授業資料の例

大前佑斗

私の専門分野の授業例を貼り付けたものです。
てきとうに貼っているので、読んでも理解不能ですが、
なにか雰囲気的なものを感じ取ってください。

機械学習の根底を、狭く、深く、
なるべく謎を残さずに、徹底的に学びます。

授業資料の例

機械学習基礎論:
深層学習（ディープラーニング）

なぜAIは犬と猫を
識別できるのか？

授業資料「深層学習」の一部 (1)

8.3.1

$\boldsymbol{x} \in \mathbb{R}^2, \boldsymbol{y} \in \mathbb{R}^2, \boldsymbol{z} \in \mathbb{R}^2, \boldsymbol{W} \in \mathbb{R}^{2\times 2}, \boldsymbol{b} \in \mathbb{R}^2$ における深さ 2 の合成ベクトル値関数

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{z} = \boldsymbol{\delta}(\boldsymbol{y}) = \begin{bmatrix} \delta(y_1) \\ \delta(y_2) \end{bmatrix}, \ \delta(y) = 1/(1+\mathrm{e}^{-y}) \tag{44}$$

を考えます（問題 7.2.1 の関数と同じ）。このヤコビ行列は、

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \boldsymbol{W}^\top \mathrm{diag}(\boldsymbol{\delta}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \odot (1_2 - \boldsymbol{\delta}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}))) \tag{45}$$

となることを示してください。

8.3.2

（問題 7.2.2 の拡張。すべての軸方向を考える問題）上のヤコビ行列を用いて、地点 $\boldsymbol{x} = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$ において、x_1, x_2 軸方向に対して、z_1, z_2 は、それぞれ局所的にみて右上がりか、右下がりか、調べてください。ただし、$\boldsymbol{W} = \begin{bmatrix} 1_2 & 1_2 \end{bmatrix}$ で、$\boldsymbol{b} = -21_2$ とします。

ヒント：というより答えですが、線形代数で紹介した知識を総動員すると、

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}} = \boldsymbol{W}^\top \mathrm{diag}(\boldsymbol{\delta}(\boldsymbol{y}) \odot (1_2 - \boldsymbol{\delta}(\boldsymbol{y}))) = \boldsymbol{W}^\top \mathrm{diag}(\boldsymbol{\delta}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \odot (1_2 - \boldsymbol{\delta}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})))$$

$$= \begin{bmatrix} 1_2 & 1_2 \end{bmatrix}^\top \mathrm{diag}(\boldsymbol{\delta}(\begin{bmatrix} 1_2 & 1_2 \end{bmatrix} 1_2 - 21_2) \odot (1_2 - \boldsymbol{\delta}(\begin{bmatrix} 1_2 & 1_2 \end{bmatrix} 1_2 - 21_2))) \tag{46}$$

$$= \begin{bmatrix} 1_2 & 1_2 \end{bmatrix}^\top \mathrm{diag}(\boldsymbol{\delta}(\begin{bmatrix} 1_2 & 1_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 21_2) \odot (1_2 - \boldsymbol{\delta}(\begin{bmatrix} 1_2 & 1_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} - 21_2))) \tag{47}$$

授業資料「深層学習」の一部 (2)

9　合成ベクトル値関数（深さ＝深層）

9.1　解説

D_0 次元の入力変数 $\boldsymbol{x}_0 \in \mathbb{R}^{D_0}$ に対して、

$$
\begin{aligned}
\boldsymbol{x}_1 &= \boldsymbol{f}_1(\boldsymbol{x}_0),\ \boldsymbol{x}_1 \in \mathbb{R}^{D_1} \\
\boldsymbol{x}_2 &= \boldsymbol{f}_2(\boldsymbol{x}_1),\ \boldsymbol{x}_2 \in \mathbb{R}^{D_2} \\
\boldsymbol{x}_3 &= \boldsymbol{f}_3(\boldsymbol{x}_2),\ \boldsymbol{x}_3 \in \mathbb{R}^{D_3} \\
&\vdots \\
\boldsymbol{x}_n &= \boldsymbol{f}_n(\boldsymbol{x}_{n-1}),\ \boldsymbol{x}_n \in \mathbb{R}^{D_n} \\
&\vdots \\
\boldsymbol{x}_N &= \boldsymbol{f}_N(\boldsymbol{x}_{N-1}),\ \boldsymbol{x}_N \in \mathbb{R}^{D_N}
\end{aligned}
\tag{61}
$$

を、深さ N の合成ベクトル値関数といいます。$\boldsymbol{x}_0 \to \boldsymbol{x}_1 \to \boldsymbol{x}_2 \cdots$ と次々に変換され、$\boldsymbol{x}_N$ を出力する関数です。N が十分に大きいとき、深層化されているため、**深層学習**の抽象表現に相当します。入力変数 $\boldsymbol{x}_0$ に対する出力変数 $\boldsymbol{x}_N$ の偏導関数は非常に距離が遠くて不安になりますが、ヤコビ行列およびヤコビ連鎖律を用いて、

$$
\frac{\partial \boldsymbol{x}_N}{\partial \boldsymbol{x}_0} = \frac{\partial \boldsymbol{x}_1}{\partial \boldsymbol{x}_0}\frac{\partial \boldsymbol{x}_2}{\partial \boldsymbol{x}_1}\frac{\partial \boldsymbol{x}_3}{\partial \boldsymbol{x}_2}\cdots\frac{\partial \boldsymbol{x}_n}{\partial \boldsymbol{x}_{n-1}}\cdots\frac{\partial \boldsymbol{x}_N}{\partial \boldsymbol{x}_{N-1}} = \prod_{n=1}^{N}\frac{\partial \boldsymbol{x}_n}{\partial \boldsymbol{x}_{n-1}} \tag{62}
$$

となります。深さ 2 のヤコビ連鎖律を抽象化し、ずらっと並べただけです。詳細は次回の授業でやりますが、この式が、**人工知能の頭が良くなる基本式**になります。

授業資料「深層学習」の一部 (4)

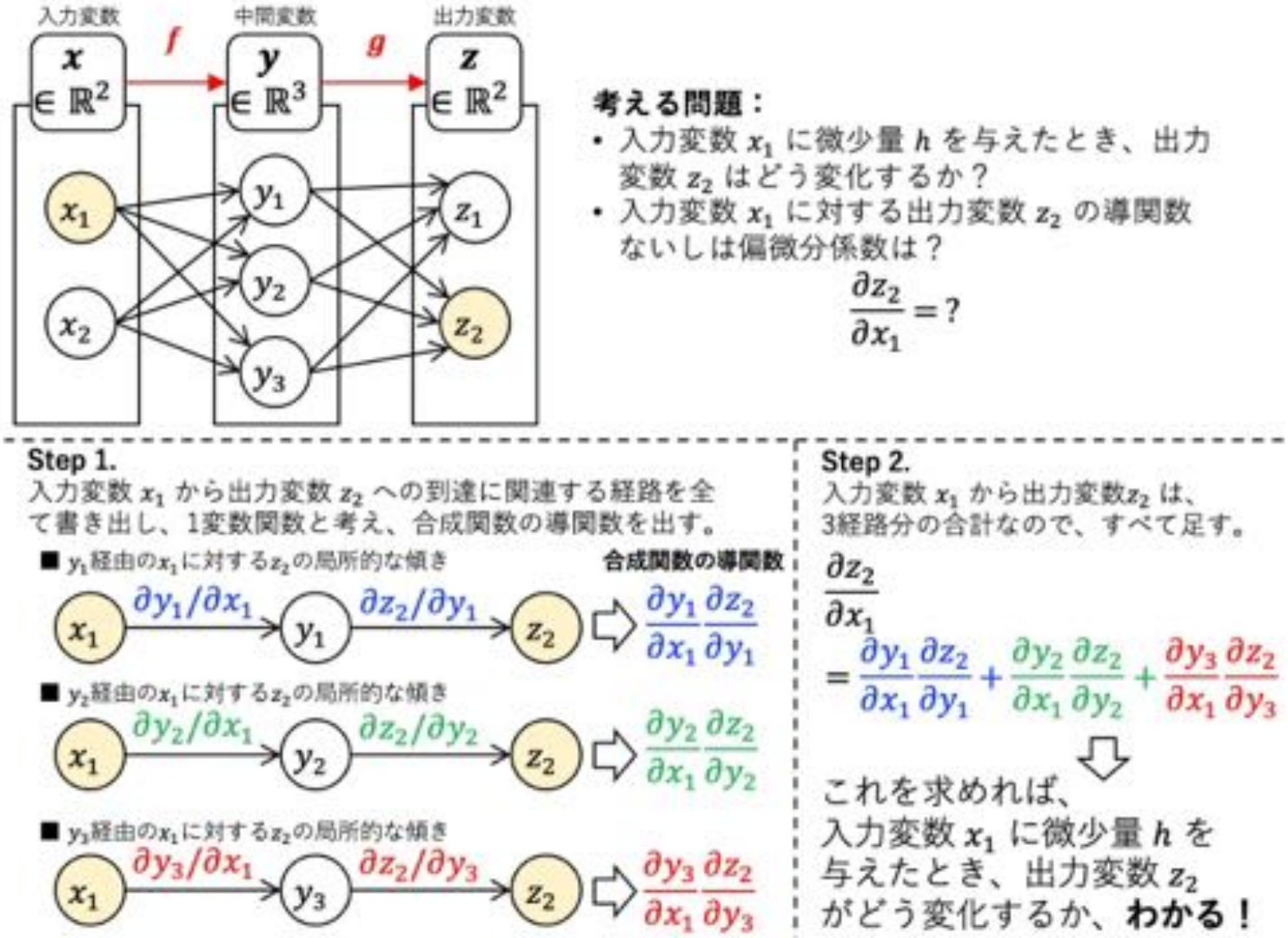


Figure 4: 合成ベクトル値関数の偏導関数

授業資料「深層学習」の一部 (5)

6.2 解説: 尤度関数

今、塩・米・価格を色々と変更しながら、コンビニで売ってみて、売れ行きが良いか悪いかを観測したとします。これを合計 I パターン試したとすると、NN への入力変数 $\boldsymbol{x}$ と正解 q（売れ行き良い/悪い）は、合計 I 個存在することになります。これを、

$$(\boldsymbol{x}_1, q_1),\ (\boldsymbol{x}_2, q_2),\ \cdots,\ (\boldsymbol{x}_I, q_I) \tag{25}$$

と表現し、これを**教師データ**とか**訓練データ**といいます[8]。$\boldsymbol{x}_i$ と q_i は、i パターン目の塩・米・価格と売れた売れないの正解ラベルです。このとき、望ましい NN とは、

- $\boldsymbol{x}_1$ を NN に与えたときに正解 q_1 が発生し、かつ、
- $\boldsymbol{x}_2$ を NN に与えたときに正解 q_2 が発生し、かつ、
- $\boldsymbol{x}_3$ を NN に与えたときに正解 q_3 が発生し、かつ、
- ⋮以下続く⋮
- $\boldsymbol{x}_I$ を NN に与えたときに正解 q_I が発生するようなモデル

となります。個別の事象の発生確率は、$P(q_1|\boldsymbol{x}_1)\cdots P(q_I|\boldsymbol{x}_I)$ で表現でき、これらが「かつ」でつながっていますから、個別事象の同時発生を考える必要があります。個別事象が同時に発生する確率を**同時確率**と呼び、個別の確率の積として得ることができます[9]。すると、良い NN とは、総乗記号を用いて、

$$\prod_{i=1}^{I} P(q_i|\boldsymbol{x}_i) = \prod_{i=1}^{I} (p_i)^{q_i}(1-p_i)^{1-q_i} \tag{26}$$

が高いモデルということができます。これを**尤度（ゆうど）関数**ないしは likelihood と呼びます。すなわち、

- 「良い NN を作る」とは、$\prod_{i=1}^{I}(p_i)^{q_i}(1-p_i)^{1-q_i}$ が高いような $\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{V}, c$ を発見することです

授業資料「深層学習」の一部 (5)

となります。$\frac{\partial L_i}{\partial p_i}$ は、クロスエントロピーの導関数であり、導関数の連鎖律・対数関数の導関数・合成関数の導関数の３つを用いれば、

$$\begin{aligned}\frac{\partial L_i}{\partial p_i} &= -\frac{\partial (q_i \log p_i)}{\partial p_i} - \frac{\partial (1-q_i)\log(1-p_i)}{\partial p_i} = -\frac{q_i}{p_i} + \frac{1-q_i}{1-p_i} \\ &= -\frac{q_i(1-p_i)}{p_i(1-p_i)} + \frac{p_i(1-q_i)}{p_i(1-p_i)} = \frac{p_i - q_i}{p_i(1-p_i)} \end{aligned} \tag{51}$$

と表現されます。これにより、必要な導関数はすべて求められたので、行列積により計算すれば、

$$\begin{aligned}\frac{\partial L_i}{\partial (\boldsymbol{W})_{:m}} &= \frac{\partial \boldsymbol{y}_i}{\partial (\boldsymbol{W})_{:m}} \frac{\partial \boldsymbol{r}_i}{\partial \boldsymbol{y}_i} \frac{\partial z_i}{\partial \boldsymbol{r}_i} \frac{\partial p_i}{\partial z_i} \frac{\partial L_i}{\partial p_i} \\ &= x_{im} \boldsymbol{I}_N \mathrm{diag}((\boldsymbol{1}_N - \boldsymbol{r}_i) \odot \boldsymbol{r}_i) \boldsymbol{V}^\top p_i(1-p_i) \frac{p_i - q_i}{p_i(1-p_i)} \\ &= x_{im}(p_i - q_i) \mathrm{diag}((\boldsymbol{1}_N - \boldsymbol{r}_i) \odot \boldsymbol{r}_i) \boldsymbol{V}^\top \end{aligned} \tag{52}$$

が得られます。そのため、元々知りたかった導関数は、

$$\begin{aligned}\frac{\partial L_i}{\partial \boldsymbol{W}} &= \begin{bmatrix} \frac{\partial L_i}{\partial (\boldsymbol{W})_{:1}} & \cdots & \frac{\partial L_i}{\partial (\boldsymbol{W})_{:M}} \end{bmatrix} \\ &= (p_i - q_i) \begin{bmatrix} x_{i1} \mathrm{diag}((\boldsymbol{1}_N - \boldsymbol{r}_i) \odot \boldsymbol{r}_i) \boldsymbol{V}^\top & \cdots & x_{iM} \mathrm{diag}((\boldsymbol{1}_N - \boldsymbol{r}_i) \odot \boldsymbol{r}_i) \boldsymbol{V}^\top \end{bmatrix} \\ &= (p_i - q_i) \mathrm{diag}((\boldsymbol{1}_N - \boldsymbol{r}_i) \odot \boldsymbol{r}_i) \boldsymbol{V}^\top \begin{bmatrix} x_{i1} & \cdots & x_{iM} \end{bmatrix} \\ &= (p_i - q_i) \mathrm{diag}((\boldsymbol{1}_N - \boldsymbol{r}_i) \odot \boldsymbol{r}_i) \boldsymbol{V}^\top (\boldsymbol{x}_i)^\top \end{aligned} \tag{53}$$

となります。$\mathrm{diag}((\boldsymbol{1}_N - \boldsymbol{r}_i) \odot \boldsymbol{r}_i) \boldsymbol{V}^\top$ を行列の中から外に出した理由は、なんとなくではなくて、行列積の定義に基づくものであることに注意してください [**証明終わり**]。**この式を使えば、クロスエントロピーが低い $\boldsymbol{W}$ を発見できるようになります！**

授業資料「深層学習」の一部 (6)

7.6.1

（問題5.2.1がベース）2つのおむすびを作って実際に売ってみました。1つ目は塩・コメ・価格が $\boldsymbol{x}_1 = \begin{bmatrix}0.5 & 0.5 & 0.0\end{bmatrix}^\top$ のときで、$q_1 = 1$ でした（実際に作ってみたら、たくさん売れました）。2つ目は塩・コメ・価格が $\boldsymbol{x}_2 = \begin{bmatrix}1.0 & 0.1 & 1.0\end{bmatrix}^\top$ のときで、$q_2 = 0$ でした（実際に作ってみたら、あまり売れませんでした）。また、シグモイド関数を活性化関数とするニューラルネットワーク

$$
\begin{aligned}
&\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}, \\
&\boldsymbol{r} = \boldsymbol{\delta}(\boldsymbol{y}), \\
&z = \boldsymbol{V}\boldsymbol{r} + c, \\
&p = \delta(z), \\
&\boldsymbol{x} \in \mathbb{R}^3, \boldsymbol{y} \in \mathbb{R}^2, \boldsymbol{r} \in \mathbb{R}^2, z \in \mathbb{R}, \boldsymbol{W} \in \mathbb{R}^{2\times 3}, \boldsymbol{b} \in \mathbb{R}^2, \boldsymbol{V} \in \mathbb{R}^{1\times 2}, c \in \mathbb{R}
\end{aligned}
\tag{66}
$$

があり、この内部パラメータは

$$
\boldsymbol{W} = \begin{bmatrix}1 & 1 & 1 \\ 1 & 1 & 1\end{bmatrix}, \boldsymbol{b} = \begin{bmatrix}1 \\ 1\end{bmatrix}, \boldsymbol{V} = \begin{bmatrix}1 & 1\end{bmatrix}, c = 1 \tag{67}
$$

であるとします。問題5.2.1の回答より、このNNのクロスエントロピーは $L = 3.06$ のようです。$\boldsymbol{b}$ をどの方向に変化させれば、クロスエントロピーの低減を期待することができるでしょうか？問題5.2.1の回答より、

$$
p_1 = 0.94,\ p_2 = 0.95,\ \boldsymbol{r}_1 = \begin{bmatrix}0.88 \\ 0.88\end{bmatrix},\ \boldsymbol{r}_2 = \begin{bmatrix}0.96 \\ 0.96\end{bmatrix} \tag{68}
$$

を用いて良いことにします。

授業資料「深層学習」の一部 (7)

5　畳み込みニューラルネットワーク

入力画像が大きくても、畳み込みやプーリング操作を繰り返せば、そのサイズはぐんぐん落ちていきます。落とした画像に対して、平坦化を施すことで、小さな次元のベクトルを得ることができ、NN に入力することが可能となります。これを実現する手法を**畳み込みニューラルネットワーク**と呼びます。例えば、

$$
\begin{aligned}
\boldsymbol{H}_1 &= \mathrm{Conv}(\boldsymbol{H}_0, \boldsymbol{F}_0), \quad \text{(入力行列 } \boldsymbol{H}_0 \text{ を畳み込んで圧縮行列 } \boldsymbol{H}_1 \text{ に)} \\
\boldsymbol{H}_2 &= \mathrm{Pool}(\boldsymbol{H}_1, M_0 \times N_0), \quad \text{(圧縮行列 } \boldsymbol{H}_1 \text{ をプーリングにより圧縮行列 } \boldsymbol{H}_2 \text{ に)} \\
\boldsymbol{H}_3 &= \mathrm{Conv}(\boldsymbol{H}_2, \boldsymbol{F}_1), \quad \text{(圧縮行列 } \boldsymbol{H}_2 \text{ を畳み込んで圧縮行列 } \boldsymbol{H}_3 \text{ に)} \\
\boldsymbol{H}_4 &= \mathrm{Pool}(\boldsymbol{H}_3, M_1 \times N_1), \quad \text{(圧縮行列 } \boldsymbol{H}_3 \text{ をプーリングにより圧縮行列 } \boldsymbol{H}_4 \text{ に)} \\
\boldsymbol{x} &= \mathrm{Flat}(\boldsymbol{H}_4), \quad \text{(圧縮行列 } \boldsymbol{H}_4 \text{ を平坦化によりベクトル } \boldsymbol{x} \text{ に)} \\
\boldsymbol{y} &= \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}, \quad \text{(入力ベクトル } \boldsymbol{x} \text{ をアフィン写像により } \boldsymbol{y} \text{ に)} \\
\boldsymbol{r} &= \boldsymbol{\delta}(\boldsymbol{y}), \quad \text{(} \boldsymbol{y} \text{ をシグモイド関数により } \boldsymbol{r} \text{ に)} \\
z &= \boldsymbol{V}\boldsymbol{r} + c, \quad \text{(ベクトル } \boldsymbol{r} \text{ をアフィン写像により } z \text{ に)} \\
p &= \delta(z), \quad \text{(推定確率 } p \text{ に)} \\
\boldsymbol{x} &\in \mathbb{R}^M, \boldsymbol{y} \in \mathbb{R}^N, \boldsymbol{r} \in \mathbb{R}^N, z \in \mathbb{R}, \boldsymbol{W} \in \mathbb{R}^{N \times M}, \boldsymbol{b} \in \mathbb{R}^N, \boldsymbol{V} \in \mathbb{R}^{1 \times N}, c \in \mathbb{R} \qquad (4)
\end{aligned}
$$

は、畳み込みニューラルネットワークの一例です。これは、畳み込みとプーリングをそれぞれ2回行い、画像を圧縮してから、平坦化し、シグモイド関数を持つ NN により何らかの推定確率を得るモデルです。例えば、

授業資料の例

機械学習基礎論:
ガウス過程とベイズ最適化

AIが新材料を発見する
とはなんなのか？

授業資料「GPBO」の一部 (1)

3.6　制約付き勾配法によるカーネルパラメータの探索

ここでは、カーネルパラメータ $\boldsymbol{\theta}$ の設定方法について言及する。なお、ノイズが考慮されたカーネル行列 $\boldsymbol{K}'$ により議論を展開していくが、ノイズを考慮しないカーネル行列 $\boldsymbol{K}$ の場合も同一の議論が成立するため、片方のみを記載する。

良いパラメータは何かを考えた場合、それは、**観測値 $\boldsymbol{y}$ の生起確率を最大化できる $\boldsymbol{\theta}$** である。カーネル行列は $\boldsymbol{\theta}$ の関数であるため、$\boldsymbol{K}'(\boldsymbol{\theta})$ と書くことにすると、$p(\boldsymbol{y})$ は共分散行列（カーネル行列）$\boldsymbol{K}'(\boldsymbol{\theta})$ を与えることで計算可能となることから、これまで言及してきた $p(\boldsymbol{y})$ を、$p(\boldsymbol{y}|\boldsymbol{K}'(\boldsymbol{\theta}))$ の条件付き確率として表現する。こうすると、

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{K}'(\boldsymbol{\theta})) &= \mathcal{N}(0, \boldsymbol{K}'(\boldsymbol{\theta})) \\ &\propto \frac{1}{\sqrt{|\boldsymbol{K}'(\boldsymbol{\theta})|}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}'(\boldsymbol{\theta})^{-1}\boldsymbol{y}\right) \end{aligned} \tag{39}$$

となる。さらに、両辺の対数を取れば、

$$\begin{aligned} \log p(\boldsymbol{y}|\boldsymbol{K}'(\boldsymbol{\theta})) &= \log\left(\frac{1}{\sqrt{|\boldsymbol{K}'(\boldsymbol{\theta})|}} \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}'(\boldsymbol{\theta})^{-1}\boldsymbol{y}\right)\right) \\ &= -\frac{1}{2}\log|\boldsymbol{K}'(\boldsymbol{\theta})| - \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}'(\boldsymbol{\theta})^{-1}\boldsymbol{y} \\ &\propto -\log|\boldsymbol{K}'(\boldsymbol{\theta})| - \boldsymbol{y}^\top \boldsymbol{K}'(\boldsymbol{\theta})^{-1}\boldsymbol{y} \end{aligned} \tag{40}$$

となる。すなわち、目的関数 L と最適化問題は、

$$\begin{aligned} L(\boldsymbol{\theta}) &= L_1(\boldsymbol{\theta}) + L_2(\boldsymbol{\theta}), \\ L_1(\boldsymbol{\theta}) &= -\log|\boldsymbol{K}'(\boldsymbol{\theta})| \\ L_2(\boldsymbol{\theta}) &= -\boldsymbol{y}^\top \boldsymbol{K}'(\boldsymbol{\theta})^{-1}\boldsymbol{y} \\ \boldsymbol{\theta}^* &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ L(\boldsymbol{\theta}) \end{aligned} \tag{41}$$

$\boldsymbol{\theta}^*$ =

=

により、採択すべきカーネルパラ
が解きやすいため、対数を付与し
　続いて、これをどのように解
対象を書き下すと、

授業資料「GPBO」の一部 (2)

Eq.45 右辺第 2 項目が分かればパラメータの近似解を採用できることになるため、これを計算する。微分は線形性[26]を有するため、Eq.41 より、

$$\frac{\partial L(\exp(\boldsymbol{\theta}'))}{\partial \theta'_i} = \frac{\partial L_1(\exp(\boldsymbol{\theta}'))}{\partial \theta'_i} + \frac{\partial L_2(\exp(\boldsymbol{\theta}'))}{\partial \theta'_i} \tag{46}$$

となる。各項の偏微分は逆行列および対数行列式の微分の公式[27]を用いると、それぞれ、

$$\frac{\partial L_1(\exp(\boldsymbol{\theta}'))}{\partial \theta'_i} = -\mathrm{tr}\left(\boldsymbol{K}'(\exp(\boldsymbol{\theta}'))^{-1}\frac{\partial \boldsymbol{K}'(\exp(\boldsymbol{\theta}'))}{\partial \theta'_i}\right), \tag{47}$$

$$\frac{\partial L_2(\exp(\boldsymbol{\theta}'))}{\partial \theta'_i} = \boldsymbol{y}^\top \boldsymbol{K}'(\exp(\boldsymbol{\theta}'))^{-1}\frac{\partial \boldsymbol{K}'(\exp(\boldsymbol{\theta}'))}{\partial \theta'_i}\boldsymbol{K}'(\exp(\boldsymbol{\theta}'))^{-1}\boldsymbol{y} \tag{48}$$

となる。すなわち勾配法を適用するには、$\boldsymbol{K}'(\exp(\boldsymbol{\theta}'))$ を θ'_i で偏微分した関数を明らかにすれば良い。$\boldsymbol{K}'$ は（観測ノイズ付き）カーネル関数を要素とする行列であったため、カーネル関数の微分を知ることができれば良い、ということになる。Eq.37 で定義されたノイズ付きカーネル関数に $\boldsymbol{\theta} = \exp(\boldsymbol{\theta}')$ を代入すると、

$$k'(\boldsymbol{x}_i, \boldsymbol{x}_j; \exp(\boldsymbol{\theta}')) = \exp(\theta'_1)\exp\left(-\frac{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2}{\exp(\theta'_2)}\right) + \exp(\theta'_3)\delta(i,j) \tag{49}$$

となる。これを θ'_1 で偏微分すると、

$$\frac{\partial k'(\boldsymbol{x}_i, \boldsymbol{x}_j; \exp(\boldsymbol{\theta}'))}{\partial \theta'_1} = \exp(\theta'_1)\exp\left(-\frac{|\boldsymbol{x}_i - \boldsymbol{x}_j|^2}{\exp(\theta'_2)}\right) \tag{50}$$

授業資料「GPBO」の一部 (3)

が成立する。すなわち、Eq.58 で定義されたブロック要素の対角成分は、転置により同一との関係がある。

これらを踏まえた上で、$p(y_*|\boldsymbol{y}) \propto p(\boldsymbol{y}, y_*) = p(\boldsymbol{y}')$ について、固定値である $\boldsymbol{y}$ を削除するように（y_* のみが残るように）解いていくと、

$$
\begin{aligned}
p(\boldsymbol{y}') &= \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}'^{*}) \\
&\propto \exp\left(-1/2 \begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\lambda}_{21}^\top \\ \boldsymbol{\lambda}_{21} & \lambda_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix}\right) \\
&= \exp\left(-1/2 \left[\boldsymbol{y}^\top y_*\right] \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\lambda}_{21}^\top \\ \boldsymbol{\lambda}_{21} & \lambda_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix}\right) \\
&= \exp\left(-1/2(\boldsymbol{y}^\top \boldsymbol{\Lambda}_{11} \boldsymbol{y} + y_* \boldsymbol{\lambda}_{21} \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{\lambda}_{21}^\top y_* + y_* \lambda_{22} y_*)\right) \\
&= \exp\left(-1/2(y_* \boldsymbol{\lambda}_{21} \boldsymbol{y} + \boldsymbol{y}^\top \boldsymbol{\lambda}_{21}^\top y_* + y_* \lambda_{22} y_*)\right) \exp\left(-1/2(\boldsymbol{y}^\top \boldsymbol{\Lambda}_{11} \boldsymbol{y})\right) \\
&\propto \exp\left(-1/2(\lambda_{22} y_*^2 + 2\boldsymbol{\lambda}_{21} \boldsymbol{y} y_*)\right) \\
&= \exp\left(-\frac{\lambda_{22}}{2}(y_*^2 + 2\boldsymbol{\lambda}_{21} \boldsymbol{y} \lambda_{22}^{-1} y_*)\right) \\
&= \exp\left(-\frac{1}{2\lambda_{22}^{-1}}\left(y_*^2 + 2\boldsymbol{\lambda}_{21} \boldsymbol{y} \lambda_{22}^{-1} y_* + (\boldsymbol{\lambda}_{21} \boldsymbol{y} \lambda_{22}^{-1})^2 - (\boldsymbol{\lambda}_{21} \boldsymbol{y} \lambda_{22}^{-1})^2\right)\right) \\
&= \exp\left(-\frac{1}{2\lambda_{22}^{-1}}\left((y_* - \boldsymbol{\lambda}_{21} \boldsymbol{y} \lambda_{22}^{-1})^2 - (\boldsymbol{\lambda}_{21} \boldsymbol{y} \lambda_{22}^{-1})^2\right)\right) \\
&= \exp\left(-\frac{(y_* - \boldsymbol{\lambda}_{21} \boldsymbol{y} \lambda_{22}^{-1})^2}{2\lambda_{22}^{-1}}\right) \exp\left(\frac{(\boldsymbol{\lambda}_{21} \boldsymbol{y} \lambda_{22}^{-1})^2}{2\lambda_{22}^{-1}}\right) \\
&\propto \exp\left(-\frac{(y_* - \boldsymbol{\lambda}_{21} \boldsymbol{y} \lambda_{22}^{-1})^2}{2\lambda_{22}^{-1}}\right) \qquad (65)
\end{aligned}
$$

授業資料「GPBO」の一部 (4)

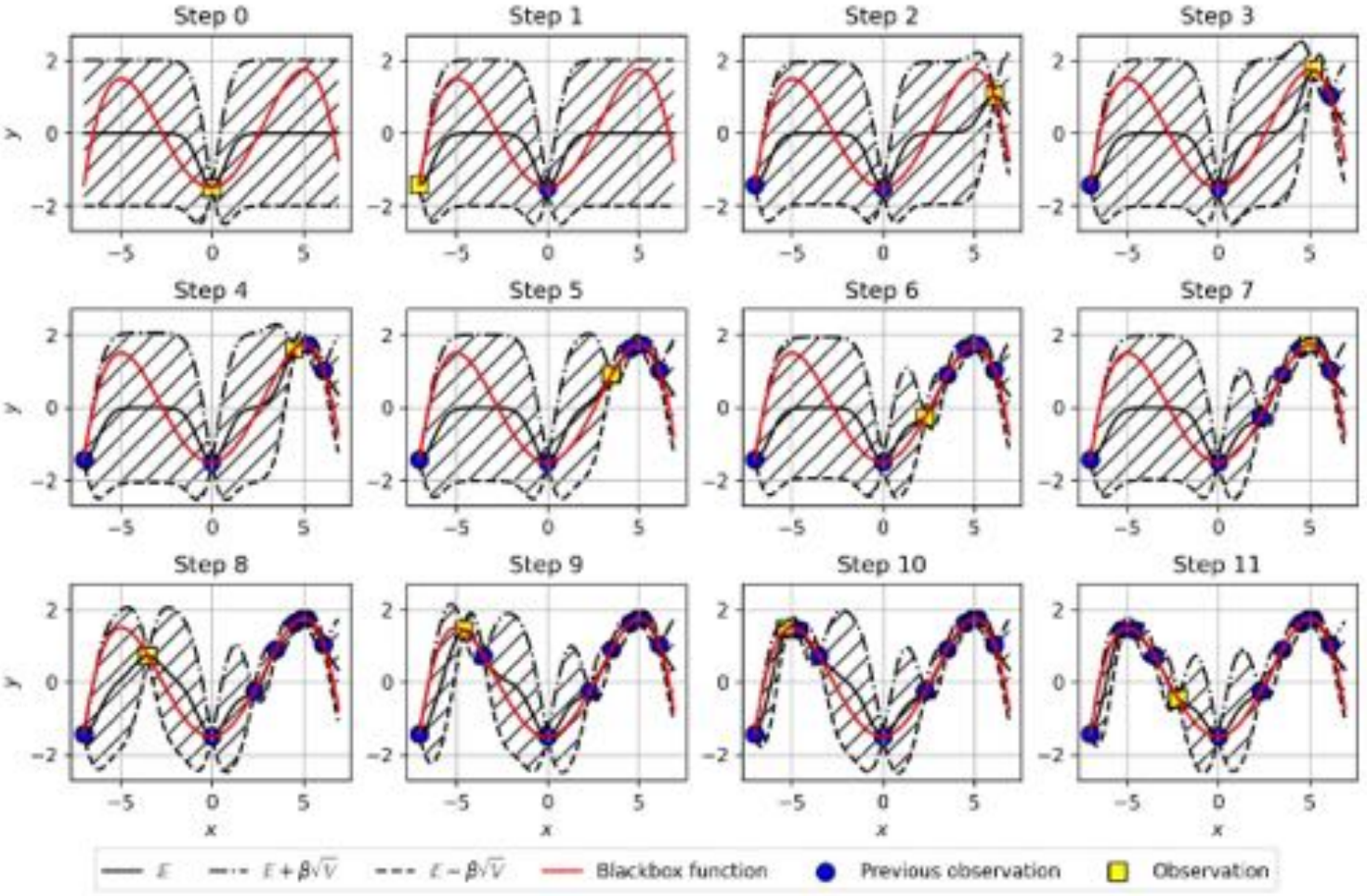


Figure 3: ベイズ最適化の実行結果（入力が 1 次元の場合）。赤線がブラックボックス関数の概形、黒線がガウス過程回帰の平均値 $\mathbb{E}[y_*|\boldsymbol{y}]$、破線がそれに $\beta\sqrt{\mathbb{V}[y_*|\boldsymbol{y}]}$ を加算/減算したもの（すなわち、上側の破線が信頼性上限関数、下側の破線が信頼性下限関数）、青マーカがこれまでの観測、黄マーカがそのステップでの観測である。なお今回は最大値の探索であるため、信頼性下限関数は使用しない。

授業資料の例

機械学習基礎論:
木構造パルツェン推定

AIが新材料を発見する
とはなんなのか？

授業資料「TPE」の一部 (1)

パルツェン推定によるベイズ最適化

大前佑斗（日本大学生産工学部）

2023.05.11

1　問題設定

ガウス過程回帰を利用せず、ノンパラメトリックな確率分布を利用してベイズ最適化を行うアプローチがある（木構造パルツェン推定などとも呼ばれる）。ガウス過程回帰ほど複雑ではないため、比較的容易に理解できるものと思われる。

今、D 次元のベクトル $\boldsymbol{x} = [x_1 \ \cdots \ x_D]^\top$ およびそれをなんらかのブラックボックス関数に入力することで得られる出力 y があるものとする。そして、具体的な入力 $\boldsymbol{x}_n$ に対する観測値 y_n が N データ得られている状況を考え、これを格納したデータセットを

$$\mathcal{D} = \{(\boldsymbol{x}_n, y_n) | n = 1, \cdots, N\} \tag{1}$$

今回はブラックボックス関数の最大化問題を対象としているため、D 次元の入力空間上において、大きな y が発生しやすく、小さな y が発生しにくい入力値 $\boldsymbol{x}$ を次に調べると良いだろう。これは、上述した確率密度関数を用いて、

$$\boldsymbol{x}_{N+1} = \underset{\boldsymbol{x} \in \Psi}{\operatorname{argmax}} \frac{p(\boldsymbol{x}|\mathcal{D}_{\boldsymbol{x}}^+)}{p(\boldsymbol{x}|\mathcal{D}_{\boldsymbol{x}}^-)} \tag{7}$$

と定式化できる。ここで、Ψ は探索空間である。その後、$N+1$ 回目の入力 $\boldsymbol{x}_{N+1}$ の観測値 y_{N+1} を得るための実験などを行う。観測値 y_{N+1} は大きな値であることが期待されるが、最大値とは限らない。そのためこの手続きを T 回実施し、$y_{N+1}, \cdots, y_{N+T}$ を得て、

授業資料「TPE」の一部 (2)

2.1 連続型変数

はじめに、$\boldsymbol{x}$ が D 次元の連続型変数により表現される場合について述べる。このため、まずは $\mathcal{D}_{\boldsymbol{x}}^{+}$ に含まれる n 番目のベクトル $\boldsymbol{x}_n^{+}$ と入力 $\boldsymbol{x}$ の類似度を測定する。同一次元のベクトル同士の類似度というと、ユークリッド距離を採択すれば良いのではと思うかもしれないが、最終的に確率密度関数へとつなげていく必要があるため、別の方法を採択する。

はじめに $\boldsymbol{x}_n^{+}$ と $\boldsymbol{x}$ について、個々の次元の類似度を測定する。具体的には、$\boldsymbol{x}$ と $\boldsymbol{x}_n^{+}$ の d 次元目の値をそれぞれ $x_d, x_{n,d}^{+}$ とし、その類似度を

$$k(x_d; x_{n,d}^{+}, h) = \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{(x_d - x_{n,d}^{+})^2}{2h^2}\right), \ h > 0 \tag{10}$$

2.2 カテゴリカル変数

(執筆中)

観測データ $x_{n,d}$ がカテゴリカル値の場合に使用する関数は、

$$k_{\mathrm{c}}(x_d; x_{n,d}, h_{\mathrm{c}}) = \begin{cases} 1 - h_{\mathrm{c}}, & x_d = x_{n,d} \\ h_{\mathrm{c}}/(N_{\mathrm{c}} - 1), & \text{otherwise} \end{cases}, \ h_{\mathrm{c}} \in [0, 1) \tag{26}$$

であり、これを Aitchison-Aitken カーネルと呼ぶ。ここで、N_{c} は x が取り得る値の数である。特に h_{c} が 0 のとき、x が観測値 $x_{n,d}$ と等しい場合に 1 を取り、そうではないときに 0 を取る関数となる。そのため、h_{c} を大きくすることで、値が分配される構造を有している。

授業資料「TPE」の一部 (3)

は正規分布と同一形状であることから、

$$\int_{-\infty}^{\infty} k(x_d; x_{n,d}^+, h) dx_d = 1 \tag{16}$$

が成立する。これを利用すれば、$p(\boldsymbol{x}|\mathcal{D}_{\boldsymbol{x}}^+)$ の定積分値が 1 であることを示すことができる。入力次元が $D = 2$ の場合についてこれを示すと、

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(\boldsymbol{x}; \boldsymbol{x}_n^+, h) d\boldsymbol{x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k(x_1; x_{n,1}^+, h) k(x_2; x_{n,2}^+, h) dx_1 dx_2 \tag{17}$$

$$= \int_{-\infty}^{\infty} k(x_2; x_{n,2}^+, h) \left(\int_{-\infty}^{\infty} k(x_1; x_{n,1}^+, h) dx_1 \right) dx_2 \tag{18}$$

$$= \int_{-\infty}^{\infty} k(x_2; x_{n,2}^+, h) dx_2 \tag{19}$$

$$= 1 \tag{20}$$

となる[3]。この計算過程を見ればわかるように、次元数がいくつであっても積分の結果は 1 となる。したがって、Eq.15 の積分は、

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\boldsymbol{x}|\mathcal{D}_{\boldsymbol{x}}^+) d\boldsymbol{x} = \frac{1}{N^+} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{n=1}^{N^+} K(\boldsymbol{x}; \boldsymbol{x}_n^+, h) d\boldsymbol{x} \tag{21}$$

$$= \frac{1}{N^+} \sum_{n=1}^{N^+} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} K(\boldsymbol{x}; \boldsymbol{x}_n^+, h) d\boldsymbol{x} \tag{22}$$

$$= \frac{1}{N^+} \sum_{n=1}^{N^+} 1 \tag{23}$$

$$= \frac{1}{N^+} \times N^+ \tag{24}$$

$$= 1 \tag{25}$$

そして研究へ…
(ここから進学希望者向け)

以降は私の研究であり、
学生はここまでは多分届きません。

でも、大学院に行くなら、
難問に挑んでみたいという気持ちが大事

なので、参考程度に載せておきます。

最近やっている研究：

人類がまだ到達していない機械学習について考える

（ハイレベルな研究者には敵いませんが...）

深層学習がうまくいく理由を明らかにする。

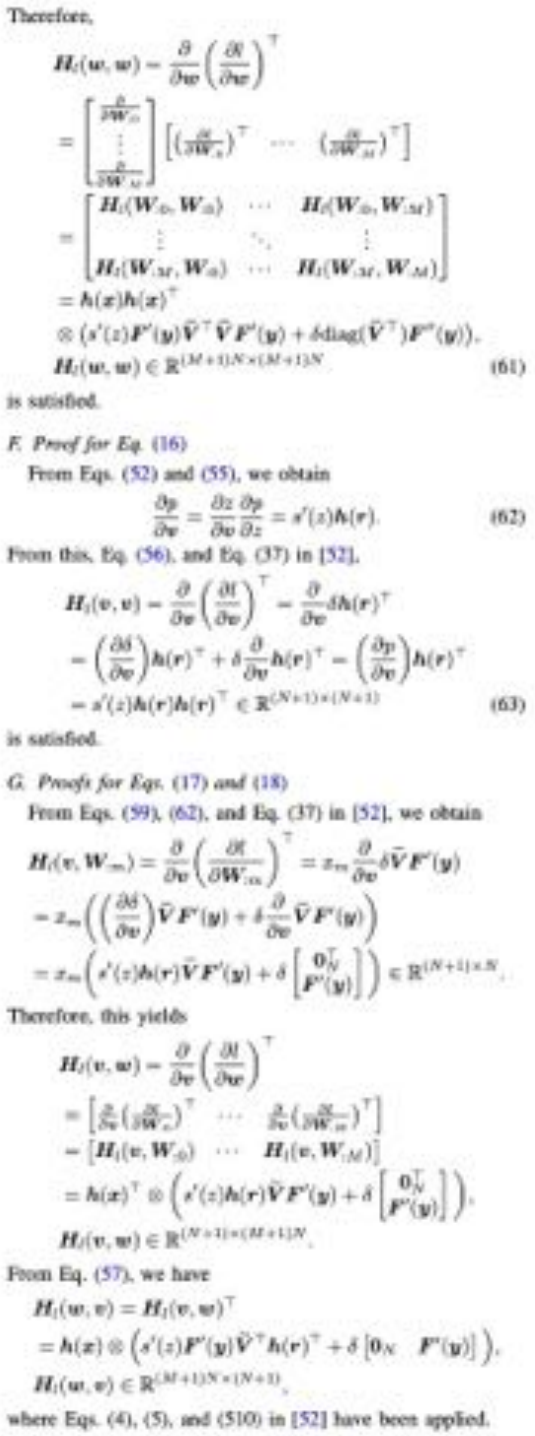
Y. Omae, Wolkowicz-Styan Upper Bound on the Hessian Eigenspectrum for Cross-Entropy Loss in Nonlinear Smooth Neural Networks

非線形平滑ニューラルネットワークにおける交差エントロピー損失のヘッセ行列固有スペクトルに関するウォルコヴィッツ-スティアン上界

詳細:

https://arxiv.org/abs/2604.10202

https://arxiv.org/pdf/2604.10202



Therefore,

$$\boldsymbol{H}_l(\boldsymbol{w},\boldsymbol{w}) = \frac{\partial}{\partial \boldsymbol{w}}\left(\frac{\partial l}{\partial \boldsymbol{w}}\right)^\top = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{W}_\square} \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{W}_M} \end{bmatrix} \left[\left(\frac{\partial l}{\partial \boldsymbol{W}_\square}\right)^\top \cdots \left(\frac{\partial l}{\partial \boldsymbol{W}_M}\right)^\top \right] = \begin{bmatrix} \boldsymbol{H}_l(\boldsymbol{W}_\square,\boldsymbol{W}_\square) & \cdots & \boldsymbol{H}_l(\boldsymbol{W}_\square,\boldsymbol{W}_M) \\ \vdots & \ddots & \vdots \\ \boldsymbol{H}_l(\boldsymbol{W}_M,\boldsymbol{W}_\square) & \cdots & \boldsymbol{H}_l(\boldsymbol{W}_M,\boldsymbol{W}_M) \end{bmatrix}$$
$$= \boldsymbol{h}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^\top \otimes (s'(z)\boldsymbol{F}'(\boldsymbol{y})\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}}\boldsymbol{F}'(\boldsymbol{y}) + \delta\mathrm{diag}(\widetilde{\boldsymbol{V}}^\top)\boldsymbol{F}''(\boldsymbol{y})),$$
$$\boldsymbol{H}_l(\boldsymbol{w},\boldsymbol{w}) \in \mathbb{R}^{(M+1)N\times(M+1)N} \quad (61)$$

is satisfied.

F. Proof for Eq. (16)

From Eqs. (52) and (55), we obtain

$$\frac{\partial p}{\partial \boldsymbol{v}} = \frac{\partial z}{\partial \boldsymbol{v}}\frac{\partial p}{\partial z} = s'(z)\boldsymbol{h}(\boldsymbol{r}). \quad (62)$$

From this, Eq. (56), and Eq. (37) in [52],

$$\boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{v}) = \frac{\partial}{\partial \boldsymbol{v}}\left(\frac{\partial l}{\partial \boldsymbol{v}}\right)^\top = \frac{\partial}{\partial \boldsymbol{v}}\delta\boldsymbol{h}(\boldsymbol{r})^\top - \left(\frac{\partial \delta}{\partial \boldsymbol{v}}\right)\boldsymbol{h}(\boldsymbol{r})^\top + \delta\frac{\partial}{\partial \boldsymbol{v}}\boldsymbol{h}(\boldsymbol{r})^\top - \left(\frac{\partial p}{\partial \boldsymbol{v}}\right)\boldsymbol{h}(\boldsymbol{r})^\top$$
$$= s'(z)\boldsymbol{h}(\boldsymbol{r})\boldsymbol{h}(\boldsymbol{r})^\top \in \mathbb{R}^{(N+1)\times(N+1)} \quad (63)$$

is satisfied.

G. Proofs for Eqs. (17) and (18)

From Eqs. (59), (62), and Eq. (37) in [52], we obtain

$$\boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{W}_m) = \frac{\partial}{\partial \boldsymbol{v}}\left(\frac{\partial l}{\partial \boldsymbol{W}_m}\right)^\top = x_m\frac{\partial}{\partial \boldsymbol{v}}\delta\widetilde{\boldsymbol{V}}\boldsymbol{F}'(\boldsymbol{y})$$
$$= x_m\left(\left(\frac{\partial \delta}{\partial \boldsymbol{v}}\right)\widetilde{\boldsymbol{V}}\boldsymbol{F}'(\boldsymbol{y}) + \delta\frac{\partial}{\partial \boldsymbol{v}}\widetilde{\boldsymbol{V}}\boldsymbol{F}'(\boldsymbol{y})\right)$$
$$= x_m\left(s'(z)\boldsymbol{h}(\boldsymbol{r})\widetilde{\boldsymbol{V}}\boldsymbol{F}'(\boldsymbol{y}) + \delta\begin{bmatrix}\boldsymbol{0}_N^\top \\ \boldsymbol{F}''(\boldsymbol{y})\end{bmatrix}\right) \in \mathbb{R}^{(N+1)\times N}.$$

Therefore, this yields

$$\boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{w}) = \frac{\partial}{\partial \boldsymbol{v}}\left(\frac{\partial l}{\partial \boldsymbol{w}}\right)^\top = \begin{bmatrix}\frac{\partial}{\partial \boldsymbol{v}}\left(\frac{\partial l}{\partial \boldsymbol{W}_\square}\right)^\top & \cdots & \frac{\partial}{\partial \boldsymbol{v}}\left(\frac{\partial l}{\partial \boldsymbol{W}_M}\right)^\top\end{bmatrix} = \begin{bmatrix}\boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{W}_\square) & \cdots & \boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{W}_M)\end{bmatrix}$$
$$= \boldsymbol{h}(\boldsymbol{x})^\top \otimes \left(s'(z)\boldsymbol{h}(\boldsymbol{r})\widetilde{\boldsymbol{V}}\boldsymbol{F}'(\boldsymbol{y}) + \delta\begin{bmatrix}\boldsymbol{0}_N^\top \\ \boldsymbol{F}''(\boldsymbol{y})\end{bmatrix}\right),$$
$$\boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{w}) \in \mathbb{R}^{(N+1)\times(M+1)N}.$$

From Eq. (57), we have

$$\boldsymbol{H}_l(\boldsymbol{w},\boldsymbol{v}) = \boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{w})^\top = \boldsymbol{h}(\boldsymbol{x}) \otimes \left(s'(z)\boldsymbol{F}'(\boldsymbol{y})\widetilde{\boldsymbol{V}}^\top\boldsymbol{h}(\boldsymbol{r})^\top + \delta\begin{bmatrix}\boldsymbol{0}_N & \boldsymbol{F}''(\boldsymbol{y})\end{bmatrix}\right),$$
$$\boldsymbol{H}_l(\boldsymbol{w},\boldsymbol{v}) \in \mathbb{R}^{(M+1)N\times(N+1)},$$

where Eqs. (4), (5), and (510) in [52] have been applied.

H. Proof for Eqs. (20) and (21)

Based on Eq. (63) and Eq. (14) in [52], we obtain

$$\mathrm{tr}(\boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{v})) = s'(z)\mathrm{tr}(\boldsymbol{h}(\boldsymbol{r})\boldsymbol{h}(\boldsymbol{r})^\top) = s'(z)\mathrm{tr}(\boldsymbol{h}(\boldsymbol{r})^\top\boldsymbol{h}(\boldsymbol{r}))$$
$$= s'(z)\boldsymbol{h}(\boldsymbol{r})^\top\boldsymbol{h}(\boldsymbol{r}) = s'(z)(1+\boldsymbol{r}^\top\boldsymbol{r}). \quad (64)$$

Employing Eq. (61) and Eq. (515) in [52] leads to

$$\mathrm{tr}(\boldsymbol{H}_l(\boldsymbol{w},\boldsymbol{w})) = \mathrm{tr}(\boldsymbol{h}(\boldsymbol{x})\boldsymbol{h}(\boldsymbol{x})^\top \otimes (s'(z)\boldsymbol{F}''(\boldsymbol{y})\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}}\boldsymbol{F}'(\boldsymbol{y}) + \delta\mathrm{diag}(\widetilde{\boldsymbol{V}}^\top)\boldsymbol{F}''(\boldsymbol{y})))$$
$$= (1+\boldsymbol{x}^\top\boldsymbol{x})\mathrm{tr}(s'(z)\boldsymbol{F}'(\boldsymbol{y})\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}}\boldsymbol{F}'(\boldsymbol{y}) + \delta\mathrm{diag}(\widetilde{\boldsymbol{V}}^\top)\boldsymbol{F}''(\boldsymbol{y}))$$
$$= (1+\boldsymbol{x}^\top\boldsymbol{x})(s'(z)\|\boldsymbol{F}'(\boldsymbol{y})\widetilde{\boldsymbol{V}}^\top\|^2 + \delta\widetilde{\boldsymbol{V}}\boldsymbol{f}''(\boldsymbol{y})). \quad (65)$$

Furthermore, from Eq. (15) in [52], the trace for all training data is expressed as

$$\mathrm{tr}(\boldsymbol{H}_L(\boldsymbol{a},\boldsymbol{a})) = \sum_{i=1}^{I}\mathrm{tr}(\boldsymbol{H}_i(\boldsymbol{a},\boldsymbol{a})),\ \boldsymbol{a}\in\{\boldsymbol{w},\boldsymbol{v}\}. \quad (66)$$

Substituting

$$(\boldsymbol{E}_I + \boldsymbol{X}^\top\boldsymbol{X}\otimes\boldsymbol{E}_I)\boldsymbol{1}_I = \begin{bmatrix}1+\boldsymbol{x}_i^\top\boldsymbol{x}_i\end{bmatrix} \in \mathbb{R}^{I\times 1},$$
$$(\boldsymbol{E}_I + \boldsymbol{R}^\top\boldsymbol{R}\otimes\boldsymbol{E}_I)\boldsymbol{1}_I = \begin{bmatrix}1+\boldsymbol{r}_i^\top\boldsymbol{r}_i\end{bmatrix} \in \mathbb{R}^{I\times 1},$$
$$s'\otimes(\boldsymbol{F}_Y''^{\odot 2}\widetilde{\boldsymbol{V}}^{\odot 2\top}) = s'\otimes((\boldsymbol{F}_Y''\otimes\boldsymbol{F}_Y'')(\widetilde{\boldsymbol{V}}\otimes\widetilde{\boldsymbol{V}})^\top)$$
$$= \begin{bmatrix}s'(z_i)\|\boldsymbol{F}'(\boldsymbol{y}_i)\widetilde{\boldsymbol{V}}^\top\|^2\end{bmatrix} \in \mathbb{R}^{I\times 1},$$
$$\delta\odot(\boldsymbol{F}_Y''\widetilde{\boldsymbol{V}}^\top) = \begin{bmatrix}\delta_i\boldsymbol{f}''(\boldsymbol{y}_i)^\top\widetilde{\boldsymbol{V}}^\top\end{bmatrix} = \begin{bmatrix}\delta_i\widetilde{\boldsymbol{V}}\boldsymbol{f}''(\boldsymbol{y}_i)\end{bmatrix} \in \mathbb{R}^{I\times 1}$$

into Eqs. (20) and (21) and simplifying the terms shows that Eq. (66) and Eqs. (20)–(21) are equivalent.

I. Proof for Eq. (22)

From Eq. (64), we obtain $\mathrm{tr}(\boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{v})) = s'(z)(1+\boldsymbol{r}^\top\boldsymbol{r})$. Since $s(z)$ is a function that outputs the estimated probability in $(0,1)$ according to Eq. (1), it follows that

$$\max_{z\in\mathbb{R}} s'(z) = \max_{p\in(0,1)} p(1-p) = \frac{1}{4}. \quad (67)$$

It holds from Eqs. (36), (39), (42), (45), and (48) that

$$\sup_{\boldsymbol{r}}(1+\boldsymbol{r}^\top\boldsymbol{r}) = \sup_{\boldsymbol{y}\in\mathbb{R}^N}(1+\boldsymbol{f}(\boldsymbol{y})^\top\boldsymbol{f}(\boldsymbol{y})) = \begin{cases} N+1, f: \text{Sigmoid/Tanh} \\ \infty, f: \text{Linear/SmoothReLU/GELU}\end{cases}$$

Therefore, we have

$$\mathrm{tr}(\boldsymbol{H}_l(\boldsymbol{v},\boldsymbol{v})) < \begin{cases}\frac{1}{4}(N+1), f: \text{Sigmoid/Tanh} \\ \infty, f: \text{Linear/SmoothReLU/GELU}\end{cases}$$

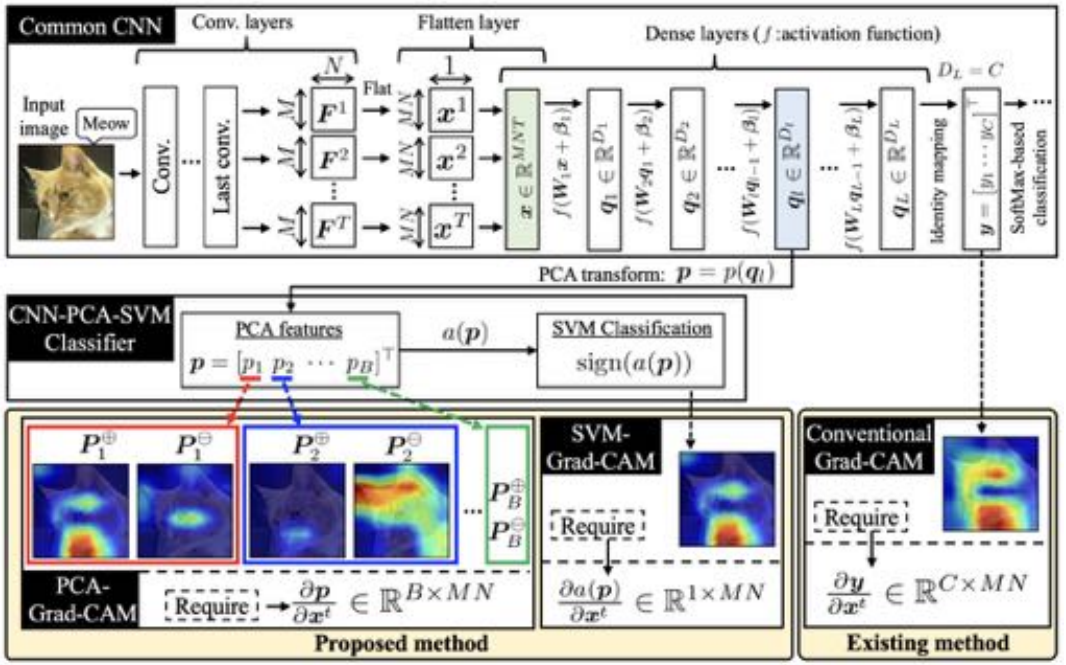
Substituting this into Eq. (66) yields Eq. (22).

最近やっている研究：

人類がまだ到達していない
機械学習について考える

（ハイレベルな研究者には敵いませんが...）

今まで誰も見れなかった
深層学習の主軸を
可視化する。



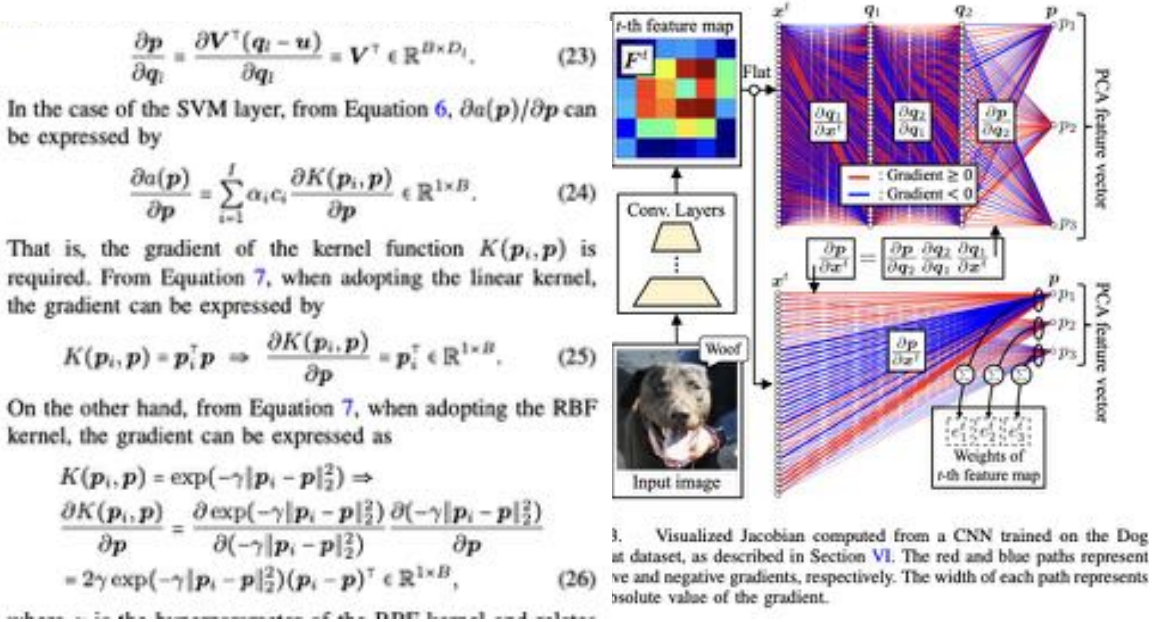
PCA- and SVM-Grad-CAM for Convolutional Neural Networks: Closed-form Jacobian Expression

畳み込みニューラルネットワークにおけるPCA-SVM-Grad-CAM: ヤコビ行列の閉形式表現

詳細:

https://arxiv.org/abs/2508.11880

https://arxiv.org/pdf/2508.11880



$$\frac{\partial p}{\partial q_l} = \frac{\partial V^\top (q_l - u)}{\partial q_l} = V^\top \in \mathbb{R}^{B\times D_l}. \tag{23}$$

In the case of the SVM layer, from Equation 6, $\partial a(p)/\partial p$ can be expressed by

$$\frac{\partial a(p)}{\partial p} = \sum_{i=1}^{I} \alpha_i c_i \frac{\partial K(p_i, p)}{\partial p} \in \mathbb{R}^{1\times B}. \tag{24}$$

That is, the gradient of the kernel function $K(p_i, p)$ is required. From Equation 7, when adopting the linear kernel, the gradient can be expressed by

$$K(p_i, p) = p_i^\top p \Rightarrow \frac{\partial K(p_i, p)}{\partial p} = p_i^\top \in \mathbb{R}^{1\times B}. \tag{25}$$

On the other hand, from Equation 7, when adopting the RBF kernel, the gradient can be expressed as

$$\begin{aligned} K(p_i, p) &= \exp(-\gamma |p_i - p|_2^2) \Rightarrow \\ \frac{\partial K(p_i, p)}{\partial p} &= \frac{\partial \exp(-\gamma |p_i - p|_2^2)}{\partial (-\gamma |p_i - p|_2^2)} \frac{\partial (-\gamma |p_i - p|_2^2)}{\partial p} \\ &= 2\gamma \exp(-\gamma |p_i - p|_2^2)(p_i - p)^\top \in \mathbb{R}^{1\times B}, \end{aligned} \tag{26}$$

3. Visualized Jacobian computed from a CNN trained on the Dog at dataset, as described in Section VI. The red and blue paths represent ve and negative gradients, respectively. The width of each path represents bsolute value of the gradient.

最近やっている研究：

人類がまだ到達していない機械学習について考える

（ハイレベルな研究者には敵いませんが...）

従来よりも良いベイズ最適化を作る。

EVI-GPBO: Estimated Variance Integration-Based Gaussian Process Bayesian Optimization

推定分散積分に基づくガウス過程ベイズ最適化

詳細:

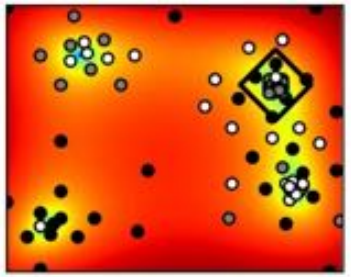
https://ieeexplore.ieee.org/document/10872956

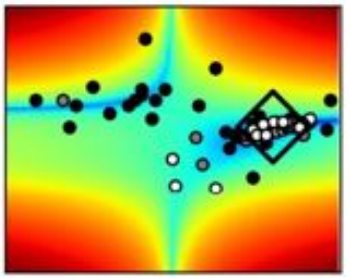
Algorithm 1 EVI-GPBO (Proposed Method)

Input: Initial observation set $D_0 = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_N, y_N)\}$, BO iterations T, gradient iterations T', learning rate r, weight coefficient τ, search space Ψ, integration range X, initial kernel parameters $\boldsymbol{v}^{(0)}$

Output: Solution $\boldsymbol{x}^\sharp$

$\boldsymbol{v}'^{(0)} \leftarrow \log \boldsymbol{v}^{(0)}$, Eq. 19
$N \leftarrow |D_0|$
for $t = 0$ **to** $T - 1$ **do**
  for $t' = 0$ **to** $T' - 1$ **do**
    $\boldsymbol{v}'^{(t'+1)} \leftarrow \boldsymbol{v}'^{(t')} + r \left.\frac{\partial L(\boldsymbol{v}, D_t)}{\partial \boldsymbol{v}'}\right|_{\boldsymbol{v} \leftarrow \exp(\boldsymbol{v}'^{(t')})}$, Eq. 23
  end for
  $\boldsymbol{v}^{(T')} \leftarrow \exp\left(\boldsymbol{v}'^{(T')}\right)$, Eq. 20
  $\beta_t \leftarrow \frac{\tau}{M} \int_X V_t(\boldsymbol{x}; \boldsymbol{v}^{(T')}) \mathrm{d}\boldsymbol{x}$, Eqs. 26, 40, and 45
  $\boldsymbol{x}_{|D_t|+1} \leftarrow \underset{\boldsymbol{x} \in \Psi}{\operatorname{argmax}} A_t\left(\boldsymbol{x}; \beta_t, \boldsymbol{v}^{(T')}\right)$, Eq. 10
  $y_{|D_t|+1} \leftarrow f(\boldsymbol{x}_{|D_t|+1}) + \epsilon$, Eq. 11
  $D_{t+1} \leftarrow D_t \cup \{(\boldsymbol{x}_{|D_t|+1}, y_{|D_t|+1})\}$, Eq. 13
  $\Psi \leftarrow \Psi \setminus \{\boldsymbol{x}_{|D_t|+1}\}$
end for
$\boldsymbol{x}^\sharp \leftarrow \boldsymbol{x}_{i^\sharp}$ s.t. $i^\sharp \leftarrow \underset{i \in \mathbb{N}_{\leq N+T}}{\operatorname{argmax}}\ y_i$, Eq. 14
return $\boldsymbol{x}^\sharp$





$$\boldsymbol{x}^* = \boldsymbol{x} - \frac{1}{2}(\boldsymbol{x}_i + \boldsymbol{x}_j), \tag{36}$$

Equation 28 can be represented as

$$\begin{aligned} & k^*(\boldsymbol{x}, \boldsymbol{x}_i, \boldsymbol{x}_j) \\ &= \zeta^2 \left(\exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2}{\eta}\right)\right)^{\frac{1}{2}} \exp\left(-\frac{2}{\eta}(\boldsymbol{x}^*)^\top (\boldsymbol{x}^*)\right) \\ &= \zeta^{\frac{3}{2}} \left(k(\boldsymbol{x}_i, \boldsymbol{x}_j) - \theta\delta_{ij}\right)^{\frac{1}{2}} \exp\left(-\frac{2}{\eta}(\boldsymbol{x}^*)^\top (\boldsymbol{x}^*)\right). \end{aligned} \tag{37}$$

With this substitution, the integration range changes from X to

$$X_{ij}^* = \prod_{p=1}^{d} \left[(x_p^*)_{ij}^{\min}, (x_p^*)_{ij}^{\max}\right], \ \boldsymbol{x}^* \in X_{ij}^*, \tag{38}$$

where, $(x_p^*)_{ij}^{\min} = x_p^{\min} - \frac{1}{2}(x_{ip} + x_{jp})$ and $(x_p^*)_{ij}^{\max} = x_p^{\max} - \frac{1}{2}(x_{ip} + x_{jp})$. The Jacobian matrix with respect to $\boldsymbol{x}^*$ can be represented as

$$- \operatorname{erf}\left[\sqrt{\frac{2}{\eta}}\left(x_p^{\min} - \frac{x_{ip} + x_{jp}}{2}\right)\right]\Bigg). \tag{41}$$

$\operatorname{erf}[z]$ is a Gaussian error function, which is defined as

$$\operatorname{erf}[z] := \frac{2}{\sqrt{\pi}} \int_0^z \exp(-t^2) \mathrm{d}t. \tag{42}$$

Note that when transforming Equation 40, we use the following relationship:

$$\begin{aligned} & \int_{(x_p^*)_{ij}^{\min}}^{(x_p^*)_{ij}^{\max}} \exp\left(-\frac{2}{\eta}(x_p^*)^2\right) \mathrm{d}x_p^* \\ &= \sqrt{\frac{\eta}{2}} \int_{\sqrt{\frac{2}{\eta}}(x_p^*)_{ij}^{\min}}^{\sqrt{\frac{2}{\eta}}(x_p^*)_{ij}^{\max}} \exp\left(-u^2\right) \mathrm{d}u \\ &= \sqrt{\frac{\eta\pi}{8}} \left(\operatorname{erf}\left[\sqrt{\frac{2}{\eta}}(x_p^*)_{ij}^{\max}\right] - \operatorname{erf}\left[\sqrt{\frac{2}{\eta}}(x_p^*)_{ij}^{\min}\right]\right). \end{aligned} \tag{43}$$