DATA CENSORING WITH SET-MEMBERSHIP ALGORITHMS

Paulo S.R. Diniz\textsuperscript{1}
Hamed Yazdanpanah\textsuperscript{1}

\textsuperscript{1}Laboratório de Sinais, Multimídia e Telecomunicações (SMT)
Departamento de Engenharia Eletrônica e de Computação (DEL)
Universidade Federal do Rio de Janeiro (UFRJ)
{diniz, hamed.yazdanpanah}@smt.ufrj.br

IEEE Global Conference on Signal and Information Processing (GlobalSIP 2017)
Presentation outline

1. Introduction

2. ST-SM-NLMS and DT-SM-NLMS Algorithms

3. Compute proper $\gamma$ to obtain the desired update rate

4. Results

5. Conclusions
Outline

1. Introduction
2. ST-SM-NLMS and DT-SM-NLMS Algorithms
3. Compute proper $\gamma$ to obtain the desired update rate
4. Results
5. Conclusions
Data Censoring

- Apply set estimation theory to censor redundant data through:
  1. Single Threshold Set-Membership Normalized LMS (ST-SM-NLMS) algorithm
  2. Double Threshold Set-Membership Normalized LMS (DT-SM-NLMS) algorithm

Set Estimation Theory

- Finds a solution to a given optimization problem. Any solution within the feasible set is acceptable.
- Examples of estimators:
  - Batch processing: few techniques (usually too complex)
  - Iterative processing: optimal-bounding-ellipsoids (OBE) and set-membership (SM) algorithms
Data Censoring

- Apply set estimation theory to censor redundant data through:
  1. Single Threshold Set-Membership Normalized LMS (ST-SM-NLMS) algorithm
  2. Double Threshold Set-Membership Normalized LMS (DT-SM-NLMS) algorithm

Set Estimation Theory

- Finds a solution to a given optimization problem → Any solution within the feasible set is acceptable
- Examples of estimators:
  - Batch processing: few techniques (usually too complex)
  - Iterative processing: optimal-bounding-ellipsoids (OBE) and set-membership (SM) algorithms
Formulation

Main Sets

- **Constraint set:**

\[ \mathcal{H}(k) \triangleq \{ \mathbf{w} \in \mathbb{R}^{N+1} : |d(k) - \mathbf{w}^T \mathbf{x}(k)| \leq \overline{\gamma} \} \]

where
- Error: \( e(k) \triangleq d(k) - \mathbf{w}^T \mathbf{x}(k) \)
- Uncertainties are modeled through \( \overline{\gamma} \)

- Feasibility set \( \Rightarrow \) set of acceptable solutions

\[ \Theta \triangleq \bigcap_{k \in \mathbb{N}} \mathcal{H}(k) \]

Problem Formulation

- **Inputs:** all data-pairs \((\mathbf{x}(k), d(k))\)

- **Target:** find \( \mathbf{w} \in \Theta \)
Challenges

- Incomplete data
  - Impossible to guarantee that all input data-pairs are available
  - Online/iterative processing:
    - Must produce an estimate every time a new input data-pair arrives
    - $\Theta$ can be iteratively estimated via $\psi(k)$

$$\psi(k) \triangleq \bigcap_{i=0}^{k} \mathcal{H}(i)$$

- $\psi(k)$ converges to $\Theta$ as $k \to \infty$
- Problem: $k \to \infty \Rightarrow$ Infinite memory and prohibitive complexity
- Solution: Use the last constraint set at each iteration $\rightarrow$ SM-NLMS algorithm
Introduction

Set-Membership Filtering (SMF)

Challenges

- **Incomplete data**
  - Impossible to guarantee that all input data-pairs are available
- **Online/iterative processing:**
  - Must produce an estimate every time a new input data-pair arrives
  - $\Theta$ can be iteratively estimated via $\psi(k)$

$$\psi(k) \triangleq \bigcap_{i=0}^{k} H(i)$$

- $\psi(k)$ converges to $\Theta$ as $k \to \infty$
- Problem: $k \to \infty \Rightarrow$ **Infinite memory and prohibitive complexity**
- Solution: Use the last constraint set at each iteration $\rightarrow$ **SM-NLMS algorithm**
Outline

1 Introduction

2 ST-SM-NLMS and DT-SM-NLMS Algorithms

3 Compute proper $\gamma$ to obtain the desired update rate

4 Results

5 Conclusions
Proposed algorithms

**ST-SM-NLMS algorithm**

\[ \mathbf{w}(k + 1) = \mathbf{w}(k) + \frac{\mu(k)}{\| \mathbf{x}(k) \|^2 + \delta} e(k) \mathbf{x}(k), \]

where

\[ \mu(k) \triangleq \begin{cases} 1 - \frac{\overline{\gamma}}{|e(k)|} & \text{if } |e(k)| > \overline{\gamma}, \\ 0 & \text{otherwise} \end{cases} . \]

**DT-SM-NLMS algorithm**

\[ \mathbf{w}(k + 1) = \mathbf{w}(k) + \frac{\mu(k)}{\| \mathbf{x}(k) \|^2 + \delta} e(k) \mathbf{x}(k), \]

where

\[ \mu(k) \triangleq \begin{cases} 1 - \frac{\overline{\gamma}}{|e(k)|} & \text{if } \overline{\gamma} < |e(k)| < \overline{\gamma}', \\ 0 & \text{otherwise} \end{cases} . \]
Proposed algorithms

**ST-SM-NLMS algorithm**

\[
\mathbf{w}(k+1) = \mathbf{w}(k) + \frac{\mu(k)}{\|\mathbf{x}(k)\|^2 + \delta} e(k)\mathbf{x}(k),
\]

where

\[
\mu(k) \triangleq \begin{cases}
1 - \frac{\gamma}{|e(k)|} & \text{if } |e(k)| > \gamma, \\
0 & \text{otherwise}
\end{cases}
\]

**DT-SM-NLMS algorithm**

\[
\mathbf{w}(k+1) = \mathbf{w}(k) + \frac{\mu(k)}{\|\mathbf{x}(k)\|^2 + \delta} e(k)\mathbf{x}(k),
\]

where

\[
\mu(k) \triangleq \begin{cases}
1 - \frac{\gamma}{|e(k)|} & \text{if } \gamma < |e(k)| < \gamma', \\
0 & \text{otherwise}
\end{cases}
\]
ST-SM-NLMS and DT-SM-NLMS Algorithms

\[ w(k) \]
\[ d(k)-w^T x(k) = \tilde{\gamma} \]
\[ \mathcal{H}(k) \]
\[ d(k)-w^T x(k) = -\tilde{\gamma} \]

Figure: Coefficient vector updating for: (a) ST-SM-NLMS; (b) DT-SM-NLMS.
Outline

1. Introduction
2. ST-SM-NLMS and DT-SM-NLMS Algorithms
3. Compute proper $\gamma$ to obtain the desired update rate
4. Results
5. Conclusions
Compute proper $\gamma$ to obtain the desired update rate

**Our goal**

- **What we want?**
  - Find the proper $\gamma$ to achieve the desired update rate $p$ (censor 100(1-$p$)% of data)
- **How?**
  - Calculate $\gamma$ such that $\mathbb{P}(|e(k)| > \gamma) = p$
- **What is the distribution of $e(k)$?**
  - Adaptive system has sufficient order $\Rightarrow e(k) \overset{d}{\sim} n(k)$ in the steady-state
  - Assume $n(k) \sim \mathcal{N}(0, \sigma_n^2) \Rightarrow e(k) \sim \mathcal{N}(\mathbb{E}[\tilde{e}(k)], \sigma_n^2 + \mathbb{E}[\tilde{e}^2(k)])$, where
    \[
    \mathbb{E}[\tilde{e}^2(k)] = \frac{(\sigma_n^2 + \gamma^2 - 2\gamma\sigma_n^2 \rho_0(k))p}{(2 - p) - 2(1 - p)\gamma \rho_0(k)},
    \]
    and
    \[
    \rho_0(k) = \sqrt{\frac{2}{\pi (2\sigma_n^2 + \gamma^2)}}.
    \]
Compute proper $\gamma$ to obtain the desired update rate

Our goal

- **What we want?**
  - Find the proper $\gamma$ to achieve the desired update rate $p$ (censor $100(1-p)\%$ of data)

- **How?**
  - Calculate $\gamma$ such that $\mathbb{P}[|e(k)| > \gamma] = p$

- What is the distribution of $e(k)$?
  
  - Adaptive system has sufficient order $\Rightarrow e(k) \overset{d}{\sim} n(k)$ in the steady-state
  
  - Assume $n(k) \overset{d}{\sim} \mathcal{N}(0, \sigma_n^2) \Rightarrow e(k) \overset{d}{\sim} \mathcal{N}(\mathbb{E}[\tilde{e}(k)], \sigma_n^2 + \mathbb{E}[\tilde{e}^2(k)])$, where

\[
\mathbb{E}[\tilde{e}^2(k)] = \frac{(\sigma_n^2 + \gamma^2 - 2\gamma^2 \rho_0(k))p}{([2 - p) - 2(1 - p)\gamma \rho_0(k)]},
\]

and

\[
\rho_0(k) = \sqrt{\frac{2}{\pi (2\sigma_n^2 + \gamma^2)}}.
\]
Our goal

- **What we want?**
  - Find the proper $\gamma$ to achieve the desired update rate $p$ (censor $100(1-p)\%$ of data)

- **How?**
  - Calculate $\gamma$ such that $\mathbb{P}[|e(k)| > \gamma] = p$

- **What is the distribution of $e(k)$?**
  - Adaptive system has sufficient order $\Rightarrow e(k) \overset{d}{\sim} n(k)$ in the steady-state
  
  - Assume $n(k) \overset{d}{\sim} \mathcal{N}(0, \sigma_n^2) \Rightarrow e(k) \overset{d}{\sim} \mathcal{N}(\mathbb{E}[\tilde{e}(k)], \sigma_n^2 + \mathbb{E}[\tilde{e}^2(k)])$, where
    
    $$\mathbb{E}[\tilde{e}^2(k)] = \frac{(\sigma_n^2 + \gamma^2 - 2\gamma\sigma_n^2\rho_0(k))p}{(2 - p) - 2(1 - p)\gamma\rho_0(k)},$$

    and
    
    $$\rho_0(k) = \sqrt{\frac{2}{\pi(2\sigma_n^2 + \gamma^2)}}.$$
Our goal

• What we want?
  • Find the proper $\overline{\gamma}$ to achieve the desired update rate $p$ (censor $100(1-p)$% of data)

• How?
  • Calculate $\overline{\gamma}$ such that $\mathbb{P}(|e(k)| > \overline{\gamma}) = p$

• What is the distribution of $e(k)$?
  • Adaptive system has sufficient order $\Rightarrow e(k) \overset{d}{\sim} n(k)$ in the steady-state
  • Assume $n(k) \overset{d}{\sim} \mathcal{N}(0, \sigma_n^2) \Rightarrow e(k) \overset{d}{\sim} \mathcal{N}(\mathbb{E}[\tilde{e}(k)], \sigma_n^2 + \mathbb{E}[\tilde{e}^2(k)])$, where

$$\mathbb{E}[\tilde{e}^2(k)] = \frac{(\sigma_n^2 + \overline{\gamma}^2 - 2\overline{\gamma}\sigma_n^2 \rho_0(k))p}{[(2 - p) - 2(1 - p)\overline{\gamma}\rho_0(k)]},$$

and

$$\rho_0(k) = \sqrt{\frac{2}{\pi(2\sigma_n^2 + \overline{\gamma}^2)}}.$$
Compute proper $\gamma$ to obtain the desired update rate

How to determine $\gamma$?

- Step 1: choose $\gamma$ such that
  \[ \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp\left(-\frac{r^2}{2\sigma_n^2}\right) dr = \frac{p}{2}; \]

- Step 2: compute $\mathbb{E}[\tilde{e}^2(k)]$ and put $\sigma_e^2 \triangleq \mathbb{E}[\tilde{e}^2(k)] + \sigma_n^2$;

- Step 3: choose $\gamma$ such that
  \[ \int_{\gamma}^{\infty} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{r^2}{2\sigma_e^2}\right) dr = \frac{p}{2}, \]
  and repeat from step 2.

Observation: in practice, we do not need repeat this algorithm more than three iterations, since the difference between two consecutive $\gamma$s becomes insignificant.
Outline

1 Introduction

2 ST-SM-NLMS and DT-SM-NLMS Algorithms

3 Compute proper $\gamma$ to obtain the desired update rate

4 Results

5 Conclusions
Scenario I: system identification

- Algorithms tested: NLMS, AC-LMS, ST-SM-NLMS algorithms;
- Input signals: BPSK, zero-mean white Gaussian noise with unit variance (WGN), AR(1) (first-order autoregressive);
- Filter order: \( N = 29 \);
- \( w(0) = [0, \cdots, 0]^T \);
- SNR: 20 dB;
- Regularization factor: \( \delta = 10^{-12} \);
- Desired update rates, \( p \): 0.1, 0.2, and 0.03;
- Step size: 0.9 and 0.004 for the NLMS and AC-LMS algorithms, respectively;
- AC-LMS censorship threshold: 1.7 (experimentally for 10% update rate);
- For \( p = 0.1, 0.2, 0.3 \) the estimated \( \bar{\gamma} = 0.1875, 0.1477, 0.1194 \), respectively.
Learning (MSE) curves

- For \( p = 0.1 \), \( \bar{\gamma} \) is estimated as \( \bar{\gamma} = 0.1875 \).

Figure: (a) BPSK; (b) WGN; (c) AR(1).
Update rates of the ST-SM-NLMS algorithm

Table: The results of update rates using the estimated $\gamma$s for the ST-SM-NLMS algorithm.

<table>
<thead>
<tr>
<th>Input signal</th>
<th>$p = 0.1$</th>
<th>$p = 0.2$</th>
<th>$p = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPSK</td>
<td>0.087</td>
<td>0.186</td>
<td>0.298</td>
</tr>
<tr>
<td>WGN</td>
<td>0.090</td>
<td>0.189</td>
<td>0.299</td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.099</td>
<td>0.202</td>
<td>0.305</td>
</tr>
</tbody>
</table>

- When $p = 0.1$, the update rates of the AC-LMS algorithm for BPSK, WGN, AR(1) input signals are 10.9%, 10.9%, 15.7%, respectively.
- In [13, Galdino et al.], $\gamma$ is estimated by $\gamma \triangleq \text{erfc}^{-1}(p)\sqrt{2(M + 1)\sigma_n} = 0.1655$, when $p = 0.1$. In this case, the update rates of the ST-SM-NLMS algorithm for BPSK, WGN, AR(1) input signals are 13.4%, 13.5%, 14.6%, respectively.
- Observation: the second column of Table shows that our estimation of $\gamma$ censors the data with more precision.
Scenario II: System Identification with the existence of an outlier signal

- Algorithms tested: NLMS, rAC-LMS, ST-SM-NLMS, DT-SM-NLMS algorithms;
- Input signals: BPSK, zero-mean white Gaussian noise with unit variance (WGN), AR(1) (first-order autoregressive);
- Filter order: $N = 29$;
- $w(0) = [0, \cdots, 0]^T$;
- SNR: 20 dB;
- Regularization factor: $\delta = 10^{-12}$;
- Desired update rates, $p$: 0.1, 0.2, and 0.03;
- Step size: 0.9 and 0.004 for the NLMS and rAC-LMS algorithms, respectively;
- rAC-LMS censorship thresholds: 3 and 10 (experimentally for 10% update rate);
- For $p = 0.1, 0.2, 0.3$ the estimated $\bar{\gamma} = 0.1875, 0.1477, 0.1194$, respectively;
- The second threshold for DT-SM-NLMS algorithm: $\bar{\gamma'} = 1$;
- Outlier signal: Bernoulli process takes 1 with probability 0.05, multiplying $\mathcal{U}(0, 50)$. 
For $p = 0.1$, $\gamma' = 1$ and $\gamma$ is estimated as $\gamma = 0.1875$.

Figure: (a) BPSK; (b) WGN; (c) AR(1).
**Update rates of the DT-SM-NLMS algorithm**

**Table:** The results of update rates using the estimated $\gamma$s for the DT-SM-NLMS algorithm.

<table>
<thead>
<tr>
<th>Input signal</th>
<th>$p = 0.1$</th>
<th>$p = 0.2$</th>
<th>$p = 0.3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPSK</td>
<td>0.090</td>
<td>0.188</td>
<td>0.292</td>
</tr>
<tr>
<td>WGN</td>
<td>0.091</td>
<td>0.190</td>
<td>0.293</td>
</tr>
<tr>
<td>AR(1)</td>
<td>0.099</td>
<td>0.196</td>
<td>0.299</td>
</tr>
</tbody>
</table>

When $p = 0.1$, the update rates of the rAC-LMS algorithm for BPSK, WGN, AR(1) input signals are 10.2%, 10.2%, 12.3%, respectively.
Outline

1. Introduction
2. ST-SM-NLMS and DT-SM-NLMS Algorithms
3. Compute proper $\gamma$ to obtain the desired update rate
4. Results
5. Conclusions
Conclusions

In this presentation:

- Revisited set estimation theory with emphasis on set-membership filtering
- Revisited single threshold set-membership NLMS algorithm to censor the data
- Proposed double threshold set-membership NLMS algorithm to censor redundant data and non-innovative data caused by outlier
- Estimated the suitable threshold parameter for the desired update rate
- By using the estimated threshold, the proposed algorithms censor the data effectively
- The proposed algorithms have better performance compared to the NLMS, AC-LMS, and rAC-LMS algorithms
- Corroborated the effectiveness of set-membership filtering in data censorship
Thank You!