—both results are 12-point star shaped. Namely, BP trained parameters \( (a_j, b_j) \) concentrate in the areas where the intensity of the ridgelet spectrum is high. Therefore, we can expect that the global minimizer can be calculated as a variant of ridgelet transforms. See §2 for more details on the experimental setup.

1.1 Integral Representation

The integral representation of a shallow neural network is defined by

\[
S[\gamma](x) := \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(a, b)\sigma(a \cdot x - b)d\lambda(a, b),
\]

where \( \gamma : \mathbb{R}^m \times \mathbb{R} \to \mathbb{C} \) denotes the coefficient function; \( \sigma : \mathbb{R} \to \mathbb{C} \) denotes an activation function such as Gaussian, hyperbolic tangent, sigmoidal function and rectified linear unit (ReLU).
The integral representation (1) is a continuum limit of an ordinary shallow neural network
\[ g(x) = \sum_{j=1}^{p} c_j \sigma(a_j \cdot x - b_j). \] (2)

Namely, the ordinary shallow neural network (2) is a Riemannian sum (i.e. numerical integration) of the integral representation (1). Therefore, in general, there are no one-to-one correspondences between (1) and (2). A simple example of one-to-one correspondence is a singular measure: \( \gamma = \sum_{j=1}^{p} c_j \delta(a_j, b_j) \) with Dirac’s \( \delta \). The numerical integration method for the integral representation is investigated by Candès (1998) and Sonoda and Murata (2014). In this paper, we refer \( S[\gamma] \) to a “neural network.”

The integral representation has two advantages: (a) the parameter is linear, and (b) the pseudo-inverse is known. First, in (2), the output parameters \( c_j \) are linear but the hidden parameters \( (a_j, b_j) \) are nonlinear, which makes the BP training non-convex. On the other hand, in (1), the output parameter \( \gamma(a, b) \) is linear and every possible hidden parameters \( (a, b) \) are integrated out. Therefore, in (1), we do not need to select which \( (a, b) \)'s to use. Instead, the weight function \( \gamma(a, b) \) automatically select \( (a, b) \)'s by weighting on them. Second, the pseudo-inverse to \( S \) is known. Namely, the ridgelet transform \( R \) is the one. As explained in the next subsection, we can use \( R \) to obtain an appropriate \( \gamma(a, b) \).

1.2 Ridgelet Transform

We recall several backgrounds of the classical ridgelet transform. We remark that in this paper, we use a modified version of this transform to prove the main result. The classical one appears as a limit of the modified one.

The (classical) ridgelet transform of \( f : \mathbb{R}^m \rightarrow \mathbb{C} \) with respect to a ridgelet function \( \rho : \mathbb{R} \rightarrow \mathbb{C} \) is defined by
\[ R[f](a, b) := \int_{\mathbb{R}^m} f(x) \rho(a \cdot x - b) dx. \] (3)

Let us consider an integral equation \( S[\gamma] = f \). Assume that \( \rho \) in \( R \) and \( \sigma \) in \( S \) satisfy a coupling condition, called the admissibility condition (Sonoda and Murata, 2017b), then the ridgelet transform \( R[f] \) is a particular solution to the equation. Namely, \( R \) is a pseudo-inverse that satisfies the reconstruction formula \( S[R[f]] = f \).

The reconstruction formula suggests the universal approximation property that a neural network \( S[\gamma] \) with a coefficient function \( \gamma = R[f] \) equals \( f \). As demonstrated in §2, the ridgelet transform can be computed by numerical integration. See Sonoda and Murata (2017b) for more details on ridgelet analysis.

1.3 BP in the Integral Representation

Let \( \mathcal{F} \) and \( \mathcal{G} \) denote Hilbert spaces of data \( f : \mathbb{R}^m \rightarrow \mathbb{C} \) and coefficient functions \( \gamma : \mathbb{R}^m \times \mathbb{R} \rightarrow \mathbb{C} \), respectively. We reformulate the BP training of an ordinary neural network as an optimization problem \( \min_{\gamma \in \mathcal{G}} L[\gamma; f, \beta] \), where
\[ L[\gamma; f, \beta] := \| S[\gamma] - f \|_\mathcal{F}^2 + \beta \| \gamma \|_\mathcal{G}^2. \] (4)
Figure 2: Loss surface $L$ of the BP training in the ordinary parameterization is (a) and that in the integral representation is (b).

While an ordinary BP training is non-convex (Figure 2a), the BP training in the integral representation is strongly convex (Figure 2b). Here, we note that due to the regularization term $\beta \|\gamma\|_2^2$, (4) becomes strong-convexity. Further investigations on other regularizations such as $L^1$ remains as our future work.

According to the Tikhonov regularization theory for linear operators, if $S$ is a Lipschitz continuous operator between Hilbert spaces, then we can find the global minimizer as

$$\arg\min_{\gamma \in G} L[\gamma; f, \beta] = (\beta + S^*S)^{-1} S^*f.$$  

(5)

Here $S^*$ is the dual operator of $S$. See Appendix A for the derivation.

In (5), we have two difficulties. One is that the Lipschitz continuity of $S$ is not obvious, and the other is that the numerical computation method for $(\beta + S^*S)^{-1} S^*f$ is not trivial. First, we note that in the infinite dimensional settings, not every linear operator is continuous, while in the finite dimensional settings, every linear operator, or a matrix, is continuous. To ensure the continuity, we specify and investigate the several properties of the function spaces $F$ and $G$ (see § 3 for more details).

Second, in (5), both $S^*$ and $(\beta + S^*S)^{-1}$ are only symbolic and lack measures for numerical computation. As explained in the next subsection, we give an explicit formula.

---

1In functional analysis, Lipschitz continuity is usually referred to as “bounded.” However, in machine learning, it is sometimes confused with “bounded function.” Thus, we use “Lipschitz continuous,” in order to avoid confusion.
1.4 Main Result

By using the modified ridgelet transform \( R^\sigma_\rho : \mathcal{F} \rightarrow \mathcal{G} \) (see (13) for the definition), we obtained an explicit expression of the global minimizer as

\[
\arg\min_{\gamma \in \mathcal{G}} L[\gamma; f, \beta] = (\beta + S^* S)^{-1} S^* f
\]

(5)

\[
= \frac{1}{\beta + 1} R^\sigma_\rho[f].
\]

(6)

That is, we have a numerically computable expression of the global minimizer.

1.5 Related Works

Recent theoretical investigations on the local minima and energy landscape start by addressing the following conjecture: Most of the local minima are either close to or equal to the global minimum (Baldi and Hornik, 1989; Dauphin et al., 2014; Choromanska et al., 2015b; Goodfellow et al., 2016). This conjecture has now been solved under many different assumptions, such as the random matrix theory (Choromanska et al., 2015a), linear activations (Kawaguchi, 2016; Lu and Kawaguchi, 2017; Yun et al., 2018), partially linear activation (Soudry and Carmon, 2016), and strictly monotone smooth activation (Nguyen and Hein, 2017). These days, some researchers are going beyond this conjecture. For example, Draxler et al. (2018) and Garipov et al. (2018) reported that the local minima are not isolated points but connected along paths. In this study, instead of analyzing the loss surface with linearity assumptions, we present an explicit expression of the global minimizer by reparameterizing the loss function in the integral representation.

Traditionally, the integral representation and ridgelet transform has been developed to estimate the approximation ability of neural networks (Barron, 1993; Murata, 1996; Candès, 1998), and to estimate the discretization errors (Kůrková, 2012). Recently, it has been applied to synthesize neural networks without BP training, by approximating the integral transform with a Riemannian sum (Sonoda and Murata, 2014; Bach, 2017a,b); to facilitate the inner mechanism of the so-called “black-box” networks (Sonoda and Murata, 2017a), and to estimate the generalization errors of deep neural networks from the decay of eigenvalues (Suzuki, 2018).

2 MOTIVATING EXAMPLE

In Figure 1, we have compared the BP trained parameters and the spectrum of a ridgelet transform. We explain the details of these two experiments. Further examples are shown in Supplements.

In Supplements, we employed 2 activation functions: tanh and ReLU; and used 5 optimization methods: LBFGS, SGD, ADAM, SGD with weight decay, and ADAM with weight decay. Note that weight decay has an equivalent effect to \( L^2 \) regularization. Every experiment is conducted with the same dataset \( \{(x_i, y_i)\}_{i=1}^{200} \), where \( x_i \in [-1, 1] \) and \( y_i = \sin 2\pi x_i \).

The results showed that irrespective of the optimization methods, every BP training result with tanh networks and ReLU networks converged to a 10-point star and 12-point star, respectively. At the same time, the ridgelet transforms for tanh and ReLU were 10-point star shaped and 12-point star shaped, respectively. Thus, we can expect that the BP minimizer may be expressed as a ridgelet transform.
In both cases, LBFGS showed the most similar pattern to the ridgelet spectrum, while the rest 4 were slightly different from LBFGS, and similar to each other irrespective of the weight decay. It suggests that both SGD and ADAM have by themselves an implicit regularization effect. In some results such as ReLU with SGD, there are hourglass-shaped noises in the center. We understand them as noise because the corresponding output parameters are relatively small, which means the hidden units are disconnected.

2.1 Scatter Plots of BP Trained Parameters

We trained \( n = 1,000 \) neural networks with \( p = 10 \) tanh units, and another \( n = 1,000 \) neural networks with \( p = 20 \) ReLU units. Namely, we have \( np = 10,000 \) or 20,000 points in the \((a, b)\)-space. The difference in \( p \) is simply due to balancing the training errors. Namely, ReLU networks doubled the unit number to achieve an equivalent accuracy to tanh networks.

2.2 Numerical Integration of Ridgelet Spectrum

For the sake of numerical integration, we employed the following classical version of the ridgelet transform:

\[
R[f](a, b) := \int_{\mathbb{R}^m} f(x)\overline{\rho(a \cdot x - b)} dx. \tag{7}
\]

Here, we employed \( \hat{\rho}(\zeta) = |\zeta|^2 \exp(-\zeta^2/2) \) when \( \sigma \) is tanh, and \( \hat{\rho}(\zeta) = |\zeta|^3 \exp(-\zeta^2/2) \) when \( \sigma \) is ReLU. (\( \hat{\cdot} \) denotes the Fourier transform.) These choices are admissible, which means that the reconstruction formula \( R[S[f]] = f \) holds. See (Sonoda and Murata, 2017b) for more details.

Given the dataset \( \{(x_i, y_i)\}_{i=1}^{200} \), we have evaluated a simple numerical integration:

\[
R[f](a, b) \approx \sum_{i=1}^{200} y_i \rho(a \cdot x_i - b) \Delta x, \tag{8}
\]

with \( \Delta x = 1/100 \), at every point \((a, b) = (-15 + p\Delta a, -15 + q\Delta b) (0 \leq p, q \leq 300) \) with grids \( \Delta a = \Delta b = 1/10 \).

Note that discretization methods for a ridgelet transform is not limited to the way described above.

3 THEORY

We start from three hyper parameters: activation function \( \sigma : \mathbb{R} \to \mathbb{C} \), ridgelet function \( \rho : \mathbb{R} \to \mathbb{C} \), and a Borel measure \( \lambda \) on \( \mathbb{R}^m \times \mathbb{R} \).

We write \( \sigma_x(a, b) = \sigma_{a,b}(x) = \sigma(a \cdot x - b) \) and \( \rho_x(a, b) = \rho_{a,b}(x) = \rho(a \cdot x - b) \).

3.1 Space \( \mathcal{G} \) of Coefficient Functions

We employ \( \mathcal{G} = L^2(\lambda) \), i.e. a Hilbert space with inner product \( \langle \gamma, \delta \rangle_\mathcal{G} := \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(a, b) \overline{\delta(a, b)} d\lambda(a, b) \).

Assume that \( \sigma_x \in \mathcal{G} \) and \( \rho_x \in \mathcal{G} \). Let \( \mathcal{G}^\phi \) be a closed subspace in \( \mathcal{G} \) of functions spanned by \( \phi_x(a, b) \), i.e. the closure of \( \text{Span} \{\phi_x\}_{x \in \mathbb{R}^m} \) in \( \mathcal{G} \). Let \( \mathcal{G}_\rho := (\mathcal{G}^\sigma \cap \mathcal{G}^\rho) \oplus (\mathcal{G}^\sigma)^\perp = \mathcal{G}^\rho + (\mathcal{G}^\sigma)^\perp \).
As explained later in Theorem 3.1 and Theorem 3.2, these auxiliary spaces are important because \( \text{Ker} S = (G) \perp \) and \( S^{-1}(H^\rho) = G^\rho. \)

In the classical setup, \( d\lambda(a, b) = d\lambda db. \) We introduce \( \lambda \) in order for \( G \) to be a Hilbert space that contains \( \sigma_x. \) Recall that even in the classical setup, for the numerical computation of \( \gamma \in G, \) we must assume a finite Borel measure \( \lambda \) either implicitly or explicitly.

For instance, commonly used activation functions such as Gaussian, tanh and rectified linear unit (ReLU) satisfy \( \sigma_x \in G, \) when \( \lambda \) is a Gauss measure.

### 3.2 RKHS Associated with \( \sigma, \rho \) and \( \lambda \)

Let

\[
k^{\sigma\rho}(x, y) := \langle \sigma_x, \rho_y \rangle_G.
\]

We say that the triplet \( \sigma, \rho, \) and \( \lambda \) is admissible when \( k^{\sigma\rho} \) is a positive definite kernel. When \( \sigma = \rho, \) the triplet is always admissible. When \( \sigma \neq \rho, \) two sufficient conditions are given in Theorem 4.1 and Theorem 4.3. In the following arguments, we always assume that \( \sigma, \rho, \) and \( \lambda \) are admissible.

Provided that \( k^{\sigma\rho} \) is positive definite, then by the Moore-Aronszajn theorem, there exists an reproducing kernel Hilbert space (RKHS) \( H^{\sigma\rho} \) with inner product \( \langle \cdot, \cdot \rangle_{\sigma\rho} \) such that the reproducing property holds:

\[
\langle f, k^{\sigma\rho}(x) \rangle_{\sigma\rho} = f(x).
\]

### 3.3 Space \( F \) of Data, or Hypothesis

We employ \( F = H^{\sigma\rho}. \) According to Theorem 3.2, \( H^{\sigma\rho} \) corresponds to the hypothesis class of the neural networks \( \{S[\gamma] : \gamma \in G^\rho\}. \) Namely, \( H^{\sigma\rho} \) is the collection of all possible functions \( f : \mathbb{R}^m \to \mathbb{R} \) that a neural network \( S[\gamma] \) with \( \gamma \in G^\rho \) can approximate.

Employing \( F = H^{\sigma\rho} \) appears to be artificial. Fortunately, as discussed in § 4.2, examples of \( k^{\sigma\rho} \) are not artificial at all, and \( H^{\sigma\rho} \) can cover a wide range of functions. Namely, every \( L^2 \) function is contained in the limit case, \( d\lambda(a, b) = d\lambda ab, \) or the uniform distribution over the whole Euclidean space.

At first, we worked on \( F = L^2(p) \) with data distribution \( p, \) which appears to be more natural. However, to show the continuity of \( S : L^2(\lambda) \to L^2(p) \) is difficult. At last, we discovered \( H^{\sigma\rho} \) and found that to show the continuity of \( S : L^2(\lambda) \to H^{\sigma\rho} \) is easier, because both \( \text{Ker} S \) and \( \text{Im} S \) are remarkably clear.

### 3.4 Integral Representation \( S : G \to F \)

First of all, we can rewrite \( S[\gamma](x) = \langle \gamma, \sigma_x \rangle_G \) and \( k^{\sigma\rho}(x, y) = S[\rho^\sigma]_{G}(x). \)

**Theorem 3.1.** Provided that \( \sigma_x \in G. \) Then, \( \text{Dom} S = G, \) \( \text{Ker} S = (G^\sigma)^\perp, \) \( \text{Im} S = H^{\sigma\rho}, \) which are summarized as

\[
\begin{array}{ccc}
\text{Dom} S = G & \xrightarrow{S} & \text{Im} S = H^{\sigma\rho} \\
\downarrow & & \downarrow \\
G/\text{Ker} S \cong G^\sigma
\end{array}
\]
The proof is given in Appendix B.1. Note that $H^{\sigma\sigma}$ is not a typo but a special case of $H^{\rho\rho}$ when $\rho = \sigma$.

**Theorem 3.2.** Provided that $\sigma_x \in \mathcal{G}$ and $\rho_x \in \mathcal{G}$. Then, $S^{-1}(H^{\rho\rho}) = \mathcal{G}_\rho$ and $H^{\rho\rho} \subset H^{\sigma\sigma}$.

The proof is immediate from $k^{\sigma\rho}(x, y) = S[\overline{\mathcal{G}}_\rho](x)$.

These theorems are the key because we can always write an arbitrary $\gamma \in \mathcal{G}_\rho$ as

$$\gamma = \sum_j c_j \rho_{x_j} \oplus \gamma^\perp, \quad (11)$$

with some $c_j \in \mathbb{C}, x_j \in \mathbb{R}^m$ and $\gamma^\perp \in (\mathcal{G}^\sigma)^\perp$; and $\gamma$ satisfies

$$S \left[ \sum_j c_j \rho_{x_j} \oplus \gamma^\perp \right] = \sum_j c_j k^{\sigma\rho}_{x_j}. \quad (12)$$

In general, basic information such as null space and image of “X-let” transform cannot be completely specified. Nevertheless, for our $S$ (and $R$), such information is remarkably clear.

**Theorem 3.3.** Provided that $\sigma_x \in \mathcal{G}$. Then $S : \mathcal{G} \rightarrow H^{\sigma\sigma}$ is Lipschitz continuous.

The proof is given in Appendix B.2, performed by using (11).

### 3.5 Ridgelet Transform on RKHS

Let $\phi = \sigma$ or $\rho$. We define the ridgelet transform of $f \in H^{\rho\rho}$ with respect to ridgelet function $\phi$ as

$$R^\rho_{\sigma\rho}[f](a, b) := \langle f, \rho_{a,b}\rangle_{\sigma\rho}. \quad (13)$$

Here, $R^\rho_{\sigma\rho}$ is a pseudo-inverse to $S$, and $R^\rho_{\sigma\rho}$ is the dual operator of $S|_{\rho}$.

**Theorem 3.4 (Pseudo-inverse).** $[SR^\rho_{\sigma\rho}] = \text{Id}_{H^{\rho\rho}}$.

$\therefore SR^\rho_{\sigma\rho}[f](x) = \langle \langle f, \rho_{a,b}\rangle_{\sigma\rho}, x \rangle \rangle_{\mathcal{G}} = \langle f, k^{\sigma\rho}_{x} \rangle_{\sigma\rho} = f(x)$.

**Theorem 3.5 (Dual Operator).** $(S|_{\rho})^* = R^\rho_{\sigma\rho}$.

$\therefore \langle \gamma, R^\rho_{\sigma\rho}[k^{\sigma\rho}_{x}] \rangle_{\mathcal{G}} = \langle \gamma, x \rangle_{\mathcal{G}} = S[\gamma](x) = \langle S\gamma, k^{\sigma\rho}_{x} \rangle_{\sigma\rho}$.

The following formula frequently appears in the proofs.

**Lemma 3.6.** $R^\rho_{\sigma\rho}[k^{\sigma\rho}_{x}] = \overline{\phi_x}$.

The proof is given in Appendix B.3. For example, we can derive translation laws: $S[R^\rho_{\sigma\rho}[k^{\sigma\rho}_{x}]] = S[\overline{\phi_x}] = k^{\sigma\phi}_{x}$ and $R^\rho_{\phi}[S[\overline{\mathcal{G}}_\rho]] = R^\rho_{\phi}[k^{\sigma\rho}_{x}] = \overline{\phi_x}$.

The following two theorems are proved by using the lemma.

**Theorem 3.7.** Dom $R^\rho_{\rho} = H^{\rho\rho}$, Im $R^\rho_{\rho} = \mathcal{G}^\rho$ and $R^\rho_{\rho} : H^{\rho\rho} \rightarrow \mathcal{G}$ is injective.

**Theorem 3.8.** Suppose that there exists $L > 0$ such that $L k^{\sigma\rho} - k^{\phi\phi}$ is positive definite. Then, the ridgelet transform $R^\rho_{\phi} : H^{\rho\rho} \rightarrow \mathcal{G}^\phi$ is Lipschitz continuous.

See Appendix B.4 for the proof.
3.6 Integral Representation of Global Minimizer

We consider an optimization problem that minimizes (4) with \( F = H^{\sigma \rho} \) and \( G = L^2(\lambda) \).

Note that the minimization in \( G \) is equivalent to the minimization in \( G_\rho \), because if \( \gamma \in G \setminus G_\rho \), then \( \| f - S[\gamma] \|_{\sigma \rho} \) diverges. In this case, \( S \) is automatically restricted to \( S|_{\rho} : G_\rho \rightarrow H^{\sigma \rho} \).

**Theorem 3.9** (Main Result). Let \( f \in H^{\sigma \rho}, \beta > 0 \), and \( L[\gamma; f, \beta] \) is given by (4). Assume that \( \sigma x \in G \) and \( \rho x \in G \). Then,

\[
\text{argmin}_{\gamma \in G} L[\gamma; f, \beta] = \frac{1}{\beta + 1} R_{\rho}^{\sigma \rho}[f]. \tag{14}
\]

The result indicates that the global minimizer is a shrink estimator. Namely, \( S[\text{argmin}_{\gamma \in G} L[\gamma; f, \beta]] = (\beta + 1)^{-1} f \). This is consistent with a common understanding that the \( L^2 \)-regularization results in a shrink estimation. The hyperparameter \( \beta \) control the intensity of the noise in \( f \).

By letting \( \beta \rightarrow 0 \), (14) converges to an ordinary ridgelet transform \( R_{\rho}^{\sigma \rho} : H^{\sigma \rho} \rightarrow G^\rho \). It means that the ridgelet transform \( R_{\rho}^{\sigma \rho} \) provides the minimum norm solution under the assumption that \( f \in H^{\sigma \rho} \). This is consistent with a general result of the Tikhonov regularization theory that \( \lim_{\beta \rightarrow 0}(\beta + S^* S)^{-1} S^* f \in (\text{Ker} S)^\perp = G^\rho \cap G^\sigma \). See Appendix A for more details.

3.7 Proof of Main Result

Because \( S^* = R_{\rho}^{\sigma \rho}, S^* S|_{\rho} \) is the projection

\[
G_\rho \ni \sum_j c_j \rho x_j \oplus \gamma^\perp \mapsto \sum_j c_j \rho x_j \in G^\rho.
\]

(15)

So, let

\[
P : G_\rho \ni \sum_j c_j \rho x_j \oplus \gamma^\perp \mapsto \gamma^\perp \in (G^\rho)^\perp.
\]

(16)

Then, \( \beta + S^* S = \beta P \oplus (\beta + 1) S^* S : G_\rho \rightarrow G \), and

\[
(\beta + S^* S)^{-1} = \frac{1}{\beta} P \oplus \frac{1}{\beta + 1} R_{\rho}^{\sigma \sigma} S.
\]

(17)

Therefore,

\[
\text{argmin}_{\gamma \in G_\rho} L[\gamma; \cdot, \beta] = (\beta + S^* S)^{-1} S^* = \left( \frac{1}{\beta} P \oplus \frac{1}{\beta + 1} R_{\rho}^{\sigma \sigma} S \right) R_{\rho}^{\sigma \rho} = \frac{1}{\beta + 1} R_{\rho}^{\sigma \sigma} S R_{\sigma}^{\rho} = \frac{1}{\beta + 1} R_{\rho}^{\sigma \rho}.
\]

Here, the third equation holds because \( P R_{\sigma}^{\rho} : H^{\sigma \rho} \rightarrow G^\rho \rightarrow 0 \); and the last equation holds because \( R_{\rho}^{\sigma \sigma} S R_{\sigma}^{\rho}[k] = \overline{\rho x} \), which is the same map as \( R_{\rho}^{\sigma \rho} \).
4 FURTHER INVESTIGATIONS ON RKHS WITH SEPARABLE $\lambda$

We investigate on the special case when the measure $\lambda(a, b)$ is separable. Namely, there exists a probability density function $\mu(a)$ such that

$$d\lambda(a, b) = \mu(a) da db. \quad (18)$$

In this case, non-integrable activation functions such as tanh and ReLU cannot satisfy $\sigma_x \in G$ for any $\mu$. This is because

$$\|\sigma_x\|_2^2 = \int_{\mathbb{R}^m} |\sigma(a \cdot x - b)|^2 \mu(a) da db = \int_{\mathbb{R}^m} \mu(a) da \int_{\mathbb{R}} |\sigma(b)|^2 db = 1 \cdot \infty.$$

Nevertheless, we investigate on the case because $k^{\sigma\rho}$ has the Fourier expression. Hereafter, $\hat{\cdot}$ denotes the Fourier transform.

**Theorem 4.1.** Let $k^{\sigma\rho}(x, y) = \langle \sigma_x, \rho_y \rangle_G$ with $d\lambda(a, b) = \mu(a) da db$. Then,

$$k^{\sigma\rho}(x, y) = \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} \rho^{\sigma\rho}(a) e^{ia \cdot (x-y)} da, \quad (19)$$

where

$$p^{\sigma\rho}(a) := (2\pi)^{m-1} \int_{\mathbb{R}} \frac{\hat{\sigma}(\zeta) \hat{\rho}(\zeta)}{|\zeta|^m} \mu \left( \frac{a}{\zeta} \right) d\zeta. \quad (20)$$

Immediately, if $p^{\sigma\rho}$ is a probability density function, then according to Bochner’s theorem, $k^{\sigma\rho}$ is a shift invariant positive definite kernel. The inner product is given by

$$\langle f, g \rangle_{\sigma\rho} = \int_{\mathbb{R}^m} \frac{\hat{f}(\xi) \hat{g}(\xi)}{p^{\sigma\rho}(\xi)} d\xi. \quad (21)$$

The proof of this is given in Appendix B.5.

In association with (19), we have the Fourier expression of ridgelet transform.

**Theorem 4.2.**

$$R^{\sigma\rho}_{\phi} [f](a, b) = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{\hat{f}(\zeta) \hat{\phi}(\zeta)}{p^{\sigma\rho}(\zeta)} e^{ib\zeta} d\zeta. \quad (22)$$

The proof is given in Appendix B.6

4.1 Examples of Admissible Triplets

For a fixed activation function $\sigma$, there are literally an infinite number of $\rho$ that is admissible with $\sigma$. This reflects a general fact that there are infinitely many different coefficient functions $\gamma_1(a, b), \gamma_2(a, b), \ldots$ of neural networks that realize the same function $S[\gamma_1](x) = S[\gamma_2](x) = \cdots$. In the following proposition, we assume that $\sigma, \rho \in L^1 \cap L^2(\mathbb{R})$ for the sake of simplicity. Note that this is not necessary.

**Theorem 4.3.** For a fixed $\sigma \in L^1 \cap L^2(\mathbb{R})$, $p^{\sigma\rho}$ is a probability density function when $\rho \in L^1 \cap L^2(\mathbb{R})$ satisfies $\int_{\mathbb{R}} \sigma(z) \rho(z) dz = 1$ and $\hat{\sigma}(\zeta) \hat{\rho}(\zeta) \geq 0$. 

10
Recall that the positivity \( \tilde{\sigma}(\zeta) \tilde{\rho}(\zeta) \geq 0 \) means that the convolution \( \sigma \ast \rho(z) \) is real even function. Obviously, there are infinitely many \( \rho \)'s that satisfy the sufficient condition in the proposition.

**Proof.** First, if \( \int_{\mathbb{R}} \sigma(z) \overline{\rho(z)} \, dz = 1 \), then \( \int_{\mathbb{R}^m} p^{(\sigma \rho)}(a) \, da = 1 \). This is immediate because \( \int_{\mathbb{R}^m} p^{(\sigma \rho)}(a) \, da = \int_{\mathbb{R}} \tilde{\sigma}(\zeta) \overline{\tilde{\rho}(\zeta)} |\zeta|^{-m} \int_{\mathbb{R}^m} \mu(a/\zeta) \, da \, d\zeta = \int_{\mathbb{R}} \tilde{\sigma}(\zeta) \overline{\tilde{\rho}(\zeta)} \int_{\mathbb{R}^m} \mu(a) \, da \, d\zeta = 1 \).

Second, if \( \tilde{\sigma}(\zeta) \overline{\tilde{\rho}(\zeta)} \geq 0 \), then \( p^{(\sigma \rho)} \geq 0 \). This is immediate because \( p^{(\sigma \rho)}(a) = \int_{\mathbb{R}} \tilde{\sigma}(\zeta) \overline{\tilde{\rho}(\zeta)} |\zeta|^{-m} \mu(a/\zeta) \, d\zeta = \int_{\mathbb{R}} \tilde{\sigma}(1/\zeta) \overline{\tilde{\rho}(1/\zeta)} |\zeta|^{m+2} \mu(\zeta a) \, d\zeta \), and this is positive for every \( a \in \mathbb{R}^m \), when \( \tilde{\sigma}(1/\zeta) \overline{\tilde{\rho}(1/\zeta)} \geq 0 \).

### 4.2 Examples of Kernels

When \( \sigma(z) = \rho(z) = \pi^{-1/2} \exp(-z^2) \) and \( \mu(a) = (2\pi)^{-m/2} \exp(-|a|^2/2) \), then \( p^{(\sigma \rho)}(a) = (2\pi)^{m/2} \exp(-|a|) \), and

\[
k^{(\sigma \rho)}(x, y) \propto (1 + |x - y|^2)^{-m/2},
\]

which means that \( k^{(\sigma \rho)} \) decays slowly, and thus the corresponding RKHS \( H^{(\sigma \rho)} \) is small.

When \( \mu(a) \) is a constant function 1, then \( p^{(\sigma \rho)}(a) \) is a constant in \( a \), and

\[
k^{(\sigma \rho)}(x, y) \propto \delta(x - y),
\]

which means that \( H^{(\sigma \rho)} \) is no more a RKHS, but coincides with \( L^2(\mathbb{R}^m) \).

These examples suggest that there is a trade-off between the localization in \( a \) and \( x \).

### 5 CONCLUSION

In this study, we obtained the integral representation \( S[\gamma] \) that attains the global minimum \( \min_{\gamma \in \mathcal{G}} L[\gamma; f, \beta] \) of the BP training problem that is reformulated in the integral representation. The interesting similarity between the distribution of BP trained parameters \( \{ (a_j, b_j) \} \) and ridgelet spectrum \( R[f](a, b) \) is now clear because the global minimum in the integral representation is attained by the shrink ridgelet transform \( (\beta + 1)^{-1} R[f](a, b) \).

#### 5.1 Contributions

(a) We reformulated the BP training in the integral representation, as in (4), which results in a strongly convex optimization problem, where the global minimizer is formally known as (5), if the integral representation \( S : \mathcal{G} \rightarrow \mathcal{F} \) is Lipschitz continuous.

(b) We proved \( S : \mathcal{G} \rightarrow \mathcal{F} \) is Lipschitz continuous when \( \mathcal{F} = H^{(\sigma \rho)} \). As discussed in § 4.2, \( H^{(\sigma \rho)} \) covers a wide range of functions. The representation (11) of \( \gamma \in \mathcal{G} \) is the key to many theorems.

(c) We showed in Theorem 3.9 that the global minimizer is given by (6). While (5) is symbolic and intractable, (6) is analytic and essentially tractable. Namely, the limit case \( d\lambda(a, b) = da \, db \) can be computed as performed in § 2.

(d) We established Fourier expressions for a special case and found many examples in § 4.
5.2 Discussions

The shrink ridgelet transform \((\beta + 1)^{-1} R\) provides an explicit expression of the global minimizer. As performed in § 2, it is basically computable by numerical integration. This implies that we may obtain a BP trained neural network, without BP training. As shown in Figure 1 and Supplements, it was really able to synthesize similar patterns to the BP results.

Some readers may be surprised at our results: the global minimizer of BP, because BP is famous for the non-convexity and local minima. However, in general, the (non-)convexity of a function \(L[\theta]\) is not conserved under the change of parameterization from \(\theta\) to another variable, say \(\omega\). Consider a toy example: \(L[\theta] = \theta^2, \theta \in [-1, 1]\), which is convex in \(\theta\); and the change of parameterization: \(\theta(\omega) = \sin(2\pi \omega), \omega \in \mathbb{R}\). Then, \(L[\theta(\omega)] = \sin^2(2\pi \omega), \omega \in \mathbb{R}\) is non-convex in \(\omega\). Of course, the training difficulty of neural networks does not vanish. One difficulty lies in the discretization of \(S[R[f]]\). This is similar to Bayesian neural networks, where the posterior sampling is difficult.

Because the relation between an integral representation and an ordinary finite network is not clear, and BP is convex in the integral representation, we cannot immediately answer the question: “Is a local minimum a global minimum?” Essentially, the ordinary BP is non-convex simply because it is formulated in a bad parametrization. A parameter provides a coordinate system in the function space, but the parameter is not a function itself. In the integral representation, the ideal learning curve is the gradient flow \(\dot{\gamma} = -\text{grad} L[\gamma]\). In contrast, the learning curve by the ordinary BP would detour, if it is transformed in \(G\). Integral representation offers a new direction to theoretical deep learning research.

We have concentrated on a traditional case: shallow neural networks with Tikhonov regularization. Analysis of more modern settings such as deep neural networks with general risk functions will be performed in future work. Currently, we are investigating the global minimizer of deep neural networks in the integral representation using the calculus of variations.

A Tikhonov Regularization Theory for Operators

Let \(H_0, H_1\) be Hilbert spaces, and \(A : H_0 \to H_1\) be a densely defined closed operator.

Consider finding \(f \in H_0\), for a given \(g \in H_1\) that satisfies

\[
Af = g.
\]

Proposition A.1. For every \(\beta > 0\),

\[
\text{argmin}_{f \in H_0} (\|Af - g\|_1^2 + \beta \|f\|_0^2) = (\beta + A^* A)^{-1} A^* g.
\]

Proof.

\[
\|Af - g\|_1^2 + \beta \|f\|_0^2
\]
\[
= (Af, Af)_1 - 2 \Re (Af, g)_1 + (g, g)_1 + \beta (f, f)_0
\]
\[
= (\sqrt{\beta} + A^*Af, \sqrt{\beta} + A^*Af)_0
\]
\[
- 2 \Re (\sqrt{\beta} + A^*Af, \sqrt{\beta} + A^*A^{-1}A^*g)_0 + (g, g)_1
\]
\[
= \|\sqrt{\beta} + A^*Af - \sqrt{\beta} + A^*A^{-1}A^*g\|_0^2 + \text{(nonnegative)}.
\]

Therefore, the objective functional attains the minimum at \(f^* = (\beta + A^* A)^{-1} A^* g\). \qed
Proposition A.2. Suppose that \( f_0 \in H_0 \) satisfies \( g = Af_0 \). Then,
\[
\lim_{\beta \to 0} (\beta + A^*A)^{-1}A^*g = \text{proj } H_0 \to (\text{Ker } A)^\perp f_0. \tag{27}
\]

Proof. Using the right continuous resolution of the identity \( \{E_\mu\}_{\mu \in \mathbb{R}} \) for \( A^*A \),
\[
(\beta + A^*A)^{-1}A^*g = \int_\mathbb{R} \frac{\mu}{\beta + \mu} dE_\mu f_0
\]
\[
\to \int_\mathbb{R} 1_{\mathbb{R}\setminus\{0\}}(0) dE_\mu g, \quad \text{as } \beta \to 0
\]
\[
= (E_0 - E_0-)f_0
\]
\[
= \text{proj } H_0 \to (\text{Ker } A)^\perp f_0. \tag*{\square}
\]

Here, \( (E_0 - E_0-)f_0 \) follows from the projection nature of \( dE_\mu f_0 \).

\section*{B \ PROOFS}

\subsection*{B.1 Theorem 3.1}

Proof. If \( \sigma_\xi \in \mathcal{G} \), then the integral \( S[\gamma] \) is always absolutely convergent for every \( \gamma \in \mathcal{G} \). Namely, \( \text{Dom } S = \mathcal{G} \). This is immediate from the Cauchy-Schwartz inequality:
\[
\int_{\mathbb{R} \times \mathbb{R}} |\gamma(a, b)\sigma(a \cdot x - b)| d\lambda(a, b) \leq ||\gamma||_\mathcal{G} \cdot ||\sigma_\xi||_\mathcal{G} < \infty.
\]

The null space \( \text{Ker } S \) is given by \( (\mathcal{G}^\sigma)^\perp \), because \( \gamma \in \text{Ker } S \) implies \( S[\gamma](x) = \langle \gamma, \overline{\sigma_\xi} \rangle_\mathcal{G} = 0 \) for every \( x \in \mathbb{R}^m \), which implies \( \gamma \in (\mathcal{G}^\sigma)^\perp \).

Therefore, the image space \( \text{Im } S \) is given by \( H^\sigma \), because \( S[\sum_i c_i \sigma_\xi_i] = \sum_i c_i k_{\sigma_i}^\sigma \) for every \( \sum_i c_i \sigma_\xi_i \) \( \in (\mathcal{G}^\sigma)^\perp = \mathcal{G} \).

The isomorphism \( \mathcal{G}/\text{Ker } S \cong \mathcal{G}^\sigma \) is immediate from \( \mathcal{G} = \mathcal{G}^\sigma \oplus (\mathcal{G}^\sigma)^\perp \) and \( \text{Ker } S = (\mathcal{G}^\sigma)^\perp. \tag*{\square}

\subsection*{B.2 Theorem 3.3}

Proof. Take an arbitrary \( \gamma = (\sum_i c_i \sigma_\xi_i) \oplus \gamma^\perp \in \mathcal{G}^\sigma \oplus (\mathcal{G}^\sigma)^\perp \). Observe that \( ||S[(\sum_i c_i \sigma_\xi_i) \oplus \gamma^\perp]||_2^2 = \| \sum_i c_i k_{\sigma}^\sigma \|^2_2 = \sum_i c_i \overline{\sigma_\xi} k_{\sigma}^\sigma(x_i, x_j) = \sum_i c_i \overline{\sigma_\xi} \rho(x_i, x_j) \rho \mathcal{G} = \sum_i c_i \sigma_\xi_i, \quad \text{and } ||\gamma||_2^2 = ||\gamma^\perp||_2^2 + \| \sum_i c_i \sigma_\xi_i \|^2_\mathcal{G}. \) Then, \( S[\gamma] \) is bounded: \( ||S[\gamma]||_{\mathcal{G}^\sigma} \leq ||\gamma||_\mathcal{G}. \tag*{\square}

\subsection*{B.3 Lemma 3.6}

Proof. Immediate because \( S[R^\rho_{\phi} | k_{x_i}^\sigma](y) = \left( \langle k_{x_i}^\sigma, \phi_{\alpha, \beta} \rangle_{\mathcal{G}}, y \rangle_\mathcal{G} = \langle k_{x_i}^\sigma, k_y^\phi \rangle_{\mathcal{G}} = \overline{k_y^\phi}(x) = \langle \overline{\phi_{x_i}}, y \rangle_\mathcal{G} = S[\overline{\phi_x}](y). \tag*{\square}

\subsection*{B.4 Theorem 3.8}

Proof. Observe that \( ||R^\rho_{\phi} (\sum_i c_i k_{x_i}^\sigma) ||_2^2 = \sum_i c_i \overline{\sigma_\xi} k_{\sigma}^\phi(x_i, x_j) \) and \( \| \sum_i c_i k_{x_i}^\sigma \|^2_2 = \sum_i c_i \overline{\sigma_\xi} k_{\sigma}^\phi(x_i, x_j) \). Hence, the boundedness: \( ||R^\rho_{\phi} (\sum_i c_i k_{x_i}^\sigma) ||_\mathcal{G} \leq L' \| \sum_i c_i k_{x_i}^\sigma \|_{\mathcal{G}^\sigma} \) is equivalent to the positive definiteness: \( Lk_{\sigma}^\rho - k_{\sigma}^\phi > 0. \tag*{\square} \)
B.5 Theorem 4.1

Proof.

\[ k^{\sigma \rho}(x, y) = \int_{\mathbb{R}^m \times \mathbb{R}} \sigma(a \cdot x - b) \rho(a \cdot y - b) \mu(a) da db \]

\[ = \int_{\mathbb{R}^m} (\sigma * \overline{\rho})(a \cdot (x - y)) \mu(a) da, \]

\[ = \frac{1}{2\pi} \int_{\mathbb{R}^m} \overline{\sigma(\zeta) \overline{\rho(\zeta)}} \exp(i a \cdot (x - y)) d\zeta \mu(a) da \]

\[ = \frac{1}{2\pi} \int_{\mathbb{R}^m} \overline{\sigma(\zeta) \overline{\rho(\zeta)}} \frac{a}{\zeta} d\zeta \exp(i a \cdot (x - y)) da, \]

\[ = \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} p^{\sigma \rho}(a) e^{i a \cdot (x - y)} da. \]

Here, * denotes convolution, and \( \overline{\cdot} \) denotes reflection: \( \overline{\rho}(b) := \rho(-b) \). Hence, according to Bochner’s theorem, \( k^{\sigma \rho} \) is positive definite if \( p^{\sigma \rho} \) is a probability density. Write \( k^{\sigma \rho}_x(y) := k(y, x) \). Recall that \( \hat{k}^{\sigma \rho}_x(\xi) = p^{\sigma \rho}(\xi) e^{-ix \cdot \xi} \). Therefore, we can check the reproducing property:

\[ \langle f, k^{\sigma \rho}_x \rangle = \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} \hat{f}(\xi) \hat{k}^{\sigma \rho}_x(\xi) \frac{p^{\sigma \rho}(\xi)}{p^{\sigma \rho}(\xi)} d\xi = \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} \hat{f}(\xi) p^{\sigma \rho}(\xi) e^{-ix \cdot \xi} d\xi = f(x). \]

B.6 Theorem 4.2

Proof. The Fourier transform of \( \phi_{ab} \) is calculated as described below. For an arbitrary fixed \( a \in \mathbb{R}^m \), let \( a u = a \) with unit vector \( u \in S^{m-1} \) and radius \( a > 0 \). Use direct sum decompositions \( x = pu \oplus x^\perp \) and \( \xi = \zeta u \oplus \xi^\perp \), where \( p, \zeta \in \mathbb{R} \) and \( x^\perp, \xi^\perp \in (\mathbb{R}u)^\perp \cong \mathbb{R}^{m-1} \).

\[ \hat{\phi}_{ab}(\xi) = \int_{\mathbb{R}^m} \phi(a \cdot x - b) e^{-ix \cdot \xi} dx \]

\[ = \int_{\mathbb{R}^{m-1}} \int_{\mathbb{R}} \phi(ap - b) e^{-ip \zeta - ix^\perp \cdot \xi^\perp} dp dx^\perp \]

\[ = (2\pi)^{m-1} \delta(\xi^\perp) \hat{\phi}(\zeta/a) / a e^{-i\zeta b/a}. \]

Hence, the ridgelet transform is calculated as follows.

\[ R^\rho_\phi[f](a, b) := \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} \hat{f}(\xi) \overline{\hat{\phi}_{ab}(\xi)} \frac{p^{\rho\sigma}(\xi)}{p^{\rho\sigma}(\xi)} d\xi \]

\[ = \frac{1}{2\pi} \int_{\mathbb{R}^{m-1}} \int_{\mathbb{R}} \frac{\hat{f}(\zeta u \oplus \xi^\perp)}{p^{\rho\sigma}(\zeta u \oplus \xi^\perp)} \]

\[ \times \delta(\xi^\perp) \phi(\zeta/a) / a e^{-i\zeta b/a} d\xi d\zeta \]

\[ = \frac{1}{2\pi} \int_{\mathbb{R}} \frac{\hat{f}(\zeta a) \phi(\zeta)}{p^{\rho\sigma}(\zeta a)} e^{ib \cdot \zeta} d\zeta. \]
Acknowledgements

This work was supported by JSPS KAKENHI 18K18113.

References

Bach, F. (2017a). Breaking the Curse of Dimensionality with Convex Neural Networks. *Journal of Machine Learning Research*, 18(19):1–53.

Bach, F. (2017b). On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions. *Journal of Machine Learning Research*, 18:1–38.

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945.

Candès, E. J. (1998). *Ridgelets: theory and applications*. PhD thesis, Stanford University.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2015a). The Loss Surfaces of Multilayer Networks. In *AISTATS*, volume 38.

Choromanska, A., LeCun, Y., and Ben Arous, G. (2015b). Open Problem: The landscape of the loss surfaces of multilayer networks. In *COLT*, volume 40, pages 1–5.

Dauphin, Y., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *NIPS*, pages 2933–2941.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. A. (2018). Essentially No Barriers in Neural Network Energy Landscape.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. (2018). Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs.

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

Kawaguchi, K. (2016). Deep Learning without Poor Local Minima. In *NIPS*, pages 586–594.

Kůrková, V. (2012). Complexity estimates based on integral transforms induced by computational units. *Neural Networks*, 33:160–167.

Lu, H. and Kawaguchi, K. (2017). Depth Creates No Bad Local Minima.

Murata, N. (1996). An integral representation of functions using three-layered networks and their approximation bounds. *Neural Networks*, 9(6):947–956.

Nguyen, Q. and Hein, M. (2017). The Loss Surface of Deep and Wide Neural Networks. In *ICML*, volume 70, pages 2603–2612.
Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 779–788.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., and Hassabis, D. (2017). Mastering the game of Go without human knowledge. Nature, 550:354–359.

Sonoda, S. and Murata, N. (2014). Sampling hidden parameters from oracle distribution. In ICANN, volume 8681, pages 539–546.

Sonoda, S. and Murata, N. (2017a). Double Continuum Limit of Deep Neural Networks. In ICML 2017 Workshop on Principled Approaches to Deep Learning, pages 1–5.

Sonoda, S. and Murata, N. (2017b). Neural network with unbounded activation functions is universal approximator. Applied and Computational Harmonic Analysis, 43(2):233–268.

Soudry, D. and Carmon, Y. (2016). No bad local minima: Data independent training error guarantees for multilayer neural networks.

Suzuki, T. (2018). Fast generalization error bound of deep learning from a kernel perspective. In AISTATS, volume 84, pages 1397–1406.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio.

Yun, C., Sra, S., and Jadbabaie, A. (2018). Global Optimality Conditions for Deep Neural Networks. In ICLR, pages 1–14.
Figure 3: Further Motivating Examples
Figure 4: Further Motivating Examples