ニューラルネットワーク

師玉 康成

平成 13 年 11 月 30 日

第1章　ニューラルネットワークについて

1.1　はじめに

この Web，VOD 教材は、階層型ニューラルネットワークについて解説します。ニューラルネットワークはパーセプトロンの研究に始まります。

その動機は，人間の大脳の情報処理メカニズムを解明したり，そのまま再現しようとするものでしたが，最近では，それを離れ，パターン認識や，制御，その他諸々の工学に現われる最適化問題でその解である非線形関数を近似する手段の一つとして位置付けられるようになりました。

無論，大脳の情報処理メカニズムに関わる分野も続けられています。今までの研究について概略ふれておきます。

人間の大脳にはニューロンと呼ばれる 140 億個の神経細胞が存在します。
その人間の大脳の情報処理メカニズムを明らかにし，あるいはそのまま再現しようとする研究が近年, 行われるようになりました。

なぜそうなったかというと，ニューロンが学習という非常に特異な機能もっているからです。 経験などのデータからだけの判断論理は適用範囲を制限されていますが，学習機能は事例データから複雑な判断論理を構築可能にします。
ここで，神経細胞の大まかの構造いついてふれておきます。

通常の細胞と基本的には同じですが，以下の大きな特徴があります。

(1) 樹状突起。これは入力端子の役目をします。

(2) 軸索。これは出力端子の役目をします。

ほかの細胞との接合部はシナプスと呼ばれ神経細胞間の情報伝達が以下のように電気的に行われます。その過程は,

(1) ，シナプス結合を通じて電気パルスが伝達し細胞体の状態が変化することにより情報が伝達されます。

(2) 変化の仕方には 2 種類あり，興奮性のものは細胞体の電気レベルが上昇し，抑制性は細胞体の電気レベルが低下します。

(3) あるしきい値を超えるとパルスを発生して出力側に接続された神経細胞体に刺激を与えます。

近年の工学的な研究では, こういう動作の特性を以下のように数学的に単純化しています。

(1) 空間的加算性

$x_1, \cdots, x_n$ によって樹状突起につながっている他のニューロンからの信号（入力信号）の強さを表すと，細胞体の内部電位 u は入力信号の線形和

$$u = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

によって定まるといわれています。これを空間的加算性といいます。$a_1, \cdots, a_n$ はそれぞれの信号が強められたり弱められたりする度合いを表します。

(2) 時間的加算性

入力信号 x_i は一般に時間の経過とともに変動する時間の関数と考えることができ，内部電位 u は現在までの各時刻における信号の時間的な線形和

$$\int_{-\infty}^{t} v_i(t-\tau)x_i(\tau)d\tau \quad (i = 1, \ldots, n)$$

で定まるものと考えられています。この性質を時間的加算性といいます。$x_i(t)$ は時刻 t での入力信号の大きさを表し，$v_i(t)$ は線形和の結合係数で，単位の強さの入力信号が時間 t 後に内部に貢献する度合いを表します。これはニューロンの中を情報が通過するのに遅延が生じることを表します。

(3) 非線形性

内部電位があるしきい値を超えるとパルスを発生し，軸策によって出力側に接続された神経細胞体に刺激（出力信号）を与えます。その刺激の大きさは入力信号の大きさ，従って内部電位にの強さによらず一定であることが実験によって知られている。即ちニューロンは入力に対して非線形な動作をします。

(4) ２値性ニューロンから出される刺激（出力信号）の大きさは入力信号の大きさによらず一定であるので，巨視的に観れば，ある時刻に信号があるかないかの形（つまりは１か０か）で把握することができます。この意味で２値的です。

以上のように動作の特性は数学的に単純化されるわけですが，その結果，設定したモデルが実物の脳とかなりかけ離れたものになっています。

しかしながら，工学が従来そうであったようにこの「人工的なニューラルネット」は興味深い知見をもたらしてくれるものと期待されています。

また，設定したモデルが生理学的に妥当なら脳に関する有益な知見を提供する可能性もあります。

現在，研究に用いられているニューラルネットの代表的な構成には

(1) 階層型ネットワーク（この教材で扱う）

(2) 相互結合型ネットワーク

がありますが, この資料では，(1) の階層型ネットワークについて解説します。

第2章　パーセプトロン

2.1　パーセプトロン

階層型ニューラルネットワークの研究はパーセプトロンに始まりました。この研究は, ある対象の特性値あらわすためにベクトル化されたデータを, 2種類に分類する方法です。それは「学習」という識別用の関数の逐次近似構成によっています。

2.1.1　パーセプトロンの特徴

受容器と連合器 パーセプトロンは受容器の素子と連合器の素子の対応関係で定義されます。これによって, 受容器のへの入力を連合器の出力に変化させます。

数学モデル

入力と出力の関係から見ると，集合 $\{0,1\}$ の m 重直積集合から $\{-1,0,1\}$ への関数です。

$$\{0,1\}^m \to \{-1,0,1\}$$

論理関数

パーセプトロンは**学習** によって，**任意**の論理関数を構成することができます。以下にその例を説明します。

例

論理関数

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \in \{0,1\}^3 \mapsto z \in \{-1,0,1\}$$

である関数を例に取ります。ただし，z は

$$z = \phi(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 - \theta)$$

です。 ここで，α_i は荷重系数，θ はしきい値です。また ϕ は量子化関数で，次のように表されます。

$$\phi(\mu) = \begin{cases} 0, & \mu < 0 \\ 1, & \mu \geq 0 \end{cases}$$

目標値　この論理関数の入力に対する目標値を以下のよう定めます。

x_1	x_2	x_3	目標値
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

この目標値を実現するパ ーセプトロンの荷重係数 α_i を「学習」という手続きによって調整します。結論から先に述べますとこの問題の解は以下で与えられます。

任意の $\theta > 0$ に対して

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} = \begin{pmatrix} -\theta \\ \theta \\ \theta \end{pmatrix}$$

次に, このような問題に関する解の存在と, 学習と呼ばれる解の逐次構成法について説明します。

2.1.2　解の存在と学習定理

まず，解の存在条件については次の**「線形分離性」**という条件が知られています。

解の存在性：線形分離性

$$\Phi(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \in \mathbf{X}^- \\ 1, & \boldsymbol{x} \in \mathbf{X}^+ \end{cases}$$

なるパーセプトロン

$$z = \Phi(\boldsymbol{x}) = \phi(\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 - \theta)$$

が存在するための必要十分条件は $\boldsymbol{R}^3$ を $\mathbf{X}^-$ と $\mathbf{X}^+$ に分離する平面 P が存在すること。(**線形分離性**)

その平面 $\mathbf{P}$ の方程式は

$$\mathbf{P} = \{(x_1, x_2, x_3) \in \boldsymbol{R}^3 : \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 - \theta = 0\}$$

で定義される。

解の構成法については次の**「学習アルゴリズム」**が知られています。

学習アルゴリズム

START　$\boldsymbol{\alpha} = 0$ とおく。

TEST

$\mathbf{X}^-$ または $\mathbf{X}^+$ に属する $\boldsymbol{x}$ を一つとる。

$\boldsymbol{x} \in \mathbf{X}^-$ ならば $\boldsymbol{x}$ に $-\boldsymbol{x}$ を代入する。

$\boldsymbol{\alpha}^T\boldsymbol{x} - \theta > 0$ ならば **TEST** へ，そうでなければ **ADD** へ

ADD $\boldsymbol{\alpha}$に $\boldsymbol{\alpha} + \boldsymbol{x}$ を代入し，**TEST** へ

さらに，この学習アルゴリズムは$\boldsymbol{\alpha}$の有限回の更新が行われ，解が求められることは次の定理で示されています。

定理 2.1.1 *[学習定理]*
$\mathbf{X}^-, \mathbf{X}^+$ は有限集合とする。問題の解 $\boldsymbol{\alpha}^*$ が存在すれば（論理関数が線形分離可能ならば）上記の学習アルゴリズムは有限回で

$$\begin{aligned} \boldsymbol{\alpha}^T\boldsymbol{x} - \theta &> 0 \quad (\forall \boldsymbol{x} \in \mathbf{X}^+), \\ \boldsymbol{\alpha}^T(-\boldsymbol{x}) - \theta &> 0 \quad (\forall \boldsymbol{x} \in \mathbf{X}^-) \end{aligned}$$

が実現される。

以上によって, 構成したい論理関数が線形分離可能ならば, それを実現するパーセプトロンは, 有限回の学習によって構築できます。以下，この定理の証明を書きます。

証明 先ず,

$$L = \max\{|\boldsymbol{x}|; \boldsymbol{x} \in \mathbf{X}^- \cup \mathbf{X}^+\}, K = |\boldsymbol{\alpha}^*|$$

とおきます。ここで，任意の $\alpha_i \quad (i = 1, 2, 3)$ について

$$\frac{\alpha_1}{|\boldsymbol{\alpha}|} \leq 1$$

ですから

$$\frac{\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}}{|\boldsymbol{\alpha}|} \leq K$$

は常に成り立っています。
ここで$\boldsymbol{\alpha}$を上記の学習強化アルゴリズムで変化する係数ベクトルとし，**ADD** で強化された回数を添え字 n として $\boldsymbol{\alpha}_{n-1}$ から $\boldsymbol{\alpha}_n$ に変化する場合の上式の分子を計算すると

$$\begin{aligned} &\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}_n \\ &= \boldsymbol{\alpha}^{*T}(\boldsymbol{\alpha}_{n-1} + \boldsymbol{x}) \\ &= \boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}_{n-1} + \boldsymbol{\alpha}^{*T}\boldsymbol{x} \\ &> \boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}_{n-1} + \theta \\ &\cdots \\ &> \boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}_0 + n\theta \\ &= n\theta \end{aligned}$$

tonarimasu となります。次に分母を調べと

$$\begin{aligned} &|\boldsymbol{\alpha}_n||\boldsymbol{\alpha}_n| \\ &= (\boldsymbol{\alpha}_{n-1} + \boldsymbol{x})^T(\boldsymbol{\alpha}_{n-1} + \boldsymbol{x}) \\ &= \boldsymbol{\alpha}_{n-1}^T\boldsymbol{\alpha}_{n-1} + 2\boldsymbol{\alpha}_{n-1}^T\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{x}) \\ &\leq \boldsymbol{\alpha}_{n-1}^T\boldsymbol{\alpha}_{n-1} + (2\theta + L^2) \\ &\cdots \\ &\leq \boldsymbol{\alpha}_{n-2}^T\boldsymbol{\alpha}_{n-2} + 2(2\theta + L^2) \end{aligned}$$

$$
\begin{aligned}
&\cdots \\
&\leq \boldsymbol{\alpha}_0^T \boldsymbol{\alpha}_0 + n(2\theta + L^2) \\
&\leq n(2\theta + L^2)
\end{aligned}
$$

です。これから

$$|\boldsymbol{\alpha}_n| \leq \sqrt{n}\sqrt{(2\theta + L^2)}$$

となります。これら分母，分子の評価式から

$$n < K^2\beta^{-2}, \quad \beta = \left(\frac{\theta}{\sqrt{2\theta + L^2}}\right)$$

を得て，n が有限であることが示されます。 □

第3章　階層型ニューラルネットワーク

3.1　階層型回路網とＢＰ法

以下は3層の階層型ニューラルネットワークの例を図示したものです。このニューラルネットワークの第1層 (input layer)、第2層 (hidden layer)、第3層 (output layer) の素子数はそれぞれ n, l, m 個である。第 i 層の j 番の素子から第 $i+1$ 層 k 番の素子への結合係数は $w_{k,j}^{(i+1)}$ で表しています。

このニューラルネットワークは n 次元空間 $\mathbf{R^n}$ の元

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \in \mathbf{R^n}$$

から m 次元空間 $\mathbf{R^m}$ の元

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \in \mathbf{R^m}$$

を対応させる写像を図示しています。

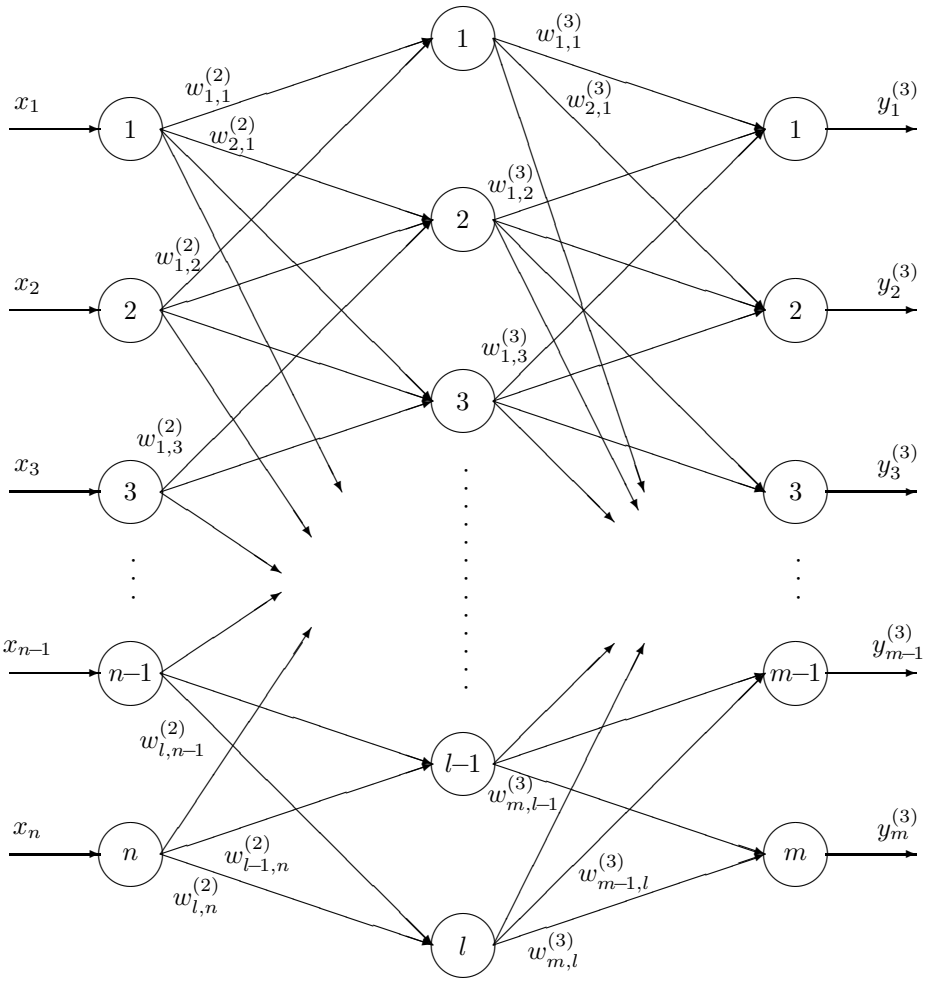

図 1　3 層の階層型ニューラルネットワーク

Fig.1　Three-layer neural network

具体的にどう

$$\mathbf{x} \in \mathbf{R^n} \mapsto \mathbf{y} \in \mathbf{R^n}$$

をどう計算するのか一般の m 層の階層型ニューラルネットワークで説明します。
図 1 の丸印○を細胞と呼びます。

各 $l \quad (l = 1, \ldots, m)$ に対し，第 l 層は n_l 個の細胞 (図中の○印) を持つものとすると，第 l 層第 k 細胞の出力 $y_k^{(l)}$ は

$$\begin{aligned} y_k^{(l)} &= f(\boldsymbol{w}_k^{(l)} \boldsymbol{y}^{(l-1)}) \\ &= f(\sum_{j=1}^{n_{l-1}} w_{kj}^{(l)} y_j^{(l-1)}) \end{aligned}$$

と定めます。
ここで $\boldsymbol{w}_k^{(l)} = (w_{k1}^{(l)}, \ldots, w_{kn_{l-1}}^{(l)})$ は第 $l-1$ 層から第 l 層第 k 細胞への信号の伝播に関する荷重ベクトルとします。f は任意に固定された関数です。　以上の計算を入力層から出力層まで繰り返し計算することにより，出力層（第 m 層）の細胞の出力 $\boldsymbol{y} = \boldsymbol{y}^{(m)}$ を求めることができます。

荷重 $\boldsymbol{w}$ を変化させれば，この階層型ネットワークが望ましい出力を出すように調整できます。

例えば，こうして得られたシステムの出力 $\boldsymbol{y}$ と出力すべき値 $\boldsymbol{z}$ との 2 乗誤差を $\boldsymbol{E}$ として，すなわち出力信号 $\boldsymbol{y}$ に対し

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{y}) &= \frac{1}{2} \|\boldsymbol{z} - \boldsymbol{y}^{(m)}\|^2 \\ &= \frac{1}{2} \sum_{i=1}^{n_m} (y_i^{(m)} - z_i)^2 \end{aligned}$$

で定義して，$\boldsymbol{E}$ の値が小さくなる方向に荷重 $\boldsymbol{w}$ を変化させる方法があります。

修正には，$\mathbf{E}$ が上のネットワークに使われる荷重のベクトル $\mathbf{w}$ について 2 階微分可能ならば,

$$\begin{aligned} &\mathbf{E}(\mathbf{w_0} + \boldsymbol{\Delta}\mathbf{w}) = \mathbf{E}(\mathbf{w_0}) + \left(\frac{\partial \mathbf{E}(\mathbf{w_0})}{\partial \mathbf{w}} \right)^T \boldsymbol{\Delta}\mathbf{w} + o(\boldsymbol{\Delta}\mathbf{w}) \\ &\frac{|o(\boldsymbol{\Delta}\mathbf{w})|}{|\boldsymbol{\Delta}\mathbf{w}|} \to 0 (\boldsymbol{\Delta}\mathbf{w} \to \mathbf{0}) \end{aligned}$$

で，$o(\boldsymbol{\Delta}\mathbf{w})$ が他の項と比較して十分小さくなるよう, 正数 $\varepsilon > 0$ を十分小さくして,

$$\boldsymbol{\Delta}\mathbf{w} = -\varepsilon \frac{\partial \mathbf{E}(\mathbf{w_0})}{\partial \mathbf{w}}$$

として

$$\mathbf{w_1} = \mathbf{w_0} + \boldsymbol{\Delta}\mathbf{w}$$

と荷重のベクトル $\mathbf{w}$ を更新します。このとき

$$\begin{aligned} &\mathbf{E}(\mathbf{w_0}) = \mathbf{E}(\mathbf{w_0} + \boldsymbol{\Delta}\mathbf{w}) \\ &\approx \mathbf{E}(\mathbf{w_0}) - \varepsilon \left(\frac{\partial \mathbf{E}(\mathbf{w_0})}{\partial \mathbf{w}} \right)^T \frac{\partial \mathbf{E}(\mathbf{w_0})}{\partial \mathbf{w}} \end{aligned}$$

から

$$\mathbf{E}(\mathbf{w_1}) \leq \mathbf{E}(\mathbf{w_0})$$

となります。

次節で説明する BP 法は，そのような手法の一つで，$\boldsymbol{E}$ について最急降下法を用いて $\boldsymbol{w}$ を変化させるものです。

$\mathbf{R^n}$ の有界な領域上の任意の連続関数が，このようなニューラルネットワークによって任意の精度で近似できることも知られています。($K.Funahashi$ や $K.Hornik$) これについては，後の章で述べます。

3.1.1 ＢＰ法の逐次修正式

以下に最急降下法を用いる$\boldsymbol{x}$の修正式を示します。

修正式

$$\begin{aligned}\Delta \boldsymbol{w}_i^{(l)} &= -c\frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial \boldsymbol{w}_i^{(l)}} \\ &= -ce_i^{(l)}\boldsymbol{y}^{(l-1)}\end{aligned}$$

として，$\boldsymbol{w}_i$ を変化させます。

ここで，$e_i^{(l)}$ は出力層（$l=m$）については

$$e_i^{(m)} = (y_i^{(m)} - z_i)f'(\boldsymbol{x}_i^{(m)}), \tag{3.1}$$

それ以外の層（$l=1,\ldots,m-1$）については

$$e_i^{(l)} = \sum_{j=1}^{n_{l+1}} e_i^{(l+1)} w_{ji}^{(l+1)} f'(x_i^{(l)}) \quad (l=1,\ldots,m-1) \tag{3.2}$$

で与えられます。

ただし，$x_i^{(l)}$ は第 l 層 i 番目の細胞に対する入力荷重和

$$x_i^{(l)} = \boldsymbol{w}_i^{(l)}\boldsymbol{y}^{(l-1)} \tag{3.3}$$

とします。これよりすぐに

$$\frac{dx_i^{(l)}}{d\boldsymbol{w}_i^{(l)}} = \boldsymbol{y}^{(l-1)} \tag{3.4}$$

が得られます。

3.1.2 逐次修正式の導出

上記の修正式の導出法を説明します。

$$\Delta \boldsymbol{w}_i^{(l)} = -c\frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial \boldsymbol{w}_i^{(l)}}$$

は合成関数の微分によって

$$\frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial \boldsymbol{w}_i^{(l)}} = \frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial x_i^{(l)}} \cdot \frac{dx_i^{(l)}}{d\boldsymbol{w}_i^{(l)}}$$

と与えられます。ここで,

$$e_i^{(l)} = \frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial x_i^{(l)}}$$

とおくと

$$e_i^{(l)} = \frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial y_i^{(l)}} \cdot \frac{dy_i^{(l)}}{dx_i^{(l)}}.$$

次に，$y_i^{(l)}$ は第 l 層 i 番目の細胞の出力であるから

$$y_i^{(l)} = f(x_i^{(l)})$$

が成り立ちます。ゆえに,

$$\frac{dy_i^{(l)}}{dx_i^{(l)}} = f'(x_i^{(l)})$$

となります。

次に，$\frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial y_i^{(l)}}$ について，$l = m$ と $l < m$ の二つの場合に分けて考えます。

- $l = m$, すなわち $\boldsymbol{y}^{(l)}$ が出力$\boldsymbol{y}$である場合：

$$\begin{aligned}
\frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial y_i^{(l)}} &= \frac{1}{2} \cdot \frac{\partial (\sum_{k=1}^{n_m} (y_k^{(m)} - z_k)^2)}{\partial y_i^{(l)}} \\
&= (y_i^{(m)} - z_i), \\
e_i^{(m)} &= \frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial y_i^{(m)}} \cdot \frac{dy_i^{(m)}}{dx_i^{(m)}} \\
&= (y_i^{(m)} - z_i) f'(x_i^{(m)})
\end{aligned}$$

- $l < m$ の場合：

  $l+1$ 層の入力荷重値 $\boldsymbol{w}^{(l+1)}$ と $e_i^{(l+1)}$ がすでに計算済みとすれば

$$\begin{aligned}
\frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial y_i^{(l)}} &= \frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial \boldsymbol{x}^{(l+1)}} \cdot \frac{\partial \boldsymbol{x}^{(l+1)}}{\partial y_i^{(l)}} \\
&= \sum_{j=1}^{n_{l+1}} \frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial x_j^{(l+1)}} \cdot \frac{\partial x_j^{(l+1)}}{\partial y_i^{(l)}} \\
&= \sum_{j=1}^{n_{l+1}} e_j^{(l+1)} w_{ji}^{(l+1)}, \\
e_i^{(l)} &= \frac{\partial \boldsymbol{E}(\boldsymbol{y})}{\partial y_i^{(l)}} \cdot \frac{dy_i^{(l)}}{dx_i^{(l)}} \\
&= \sum_{j=1}^{n_{l+1}} e_j^{(l+1)} w_{ji}^{(l+1)} f'(x_i^{(l)}) \\
&= \sum_{j=1}^{n_{l+1}} e_j^{(l+1)} w_{ji}^{(l+1)} f'(\boldsymbol{w}^{(l)} \boldsymbol{y}^{(l-1)})
\end{aligned}$$

を得ます。
以上をまとめれば

$$\Delta \mathbf{w}_i^{(l)} = -c e_i^{(l)} \boldsymbol{y}^{(l-1)}$$

です。 □

第4章　近似問題と最適化問題

4.1　近似問題

前節では，システムの出力 $\boldsymbol{y}$ と出力すべき値 $\boldsymbol{z}$ との２乗誤差を $\boldsymbol{E}$ として，すなわち出力信号 $\boldsymbol{y}$ に対し

$$\begin{aligned}\boldsymbol{E}(\boldsymbol{y}) &= \frac{1}{2}\|\boldsymbol{z}-\boldsymbol{y}^{(m)}\|^2 \\ &= \frac{1}{2}\sum_{i=1}^{n_m}(y_i^{(m)}-z_i)^2\end{aligned}$$

で定義して，$\boldsymbol{E}$ の値が小さくなる方向に荷重 $\boldsymbol{w}$ を変化させる方法を説明しました。

このように，出力すべき値が予め与えられる場合は，その値を教師値といい，その値との誤差の最小化問題を，教師付き問題といいます。

$K.Funahashi$ や $K.Hornik$ によって連続関数のニューラルネットワークによる一様近似可能性が研究されています。その研究成果によれば，　$C(B_D,\mathbf{R}^m)$ に属する任意の連続関数は l を十分大きくとれば $N(\Delta,K,l)$ の元によって一様近似可能であることがわかっています。証明なしで使いますが，以下の定理が知られています。

定理 4.1.1

任意の $C(B_D,\mathbf{R}^m)$ の元μ　と任意の正数$\varepsilon'>0$ に対してある $l_{\mu,\varepsilon'}$とそれに対応した，
$N(\Delta,K,l_{\mu,\varepsilon'})$ の元ρが存在して
$\max_{x\in B_D}\|\mu(x)-\rho(x)\|<\varepsilon'$

詳しくは以下の文献にあります。
$K.Funahashi : On the Approximate Realization of Continuous Mappings by Neural Networks,$
$Neural Networks, Vol.2, pp.183-192, 1989$
$K.Hornik : Some New Results on Neural Network Approximation,$
$Neural Networks, Vol.6, pp.1069-1072, 1993.$

4.2　ネットワーク集合上の最小化問題

階層型ニューラルネットワークよって表現される関数，これを ρ で表すことにしますが，これに何らかの評価を与える関数

$$\rho \mapsto J(\rho)$$

が与えられた場合，$J(\rho)$ を最小にする ρ は存在するでしょうか？　この節ではこの問題を調べます。

結論から言えば，特に，結合係数としきい値の絶対値，シグモイド関数の最大勾配に制約がある階層型ネットワークから ρ が造られるのであれば，その制約の範囲内で $J(\rho)$ を最小にする ρ は存在することは比較的簡単に判ります。

これは，よく知られた，以下の定理の応用です。

定理 4.2.1　X を $\mathbf{R^n}$ の有界閉集合，J を $\mathbf{R^n}$ 上の連続関数とする。このとき， ρ は X で最大値，最小値をとる。すなわち，$\rho(X)$ は最大値，および最小値をもつ。

ρ は階層型ネットワークが表現する関数で，$\mathbf{R^n}$ の元ではありませんが，同様な議論を行うことができます。

図 1 に表されるニューラルネットワークについて、特にそれらの結合係数としきい値の絶対値が $K > 0$ 以下で、最大勾配が $\Delta > 0$ 以下のシグモイド関数を各素子共通にもち、第 2 層の素子が l 個の階層型ニューラルネットワーク集合を $N(\Delta, K, l)$ とします。

これに使うシグモイド関数は，例えば

$$\psi(x) = \frac{1}{1 + e^{-x}}$$

とおき、 $\phi_\delta(x) = \psi(\delta x)$ で定義される ϕ_δ を用います。

このとき

$$|\phi_\delta(x') - \phi_\delta(x)| \leq |\delta x' - \delta x| = \delta |x' - x| \leq \Delta |x' - x|$$

です。$N(\Delta, K, l)$ は次式で定義されます。

$$\begin{aligned}
N(\Delta, K, l) \stackrel{\triangle}{=} \{ & \rho : x \in B_D \mapsto \rho(x) \in \mathbf{R}^m; \\
& |w^{(2)}_{k,j}|, |\theta^{(2)}_k| \leq K, \quad (1 \leq j \leq n, 1 \leq k \leq l), \\
& |w^{(3)}_{k,j}| \leq K, \quad (1 \leq j \leq l, 1 \leq k \leq m), \\
& 0 \leq \delta \leq \Delta, \\
& \rho(x) = y^{(3)}; \\
& \quad y^{(3)}_k = w^{(3)}_k \cdot y^{(2)}, \quad (1 \leq k \leq m), \\
& \quad y^{(2)}_j = \phi_\delta(w^{(2)}_j \cdot x - \theta^{(2)}_j), \quad (1 \leq j \leq l) \ \}
\end{aligned} \tag{4.1}$$

ただし $w^{(2)}_k$ 【$w^{(3)}_k$】 は第 1 層 【第 2 層】 の各素子から第 2 層 【第 3 層】k 番の素子への結合係数のベクトル

$$\left.\begin{aligned}
w^{(2)}_k &= (w^{(2)}_{k,1}, w^{(2)}_{k,2}, \cdots, w^{(2)}_{k,n})^T \\
&\quad (1 \leq k \leq l) \\
w^{(3)}_k &= (w^{(3)}_{k,1}, w^{(3)}_{k,2}, \cdots, w^{(3)}_{k,l})^T \\
&\quad (1 \leq k \leq m)
\end{aligned}\right\} \tag{4.2}$$

を表し、$\theta^{(2)}_k$ は第 2 層 k 番の素子のしきい値を表す。

$$\begin{aligned}
y^{(1)} &= x \\
y^{(2)} &= (y^{(2)}_1, y^{(2)}_2, \cdots, y^{(2)}_l)^T \\
y^{(3)} &= (y^{(3)}_1, y^{(3)}_2, \cdots, y^{(3)}_m)^T
\end{aligned}$$

は各層の出力ベクトルを表しています。

4.2.1 コンパクト性

この $N(\Delta, K, l)$ が $\mathbf{R^n}$ の有界閉集合と同様な性質をもっていることを示します。
$J(\rho)$ はこの $N(\Delta, K, l)$ から $\mathbf{R}$ への写像

$$\rho \in N(\Delta, K, l) \mapsto J(\rho) \in \mathbf{R}$$

として扱います。J には連像性が要求されます。先ず，$N(\Delta, K, l)$ には距離が定義されます。

$$B_D \subseteq \{x \in \mathbf{R}^n; \|x \mid \leq D\} \tag{4.3}$$

から $\mathbf{R^m}$ への連続関数全体の集合 $C(B_D, \mathbf{R}^m)$ の部分集合ですが，$C(B_D, \mathbf{R}^m)$ には

$$d(g, h) = \max_{x \in B_D} \|f(x) - g(x)\|$$

が定義されています。ここで，$\|\cdots\|$ は $\mathbf{R^m}$ のベクトル

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_m \end{pmatrix} c \in \mathbf{R^m}$$

に対して

$$\|\mathbf{y}\| = \sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}$$

で定義されます。

上記の距離 $d(g, h)$ は $\mathbf{R^m}$ の距離同様，以下の性質を充たしています。

任意の $f, g \in C(B_D, \mathbf{R}^m)$ に対して $d(f, g) \geq 0$
任意の $f, g \in C(B_D, \mathbf{R}^m)$ に対して $d(f, g) = 0 \Leftrightarrow f = g$
任意の $f, g, h \in C(B_D, \mathbf{R}^m)$ に対して $d(f, h) \leq d(f, g) + d(g, h)$

このような距離が定義される集合を距離空間と言います。

X, Y が距離 d_X, d_Y が定義されている距離空間とします。

定義 4.2.1 $x_0 \in X$ とし、r を正の数とするとき、

$$S_X(x_0, r) \stackrel{\triangle}{=} \{x : d_X(x, x_0) < r,\ x \in X\} \tag{4.4}$$

で定義される X の部分集合 $S_r(x_0)$ を中心が x_0 で半径が r の開球といいます。

無論，距離空間 Y でも，開球が定義されます。

定義 4.2.2 $U \subset X$ として，$\forall x \in U$ に対して、x を中心とする十分小さい開球 $S_X(x, r)$ をとり，$S_X(x, r) \subset U$ となるようにできるとき、U を X の開集合であるといいます。

定義 4.2.3 $f : X \to Y$ とします。f が $x_0 \in X$ で連続であるとは、

$f(x_0)$ を中心とする任意の半径 ε の開球 $S_Y(f(x_0), \varepsilon)$ に対し、x_0 を中心とする半径 δ の開球 $S_X(x_0, \delta)$ が存在し、

$$f(S_X(x_0, \delta)) \subset S_Y(f(x_0), \varepsilon) \tag{4.5}$$

が成り立つことを言います。

定理 4.2.2 $f : X \to Y$ とします。

$$f \text{ が } X \text{ で連続} \Leftrightarrow [V \text{ が } Y \text{ の開集合ならば、} f^{-1}(V) \text{ が } X \text{ の開集合}]$$

(証明) $(\Rightarrow)$ f が X で連続で、V を Y の開集合と仮定して、$f^{-1}(V)$ が X の開集合となることを証明します。$f^{-1}(V) = \phi$ なら，これは開集合です。そこで、$f^{-1}(V) \neq \phi$ とします。

$x \in f^{-1}(V)$ とすると、 $f(x) \in V$ である。V は開集合であるから、 $f(x)$ を中心とする開球

$$S_Y(f(x), \varepsilon) \subset V$$

が存在します。 f は連続ですから，開球 $S_X(x, \delta)$ が存在して,

$$f(S_X(x, \delta)) \subset S_Y(f(x), \varepsilon)$$

$$\therefore \quad f(S_X(x, \delta)) \subset V$$
$$\therefore \quad S_X(x, \delta) \subset f^{-1}(V)$$

以上より、$f^{-1}(V)$ は開集合です。

$(\Leftarrow)$ V が Y の任意の開集合ならば、常に $f^{-1}(V)$ が X の開集合であると仮定します。

$\forall x \in X$ とし, $f(x)$ を中心とする任意の開球 $S_Y(f(x), \varepsilon)$ をとります。$S_Y(f(x), \varepsilon)$ は開集合ですから, 仮定よりその逆像 U は開集合で x を含む。よって, 定義により $S_\delta(x) \subset U$ が存在します。

$$\therefore \quad S_X(x, \delta) \subset U = f^{-1}(S_Y(f(x), \varepsilon))$$
$$\therefore \quad f(S_X(x, \delta)) \subset (S_Y(f(x), \varepsilon))$$

よって、f は点 x で連続になります。$x \in X$ は任意でしたから, f は X で連続です。□

定義 4.2.4 X を距離空間とします。X の開集合の集合 $\mathcal{F}$, これを開集合族と言いますが, が X の部分集合 $X' \subseteq X$ を覆っているとき, すなわち,

$$X' = \subseteq \bigcup_{U \in calF} U \tag{4.6}$$

となっているとき, $\mathcal{F}$ を X' の開被覆といいます。

定義 4.2.5 距離空間 X の部分集合 X' に任意の開被覆 $\mathcal{F}$ が与えられたとき、それから適当な有限個 $U_1, U_2, \cdots, U_n \in \mathcal{F}$ を選んで

$$X' \subseteq \bigcup_{i=1}^{n} U_i \tag{4.7}$$

とできるとき,X' はコンパクト集合であるといいます。

コンパクト集合の例としては,$\mathbf{R}$ の閉区間 $[a,b]$ またそれの 2 重直積

$$[a,b]\times[a,b]\stackrel{\triangle}{=}\{\mathbf{x}=(x_1,x_2)|x_1,x_2\in[a,b]\}$$

n 重直積

$$[a,b]^n\stackrel{\triangle}{=}\{\mathbf{x}=(x_1,x_2,\cdots,x_n)|x_1,x_2,\cdots,x_n\in[a,b]\}$$

$\mathbf{R^n}$ の閉球の閉球

$$S_{R^n}(x_0,r)\stackrel{\triangle}{=}\{x:\|x-x_0\|\le r,\,x\in R^n\} \tag{4.8}$$

があります。

さて, 連続写像とコンパクト集合の関係については以下が知られています。

定理 4.2.3 X,Y が距離空間とする。
$X'\subseteq X$ がコンパクトで写像 $f:X\to Y$ が連続ならば、 X' の像 $f(X)$ はコンパクトである。

証明
X,Y が距離 d_X,d_Y が定義されている距離空間とします。
$\mathcal{G}$ を $f(X')$ の開被覆とすると，開集合の連続写像による逆像は開集合ですので,

$$\mathcal{F}=\{f^{-1}(V),V\in\mathcal{G}\}$$

は X' の開被覆です。X' はコンパクトでしたからこれから有限個

$$U_1,U_2,\cdots,U_n\in\mathcal{F}$$

を選んで, X' を覆うことができます。すなわち,

$$X'\subset\bigcup_{i=1}^{n}U_i$$

そうすると、

$$\begin{aligned}f(X') &\subset f(\bigcup_{i=1}^{n}U_i)\\ &= \bigcup_{i=1}^{n}f(U_i)\\ &= \bigcup_{i=1}^{n}V_i\end{aligned}$$

となり、

$$V_1,V_2,\cdots,V_n\in\mathcal{G}$$

が $f(X')$ の有限個の部分開被覆となって、$f(X')$ がコンパクト集合であることがわかります。□

また，以下の定理も知られています。

定理 4.2.4 X を距離空間とする。
$X'\subseteq X$ がコンパクトならば，$\{x_n\}$ を X' の任意の無限列とするとき, $\{x_n\}$ の無限部分列 $\{x_{n_k}\}$ と $x_\infty\in X'$ が存在して x_{n_k} は x_∞ に収束する。

上の二つの定理を用いて,

定理 4.2.5 X を距離空間とする。
$X' \subseteq X$ がコンパクトで写像 $f : X \to \mathbf{R}$ が連続ならば、 X' 上での像 f は最大・最小値をもつ。

証明 まず，X' の像 $f(X')$ は X' がコンパクトで、f が連続写像なので，$\mathbf{R}$ のコンパクト集合です。

そこで，$f(X')$ の上限，下限

$$\sup f(X'), \inf f(X')$$

をとると，それぞれ，$f(X')$ の元の無限点列

$$\{x_n\}, \{y_m\}$$

が存在して,

$$x_n \to \sup f(X')(n \to \infty)$$

$$y_m \to \inf f(X')(m \to \infty)$$

です。ここで，$f(X')$ はコンパクト集合でしたから,

$$\{x_n\}, \{y_m\}$$

の部分点列

$$\{x_{n_k}\}, \{y_{m_l}\}$$

と，$f(X')$ の元

$$x_\infty, y_\infty \in f(X')$$

が存在して,

$$x_{n_k} \to x_\infty(k \to \infty)$$

$$y_{m_l} \to x_\infty(l \to \infty)$$

です。実数列とその部分列の極限は一致するので,

$$\sup f(X') = x_\infty \in f(X'), \inf f(X') = x_\infty \in f(X')$$

で，結局,

$$\max f(X') = \sup f(X'), \min f(X') = \inf f(X') \ \square$$

$N(\Delta, K, l)$ のコンパクト性について述べてます。

定理 4.2.6 $N(\Delta, K, l)$ は有界閉集合 $B_D \subseteq \mathbf{R}^n$ 上の連続関数全体 $C(B_D, \mathbf{R}^m)$ の部分集合として, コンパクトである。

(証明)

結合係数のベクトル $w_k^{(2)}, w_k^{(3)}$ と、しきい値 $\theta_k^{(2)}$ 及びシグモイド関数の最大勾配 δ による多重対

$$\begin{aligned}(w, \theta, \delta) = (&w_1^{(2)}, w_2^{(2)}, \cdots, w_l^{(2)}, \\ &w_1^{(3)}, w_2^{(3)}, \cdots, w_m^{(3)}, \theta_1^{(2)}, \theta_2^{(2)}, \cdots, \theta_l^{(2)}, \delta)\end{aligned} \tag{4.9}$$

から $N(\Delta,K,l)$ の元 ρ への対応が $\mathbf{R^n}$の コンパクト集合 $[-K,K]^{2l+m}\times[0,\Delta]$ から $N(\Delta,K,l)$ の上への写像

$$\Omega:[-K,K]^{2l+m}\times[0,\Delta]\to N(\Delta,K,l) \tag{4.10}$$

を定義していることが判ります。
コンパクト集合の連続写像による像はコンパクト集合であるから、この全写 Ω が連続写像であることを示せばよいわけです。そこで $[-K,K]^{2l+m}\times[0,\Delta]$ の別の元として結合係数 $v_{k,j}^{(i)}$ としきい値 $o_k^{(i)}$ 及びシグモイド関数の最大勾配 δ' による多重対 (v,o,δ') をとり、これに Ω により $g\in N(\Delta,K,l)$ が対応するものとする。すなわち、(4.1), (4.2) 式と同様に

$$\left.\begin{array}{ll} v_k^{(2)}=(v_{k,1}^{(2)},v_{k,2}^{(2)},\cdots,v_{k,n}^{(2)})^T & (1\le k\le l)\\ v_k^{(3)}=(v_{k,1}^{(3)},v_{k,2}^{(3)},\cdots,v_{k,l}^{(3)})^T & (1\le k\le m)\\ z^{(2)}=(z_1^{(2)},z_2^{(2)},\cdots,z_l^{(2)})^T & \\ g(x)=(z_1^{(3)},z_2^{(3)},\cdots,z_m^{(3)})^T & \\ z_k^{(3)}=v_k^{(3)}\cdot z^{(2)} & (1\le k\le m)\\ z_j^{(2)}=\phi_{\delta'}(v_j^{(2)}\cdot x-o_j^{(2)}) & (1\le j\le l) \end{array}\right\} \tag{4.11}$$

とします。ここで $|\rho(x)-g(x)|$ を評価します。$x\in B_D$ について $|x|\le D$ が成り立つから (4.1), (4.11) 式により

$$\begin{aligned} &|y_k^{(2)}-z_k^{(2)}|\\ &\quad\le\Delta|(w_k^{(2)}-v_k^{(2)})\cdot x-(\theta_k^{(2)}-o_k^{(2)})|\\ &\qquad+|\delta-\delta'||v_k^{(2)}\cdot x-o_k^{(2)}|\\ &\quad\le\Delta(|w-v|D+|\theta-o|)+|\delta-\delta'|K(D+1) \end{aligned} \tag{4.12}$$

です。また ψ は有界でしたから

$$|y^{(2)}|\le U_\psi \tag{4.13}$$

ただし

$$U_\psi\stackrel{\triangle}{=}\sup\{|\phi_\delta(u)|;0\le\delta\le\Delta,u\in R\}<\infty \tag{4.14}$$

これから、

$$\begin{aligned} &|y_k^{(3)}-z_k^{(3)}|\\ &\quad\le|(w_k^{(3)}-v_k^{(3)})\cdot y^{(2)}|+|v_k^{(3)}\cdot(y^{(2)}-z^{(2)})|\\ &\quad\le|w-v|U_\psi+\Delta K(|w-v|D+|\theta-o|)\\ &\qquad+K^2|\delta-\delta'|(D+1) \end{aligned} \tag{4.15}$$

を得ます。結局、任意の $x\in B_D$ について

$$\begin{aligned} |\rho(x)-g(x)|&=|y^{(3)}-z^{(3)}|\\ &\le\varepsilon(|w-v|,|\theta-o|,|\delta-\delta'|) \end{aligned}$$

ただし、

$$\begin{aligned} &\varepsilon(|w-v|,|\theta-o|,|\delta-\delta'|)\\ &\quad\stackrel{\triangle}{=}|w-v|U_\psi+\Delta K(|w-v|D+|\theta-o|)\\ &\qquad+K^2|\delta-\delta'|(D+1) \end{aligned} \tag{4.16}$$

を得ます。従って

$$\begin{aligned}|\rho - g| &\stackrel{\triangle}{=} \sup\{|\rho(x) - g(x)|, x \in B_D\} \\ &\leq \varepsilon(|w - v|, |\theta - o|, |\delta - \delta'|)\end{aligned} \tag{4.17}$$

が成り立ちます。

$$\begin{aligned}\varepsilon(|w - v|, |\theta - o|, |\delta - \delta'|) \to 0 \\ (|w - v|, |\theta - o|, |\delta - \delta'| \to 0)\end{aligned}$$

であるから写像 Ω は連続写像です。すなわち $N(\Delta, K, l) = \Omega([-K, K]^{2l+m} \times [0, \Delta])$ はコンパクト集合です。□

この命題によって、シグモイド関数の最大勾配と結合係数、しきい値がそれぞれある正数　$\Delta, K > 0$ 以下の階層型ネットワークのクラスを表す関数集合 $N(\Delta, K, l)$ は $C(B_D, \mathbf{R}^m)$ 内でそのコンパクト性が保証されます。

$\Delta, K > 0$ を十分大きくとればよく用いれている殆どの階層型ネットワークがこの集合に含まれると考えてよいでしょう。

ここで $C(B_D, \mathbf{R}^m)$ で連続関数 J

$$\rho \in C(B_D, \mathbf{R}^m) \mapsto J(\rho)$$

についての最小化問題が解を持つとします。

前節の定理 4.1.1 によれば,

> 任意の $C(B_D, \mathbf{R}^m)$ の元μ　と任意の正数$\varepsilon' > 0$ に対してある $l_{\mu,\varepsilon'}$とそれに対応した,
> $N(\Delta, K, l_{\mu,\varepsilon'})$ の元ρが存在して
> $\max_{x \in B_D} \|\mu(x) - \rho(x)\| < \varepsilon'$

が成立っています。そこで，J の $C(B_D, \mathbf{R}^m)$ 上での連続性から

> 任意の $C(B_D, \mathbf{R}^m)$ の元μ　と任意の正数$\varepsilon > 0$ に対してある $l_{\mu,\varepsilon}$とそれに対応した,
> $N(\Delta, K, l_{\mu,\varepsilon})$ の元ρが存在して
> $J(\rho) - \varepsilon < J(\mu)$

ところで各 $N(\Delta, K, l)$ には最小化元が存在するので、結局

$$\inf_l \min_{\rho \in N(\Delta,K,l)} J(\rho) \leq \inf_{\mu \in C(B_D, \mathbf{R}^m)} J(\mu)$$

一方 $N(\Delta, K, l) \subseteq C(B_D, \mathbf{R}^m)$ ゆえ

$$\min_{\rho \in N(\Delta,K,l)} J(\rho) \geq \inf_{\mu \in C(B_D, \mathbf{R}^m)} J(\mu)$$

です。したがって

$$\inf_l \min_{\rho \in N(\Delta,K,l)} J(\rho) = \inf_{\mu \in C(B_D, \mathbf{R}^m)} J(\mu)$$

とります。

ところで、$l_1 \leq l_2$ について

$$N(\Delta, K, l_1) \subseteq N(\Delta, K, l_2)$$

なので

$$\min_{\rho\in N(\Delta,K,l_1)} J(\rho) \geq \min_{\rho\in N(\Delta,K,l_2)} J(\rho)$$

が得られ，したがって

$$\lim_{l\to\infty} \min_{\rho\in N(\Delta,K,l)} J(\rho) = \inf_{\mu\in C(B_D,\mathbf{R}^m)} J(\mu)$$

これは中間層の素子数 l を十分大きくとり $N(\Delta,K,l)$ 内の最小化元を求めれば

$$\inf_{\mu\in C(B_D,\mathbf{R}^m)} J(\mu)$$

の近似が得られることを示しています。