IVA と DNN を近接平均化した優決定 BSS における DNN のリプシッツ定数に関する検討 *

◎ 松本和樹 (早大), 山田宏樹, 矢田部浩平 (農工大)

1 はじめに

Plug-and-Play (PnP) は最適化アルゴリズムに DNN を取り入れる枠組みの一つである．これに基づき，我々は主双対近接分離 (PDS) アルゴリズムに DNN と独立ベクトル分析 (IVA) の近接平均 (PA) を PnP する手法 (PA-BSS) を提案した [1, 2]．PA-BSS は DNN と IVA の利点を組み合わせ，単純な雑音除去 DNN を用いて高精度な分離行列推定を実現する．

本稿では，画像処理分野の PnP で重視される DNN のリプシッツ定数に着目し，線形層や畳み込み層のリプシッツ定数を正規化する Spectral Normalization (SN)[3] を PA-BSS に導入する．実験の結果，SN は分離性能の向上をもたらさないものの，数値安定性の観点では有用であることが確認された．

2 PA-BSS [1,2]

PA-BSS は雑音除去 DNN を優決定 BSS に活用するための枠組みであり，時間周波数領域における M 音源の N チャネル観測信号 $\mathbf{x}[f,t] \in \mathbb{C}^N$ に対し，分離行列 $\mathbf{W}[f] \in \mathbb{C}^{M\times N}$ を推定することで分離音 $\mathbf{y}[f,t] = \mathbf{W}[f]\,\mathbf{x}[f,t] \in \mathbb{C}^M$ を得る．ただし，$1 \le t \le T$, $1 \le f \le F$ は時間および周波数のインデックスである．PA-BSS は **Alg. 1** に示す通りで，最小化問題

$$\min_{(\mathbf{W}[f])_{f=1}^F} \mathcal{P}(\mathbf{y}) - 2\sum_{f=1}^{F} \log|\det \mathbf{W}[f]| \qquad (1)$$

を PDS を用いて解く更新式に対し，PnP と PA の枠組みに基づき DNN を取り入れている．具体的には，音源モデルに依存する正則化項 $\mathcal{P}$ の近接作用素

$$\mathrm{prox}_{\mu\mathcal{P}}(\mathbf{y}) = \arg\min_{\mathbf{v}} \left(\mathcal{P}(\mathbf{v}) + \frac{1}{2\mu}\|\mathbf{v}-\mathbf{y}\|_2^2\right) \qquad (2)$$

を，IVA に対応する近接作用素 $\mathrm{prox}_{\ell_{2,1}}$ と雑音除去 DNN の重み付き平均で置き換えている．ただし，$\mathbf{X}$ は観測信号からなる行列，$\mathbf{w}$ はベクトル化した分離行列，$\boldsymbol{\xi}$ は双対変数，$\mathbf{z}$ は一時変数，μ_1, μ_2 はステップサイズである．平均化率を $\alpha \in (0,1)$ とすれば，IVA と DNN を組み合わせた音源モデルが得られる [1, 2].

3 SN を施した DNN の学習と分析

画像処理分野の PnP では，アルゴリズムの安定性を向上させるため，DNN の各層のリプシッツ定数に対し正規化を施す場合がある [4]．ただし，作用素 $\mathcal{M}$ のリプシッツ定数が k であるとは，任意の $\mathbf{x}, \mathbf{y}$ に対し $\|\mathcal{M}(\mathbf{x}) - \mathcal{M}(\mathbf{y})\| \le k\|\mathbf{x}-\mathbf{y}\|$ を満たすことをいう．

Algorithm 1 PA-BSS

Require: $\mathbf{X}$, $\mathbf{w}^{[1]}$, $\boldsymbol{\xi}^{[1]}, \mu_1, \mu_2, \alpha$
Ensure: $\mathbf{w}^{[i+1]}$

1: **for** $i = 1, 2, \cdots,$ NumIteration **do**
2: $\quad \mathbf{w}^{[i+1]} = \mathrm{prox}_{-2\mu_1 \sum \log|\det(\cdot[f])|}\left(\mathbf{w}^{[i]} - \mu_1\mu_2 \mathbf{X}^{\mathsf{H}}\boldsymbol{\xi}^{[i]}\right)$
3: $\quad \mathbf{z}^{[i]} = \boldsymbol{\xi}^{[i]} + \mathbf{X}(2\mathbf{w}^{[i+1]} - \mathbf{w}^{[i]})$
4: $\quad \boldsymbol{\xi}^{[i]} = \mathbf{z}^{[i]} - \left((1-\alpha)\ \mathrm{prox}_{\frac{1}{\mu_2}\ell_{2,1}}(\mathbf{z}^{[i]}) + \alpha\,\mathrm{DNN}(\mathbf{z}^{[i]})\right)$
5: **end for**

本稿では，DNN のリプシッツ定数と PA-BSS の安定性の関連を調査する目的で，畳み込み層に対するリプシッツ定数の正規化が分離性能や分離行列の収束性におよぼす影響を調査する．まずは DNN の学習方法や得られた DNN の性質について述べる．

3.1 SN による畳み込み層の正規化

本稿では，画像生成の安定化で有効性が確認されている SN[3] に基づき正規化を行う．SN は畳み込み層のリプシッツ定数を厳密に 1 に正規化しないものの [4]，数値安定性の向上は期待できる．また，本稿では学習を安定化する目的で正規化の強さ r を導入する．i 層目の畳み込み層における j 個目の重み行列 $\mathbf{K}[i,j] \in \mathbb{R}^{W\times H}$ に対する操作は以下の式で書ける．

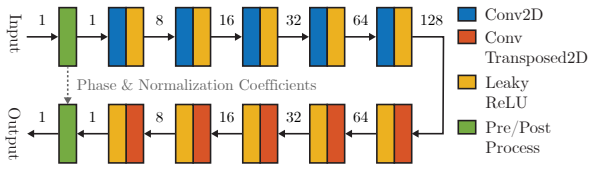
$$\begin{aligned} \mathbf{v}[i,j] &\leftarrow \mathbf{K}[i,j]^{\mathsf{T}}\mathbf{u}[i,j]/\|\mathbf{u}[i,j]\|_2 \\ \mathbf{u}[i,j] &\leftarrow \mathbf{K}[i,j]\mathbf{v}[i,j]/\|\mathbf{v}[i,j]\|_2 \\ \mathbf{K}[i,j] &\leftarrow \mathbf{K}[i,j]/\left(\mathbf{u}[i,j]^{\mathsf{T}}\mathbf{K}[i,j]\mathbf{v}[i,j]\right)^r \end{aligned} \qquad (3)$$

ただし，$\mathbf{u}[i,j] \in \mathbb{R}^W, \mathbf{v}[i,j] \in \mathbb{R}^H$ はランダムに初期化されたベクトルである．正規化の強さを $r \in (0,1]$ とすれば，重み行列の最大特異値は徐々に 1 に近づく．

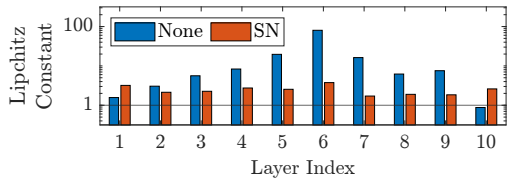
3.2 雑音除去 DNN の学習と分析

本稿では図–1 に示す DNN を用いる．DNN は目的音声の振幅スペクトログラム $\mathbf{S} \in \mathbb{R}_+^{F\times T}$ に対し，妨害音声 $\mathbf{I} \in \mathbb{R}_+^{F\times T}$ と音声整形雑音 $\mathbf{Z} \in \mathbb{R}_+^{F\times T}$ にランダムな係数 $a_{\mathrm{I}}, a_{\mathrm{Z}} \in [0, 0.5]$ を掛けて足した混合音 $\mathbf{M} = \mathbf{S} + a_{\mathrm{I}}\mathbf{I} + a_{\mathrm{Z}}\mathbf{Z}$ から目的音声 $\mathbf{S}$ を推定するように学習した．損失関数としては，時間周波数領域における平均二乗誤差を用いた．音声は Libri-TTS-R の `train100`[5] からランダム抽出し，100 epoch 学習した．最適化にはバッチサイズ 32，学習率 0.001 の

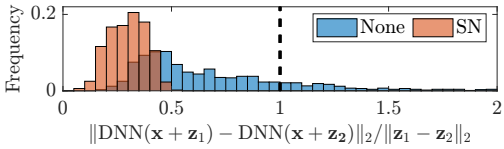
* Research on Lipschitz constants of DNNs proximal-averaged with IVA in determined BSS. By Kazuki MATSUMOTO (Waseda Univ.), Koki YAMADA and Kohei YATABE (TUAT).



図–1　実験に用いた DNN．畳み込み層のカーネルサイズは 5×3，ストライドは $[2, 2]$ である．信号線の上部にはその時点における特徴量のチャネル数を示している．前処理は入力の絶対値を取り，周波数ごとにパワー正規化を施すもので，後処理は前処理と逆の操作に対応する．



図–2　畳み込み層のリプシッツ定数．realSN[4] に倣い，各畳み込み層のリプシッツ定数を冪乗法で計算した．



図–3　雑音を加えた音声 $\mathbf{x}+\mathbf{z}_1$, $\mathbf{x}+\mathbf{z}_2$ の拡大率の分布．サンプル数は 1000 で，雑音の振幅はランダムに決定した．

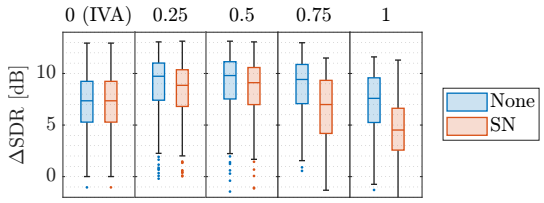
Adam を用いた．SN のパラメータ r に関しては，正規化なしの場合 (None) は 0 で固定し，正規化ありの場合 (SN) は 80 epoch 以前は 0，以降は 0.01 とした．

得られた DNN を分析したところ，SN による安定性の向上が確認された．具体的には，SN の導入は各層のリプシッツ定数を抑え（図–2），入出力の前後で距離が拡大するケースが生じなくなった（図–3）．

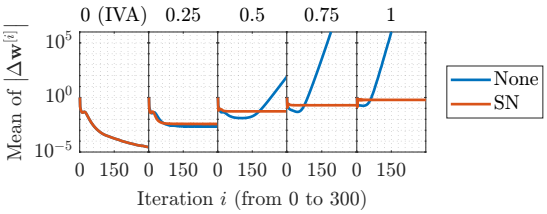
4　分離行列の収束性と性能の評価

次に，学習した DNN を用いて PA-BSS による音源分離の性能評価を行った．SiSEC 2011 の `dev1` 中の 8 音源に対して室内インパルス応答を畳み込むことで 2 チャネルの 2 話者混合音を計 224 組作成した．音源方向は $(-45°, 30°)$，$(-75°, 30°)$，$(-45°, 60°)$，$(-75°, 60°)$ の 4 組，音源距離は 1 m，マイク間隔は 8 cm，残響時間は 0.16 秒とした．ステップサイズは $\mu_1 = \mu_2 = 1$，反復回数は 300 回とした．平均化率 α には 0, 0.25, 0.5, 0.75, 1 を用いた．前処理には白色化を，後処理にはプロジェクションバックを用いた．

まずは分離性能の観点から結果を考察する．図–4 に平均化率 α ごとの ΔSDR を示す．SN の有無にかかわらず，DNN と IVA を適度に平均化した $\alpha = 0.25,\ 0.5$ は，IVA ($\alpha = 0$) および DNN 単体の PnP ($\alpha = 1$) よりも高い分離性能を実現した．このことから，DNN と IVA を近接平均化する重要性が再確認された．一方，正規化の有無に着目すると，SN は None と比較して分離性能が劣った．このことから，SN が DNN



図–4　正規化手法および平均化率 α ごとの分離性能．



図–5　正規化手法および平均化率 α ごとの分離行列変化量の推移．分離行列が収束すれば変化量は 0 に収束する．

のモデリング能力を損なう可能性が示唆された．

次に，数値安定性の観点から比較を行う．図–5 に分離行列の変化量の推移を示す．ただし，変化量は反復前後の分離行列の差分 $\Delta \mathbf{w}^{[i]} = \mathbf{w}^{[i]} - \mathbf{w}^{[i-1]}$ の絶対値の平均値である．結果として，$\alpha = 0$ (IVA) のとき，分離行列が収束することが分かった．また，$\alpha > 0$ のとき，SN では変化量が一定の値に留まる一方で，None では変化量が発散するという違いが現れた．これらのことから，SN は PA-BSS における分離行列の発散を防ぐことが確認された．なお，None の数値不安定性が分離性能に現れないのは，最適化時に発散した分離行列がプロジェクションバックで正規化されるためだと考えられる．しかしながら，分離行列の発散は演算精度に悪影響を及ぼすため，分離性能を損なわずに数値安定性を確保する対策が望まれる．

5　おわりに

本稿では PA-BSS における畳み込み層のリプシッツ定数の正規化に関して調査した．結果として，SN による正規化は分離性能向上に寄与しないものの，数値安定性を向上させることが分かった．今後は安定性と分離性能を両立可能な枠組みを模索する．

参考文献

[1] 松本和樹, 矢田部浩平, “主-双対近接分離法に DNN を Plug-and-Play した優決定ブラインド音源分離,” 音講論集, pp. 147–148 (2023.9).

[2] K. Matsumoto and K. Yatabe, “Determined BSS by combination of IVA and DNN via proximal averaging,” *Proc. IEEE ICASSP* (2024).

[3] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral Normalization for Generative Adversarial Networks,” *Proc. ICLR* (2018).

[4] E. K. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, “Plug-and-Play Methods Provably Converge with Properly Trained Denoisers,” *Proc. ICML*, pp. 5546–5557 (2019).

[5] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, “LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus,” *Proc. Interspeech*, pp. 5496–5500 (2023).