# DATA CENSORING WITH SET-MEMBERSHIP ALGORITHMS

Paulo S.R. Diniz[1]
Hamed Yazdanpanah[1]

[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)

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

# Presentation outline

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

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

UFRJ

# Outline

# Content

## 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

# Content

## 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 $\longrightarrow$ 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:

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

  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 $\mathbf{w} \in \Theta$

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

UFRJ

# 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 $\longrightarrow$ SM-NLMS algorithm

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

# 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 $k \rightarrow \infty$
    - Problem: $k \rightarrow \infty \Rightarrow$ Infinite memory and prohibitive complexity
    - Solution: Use the last constraint set at each iteration $\longrightarrow$ SM-NLMS algorithm

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

# Outline

# 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

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

## 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

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

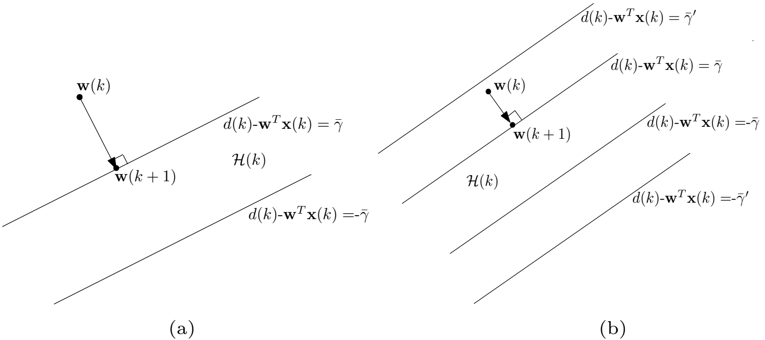



Figure: Coefficient vector updating for: (a) ST-SM-NLMS; (b) DT-SM-NLMS.

# Outline



Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

UFRJ

# 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

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

and

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

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

UFRJ

# 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

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

and

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

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

# 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

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

and

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

# 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

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

and

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

# How to determine $\overline{\gamma}$?

- Step 1: choose $\overline{\gamma}$ such that

$$\int_{\overline{\gamma}}^{\infty} \frac{1}{\sqrt{2\pi\sigma_n^2}} \exp(-\frac{r^2}{2\sigma_n^2})dr = \frac{p}{2};$$

- 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 $\overline{\gamma}$ such that

$$\int_{\overline{\gamma}}^{\infty} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp(-\frac{r^2}{2\sigma_e^2})dr = \frac{p}{2},$$

  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 $\overline{\gamma}$s becomes insignificant.

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

# Outline



Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

UFRJ

# Scenario I: system identification

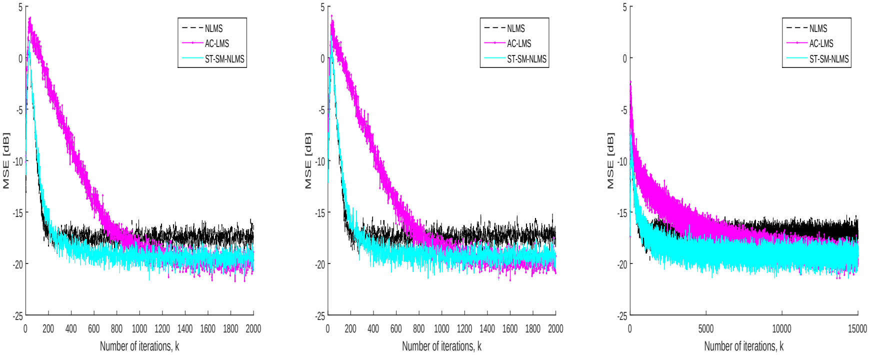
- Algorithms tested: NLMS, AC-LMS, ST-SM-NLMS algorithms;
- Input signals: BPSK, zero-mean white Gaussian noise with unit variance (WGN), AR(1) (first-order autoregressive);
- Filter order: $N = 29$ ;
- $\mathbf{w}(0) = [0, \cdots, 0]^T$;
- SNR: 20 dB;
- Regularization factor: $\delta = 10^{-12}$;
- 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 $\overline{\gamma}$=0.1875, 0.1477, 0.1194, respectively.

SMT
Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

UFRJ

# Learning (MSE) curves

- For $p = 0.1$, $\overline{\gamma}$ is estimated as $\overline{\gamma} = 0.1875$.




(a) (b) (c)

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 $\overline{\gamma}$s for the ST-SM-NLMS algorithm.

| Input signal | $p = 0.1$ | $p = 0.2$ | $p = 0.3$ |
|---|---|---|---|
| BPSK | 0.087 | 0.186 | 0.298 |
| WGN | 0.090 | 0.189 | 0.299 |
| AR(1) | 0.099 | 0.202 | 0.305 |

- 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.], $\overline{\gamma}$ is estimated by $\overline{\gamma} \triangleq \text{erfc}^{-1}(p)\sqrt{2(M+1)}\sigma_n = 0.1655$, 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 $\overline{\gamma}$ censors the data with more precision.

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

# Scenario II: System Identification with the existence of an outlier signal

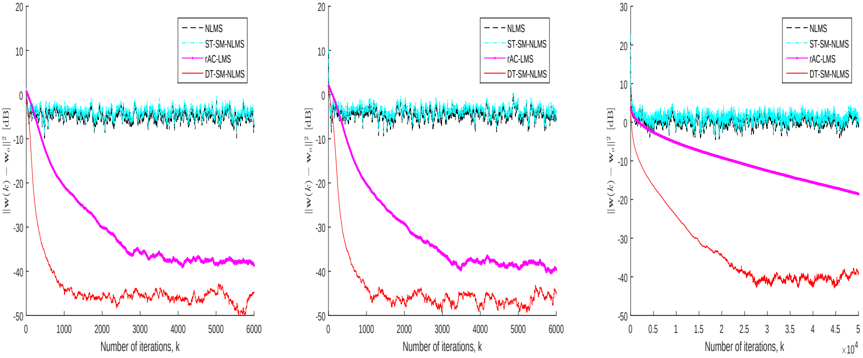
- Algorithms tested: NLMS, rAC-LMS, ST-SM-NLMS, DT-SM-NLMS algorithms;
- Input signals: BPSK, zero-mean white Gaussian noise with unit variance (WGN), AR(1) (first-order autoregressive);
- Filter order: $N = 29$ ;
- $\mathbf{w}(0) = [0, \cdots, 0]^T$;
- SNR: 20 dB;
- Regularization factor: $\delta = 10^{-12}$;
- 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 $\overline{\gamma}$=0.1875, 0.1477, 0.1194, respectively;
- The second threshold for DT-SM-NLMS algorithm: $\overline{\gamma}' = 1$;
- Outlier signal: Bernoulli process takes 1 with probability 0.05, multiplying $\mathcal{U}(0, 50)$.

SMT
Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

UFRJ

# Misalignment curves

- For $p = 0.1$, $\overline{\gamma}' = 1$ and $\overline{\gamma}$ is estimated as $\overline{\gamma} = 0.1875$.




(a) (b) (c)

Figure: (a) BPSK; (b) WGN; (c) AR(1).

Sinais, Multimídia e Telecomunicações
COPPE/Poli/UFRJ

# Update rates of the DT-SM-NLMS algorithm

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

| Input signal | $p = 0.1$ | $p = 0.2$ | $p = 0.3$ |
|---|---|---|---|
| BPSK | 0.090 | 0.188 | 0.292 |
| WGN | 0.091 | 0.190 | 0.293 |
| AR(1) | 0.099 | 0.196 | 0.299 |

- 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

# 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!