Normalized Least-Mean-Square Algorithms with Minimax Concave Penalty

Hiroyuki Kaneko, Masahiro Yukawa

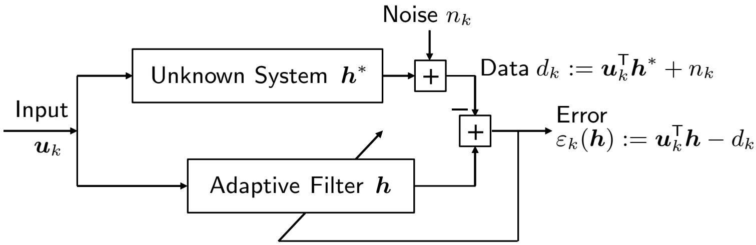
Dept. EEE, Keio University, Japan

ICASSP 2020, May 4-8 2020

Outline

Outline

Adaptive Filtering Problem



($k \in \mathbb{N}$: Time)

Goal

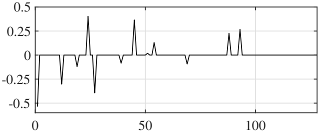
Estimate $\boldsymbol{h}^*$ by $\boldsymbol{h}_k := [h_1^{(k)}, h_2^{(k)}, \cdots, h_N^{(k)}]^{\mathsf{T}} \in \mathbb{R}^N$

Application

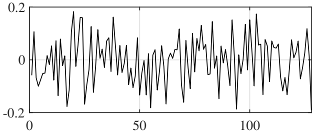
- Typical: Echo cancellation, Channel estimation
- Recent: Analysis of big data, Deep Learning

Motivation and Goal of This Study

In many applications, the unknown $\boldsymbol{h}^*$ is sparse. (e.g. echo path, comm. channel)



Sparse signal



Dense signal

$\rightarrow$ Sparsity-aware adaptive filtering algorithms

	Penalty	Estimation bias
Previous algorithms	convex (based on the ℓ_1 norm)	large
Proposed algorithms	weakly convex	small

Large biases come from underestimates of large-amplitude components.

Goal of This Study

Reduce the estimation bias with $\underbrace{\text{the minimax concave penalty}}_{\text{weakly convex}}$

Related Works

Convex Penalty (based on the ℓ_1 norm)

- ZA-LMS (Y. Chen, Y. Gu, and A. O. Hero, '09)
- OSCD, OCCD (D. Angelosante, J. A. Bazerque, and G. B. Giannakis, '10)
- APFBS (Y. Murakami, M. Yamagishi, M. Yukawa, and I. Yamada, '10)
- SPARLS (B. Babadi, N. Kalouptsidis, and V. Tarokh, '10)
- Prox-SVRG (L. Xiao and T. Zhang, '14)
- ASVB-MPL (K. E. Themelis, A. A. Rontogiannis, and K. D. Koutroumbas, '14)
- SVRG-ADMM (S. Zheng and J. T. Kwok, '16)
- SDA with linear-convergence-rate guarantees (N. Flammarion and F. Bach, '17)
- S-FM-HSDM(CRegLS) (K. Slavakis, '18)

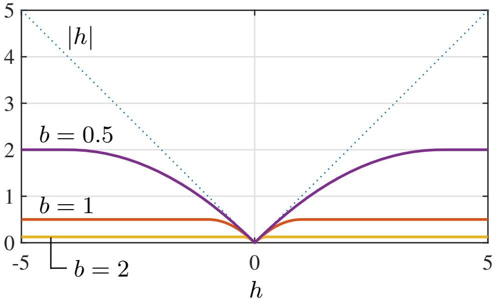
Nonconvex Penalty (Nonconvex Cost)

- ℓ_0-NLMS (Y. Gu, J. Jin, and S. Mei, '09)
- ℓ_0-RLS (E. M. Eksioglu and A. K. Tanc, '11)
- ℓ_p-norm penalized LMS (O. Taheri and S. A. Vorobyov, '11)
- Stochastic MM (E. Chouzenoux and J. C. Pesquet, '17)

Minimax Concave (MC) Penalty

MC Penalty Function

$$\psi_{\mathrm{MC}}(\boldsymbol{h}) := \|\boldsymbol{h}\|_1 - \min_{\boldsymbol{x}\in\mathbb{R}^N}\left(\|\boldsymbol{x}\|_1 + \frac{b^2}{2}\|\boldsymbol{h}-\boldsymbol{x}\|_2^2\right), \quad \boldsymbol{h}\in\mathbb{R}^N,\ b>0 \quad (1)$$



ψ_{MC}: Constant for large-amplitude components
$\rightarrow$ Reduce the estimation bias

C. H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," The Annals of statistics, 2010.
I. Selesnick, "Sparse regularization via convex analysis," IEEE Trans. Signal Processing, 2017.

Generalized Proximity Operator

Definition 1

- Given a possibly nonconvex function $\psi : \mathbb{R}^N \to (-\infty, +\infty]$, define

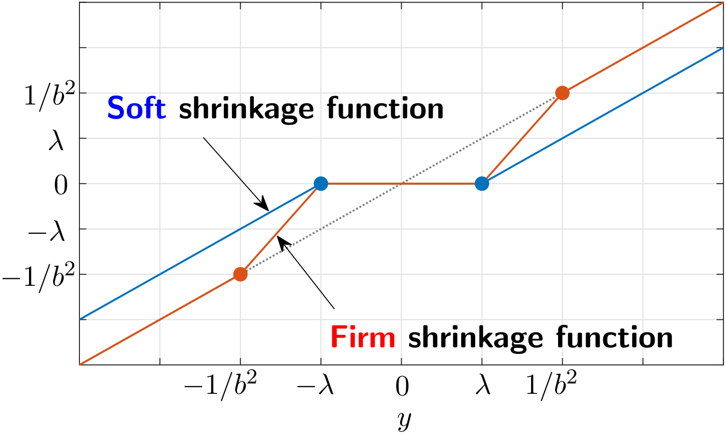
$$T_{\alpha\psi}(\boldsymbol{x}) := \arg\min_{\boldsymbol{y}\in\mathbb{R}^N} \left(\psi(\boldsymbol{y}) + \frac{1}{2\alpha}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\right), \quad \alpha > 0, \qquad (2)$$

  as long as the minimizer exists uniquely.

- $T_{\alpha\psi} = \mathrm{prox}_{\alpha\psi}$ (prox: Proximity operator) if ψ is proper l.s.c. convex.

I. Bayram, "On the convergence of the iterative shrinkage/thresholding algorithm with a weakly convex penalty," IEEE Trans. Signal Processing, 2015.

Shrinkage Functions



- $\mathrm{prox}_{\lambda\|\cdot\|_1}(y) = \mathrm{soft}(y;\lambda)$
- $T_{\lambda\psi_{\mathrm{MC}}}(y) = \mathrm{firm}(y;\lambda,1/b^2)$

Outline

Proposed Formulation

Instantaneous Cost Function

$$\underbrace{J_k(\boldsymbol{h})}_{\text{Instantaneous cost}} := \underbrace{f_k(\boldsymbol{h})}_{\text{Instantaneous loss}} + \underbrace{\lambda\psi_{\mathrm{MC}}(\boldsymbol{h})}_{\text{Penalty}}, \quad \lambda > 0 \quad (3)$$

$$f_k(\boldsymbol{h}) := \frac{(\boldsymbol{u}_k^{\mathsf{T}}\boldsymbol{h} - d_k)^2}{2\|\boldsymbol{u}_k\|_2^2} = \frac{1}{2}\left(\tilde{\boldsymbol{u}}_k^{\mathsf{T}}(\boldsymbol{h} - \boldsymbol{h}^*) - \tilde{n}_k\right)^2$$

$$\tilde{\boldsymbol{u}}_k := \frac{\boldsymbol{u}_k}{\|\boldsymbol{u}_k\|_2}$$

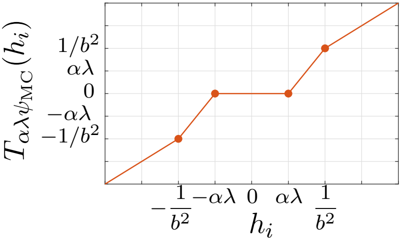
$$\tilde{n}_k := \frac{n_k}{\|\boldsymbol{u}_k\|_2}$$

Key Point

(i) Cost $J(\boldsymbol{h}) := \mathrm{E}[J_k(\boldsymbol{h})]$: convex under $\lambda b^2 \leq \lambda_{\min}(\boldsymbol{R})$ $(\boldsymbol{R} := \mathrm{E}[\tilde{\boldsymbol{u}}_k\tilde{\boldsymbol{u}}_k^{\mathsf{T}}])$

(ii) Instantaneous Cost $J_k(\boldsymbol{h})$: nonconvex for all $\lambda, b > 0$

Firm-Shrinkage NLMS (FS-NLMS) Algorithm (1/2)



$$J(\boldsymbol{h}) = f(\boldsymbol{h}) + \lambda\psi_{\rm MC}(\boldsymbol{h}), \quad f(\boldsymbol{h}) := {\rm E}[f_k(\boldsymbol{h})] \tag{4}$$

Properties (I. Bayram, '15)

Under $0 < \alpha < 2/(\lambda_{\max}(\boldsymbol{R}) + \lambda b^2)$ and $0 \leq \lambda b^2 \leq \lambda_{\min}(\boldsymbol{R})$ ($\Rightarrow 1/b^2 > \alpha\lambda$),

- $T_{\alpha\lambda\psi_{\rm MC}}((1-\alpha\lambda b^2)\cdot)$ is a 1/2-averaged nonexpansive mapping and $(1-\alpha\lambda b^2)^{-1}({\rm Id} - \alpha\nabla f)$ is a β-averaged nonexpansive mapping, where $\beta := \alpha(\lambda_{\max}(\boldsymbol{R}) + \lambda b^2)/2$.
- The composite mapping $T_{\alpha\lambda\psi_{\rm MC}} \circ ({\rm Id} - \alpha\nabla f)$ is averaged nonexpansive.
- $J(\boldsymbol{h})$ can be minimized by the Krasnosel'skiĭ-Mann iteration.

Firm-Shrinkage NLMS (FS-NLMS) Algorithm (2/2)

FS-NLMS Algorithm

For an arbitrarily chosen initial vector $\boldsymbol{h}_0 \in \mathbb{R}^N$, generate a sequence $(\boldsymbol{h}_k)_{k\in\mathbb{N}}$ by:

$$\begin{aligned} \boldsymbol{h}_{k+1} :=& T_{\alpha\lambda\psi_{\mathrm{MC}}} \circ (\mathrm{Id} - \alpha\nabla f_k)(\boldsymbol{h}_k) \\ =& T_{\alpha\lambda\psi_{\mathrm{MC}}} \left(\boldsymbol{h}_k - \frac{\alpha}{\|\boldsymbol{u}_k\|_2^2} (\boldsymbol{u}_k^{\mathsf{T}} \boldsymbol{h}_k - d_k) \boldsymbol{u}_k \right) \end{aligned} \quad (5)$$

$\alpha \in (0, 2)$: Step size

Remark 1

ψ_{MC} is nonconvex, although FS-NLMS resembles APFBS (Murakami, Yamagishi, Yukawa, and Yamada, '10).

Twin-Shrinkage NLMS (TS-NLMS) Algorithm (1/2)

Moreau Decomposition

$$(f \boxdot q) + (f^* \boxdot q) = q \qquad (\boxdot\text{: Infimal Convolution, } q(\boldsymbol{h}) := \tfrac{1}{2}\|\boldsymbol{h}\|_2^2)$$

By Moreau decomposition, provided $J(\boldsymbol{h})$ is convex ($\lambda b^2 \leq \lambda_{\min}(\boldsymbol{R})$),

$$\overbrace{J(\boldsymbol{h})}^{\text{convex}} = \overbrace{f(\boldsymbol{h})}^{\text{convex \& smooth}} + \overbrace{\lambda\psi_{\mathrm{MC}}(\boldsymbol{h})}^{\text{nonconvex}} \tag{6}$$

(7)

- Need to use the instantaneous nonconvex J_k
- No use of the firm shrinkage

Twin-Shrinkage NLMS (TS-NLMS) Algorithm (1/2)

Moreau Decomposition

$$(f \boxdot q) + (f^* \boxdot q) = q \qquad (\boxdot\text{: Infimal Convolution, } q(\boldsymbol{h}) := \tfrac{1}{2}\|\boldsymbol{h}\|_2^2)$$

By Moreau decomposition, provided $J(\boldsymbol{h})$ is convex $(\lambda b^2 \leq \lambda_{\min}(\boldsymbol{R}))$,

$$\overbrace{J(\boldsymbol{h})}^{\text{convex}} = \overbrace{f(\boldsymbol{h})}^{\text{convex \& smooth}} + \overbrace{\lambda\psi_{\mathrm{MC}}(\boldsymbol{h})}^{\text{nonconvex}} \tag{6}$$

$$= \underbrace{f(\boldsymbol{h}) + \lambda g(\boldsymbol{h})}_{\text{convex \& smooth}} + \underbrace{\lambda\|\boldsymbol{h}\|_1}_{\text{convex \& nonsmooth}} \tag{7}$$

$$g(\boldsymbol{h}) := b^2 \left(\left(\frac{1}{b^2} \|\cdot\|_1 \right)^* \boxdot q - q \right)(\boldsymbol{h})$$

- Need to use the instantaneous nonconvex J_k
- No use of the firm shrinkage

Twin-Shrinkage NLMS (TS-NLMS) Algorithm (2/2)

TS-NLMS Algorithm

For an arbitrarily chosen initial vector $\boldsymbol{h}_0 \in \mathbb{R}^N$, generate a sequence $(\boldsymbol{h}_k)_{k\in\mathbb{N}}$ by:

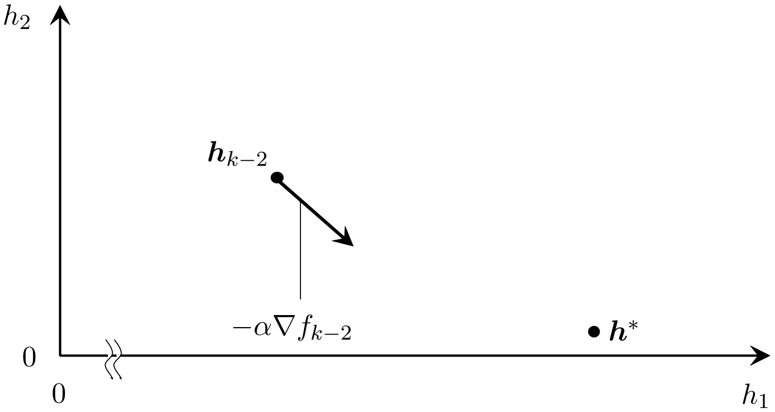
$$\boldsymbol{h}_{k+1} := \sum_{i=1}^{N} \mathrm{soft}\left(h_i^{(k)} - \alpha[\nabla(f_k + \lambda g)(\boldsymbol{h}_k)]_i\,;\ \alpha\lambda\right)\boldsymbol{e}_i \tag{8}$$

- $\alpha \in (0, 2)$: Step size
- $\{\boldsymbol{e}_i\}_{i=1}^{N}$: Standard basis

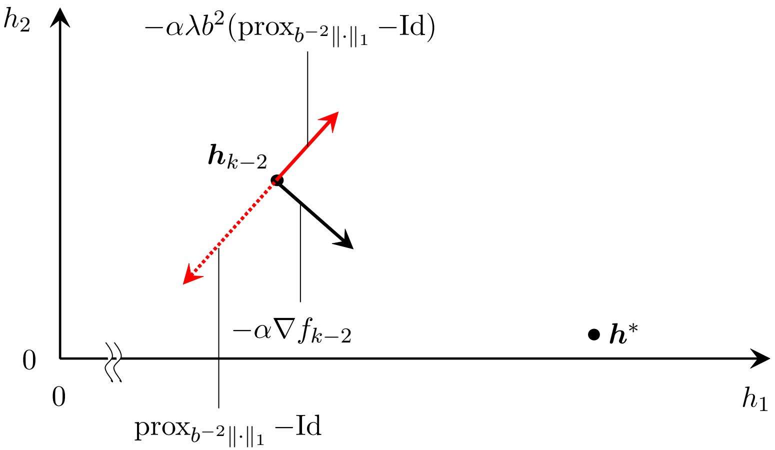
By Moreau decomposition $\nabla(f^* \boxdot q) = \mathrm{Id} - \mathrm{prox}_{f^*} = \mathrm{prox}_f$,

$$\nabla(f_k + \lambda g)(\boldsymbol{h}_k) = \frac{(\boldsymbol{u}_k^\mathsf{T}\boldsymbol{h}_k - d_k)\boldsymbol{u}_k}{\|\boldsymbol{u}_k\|_2^2} + \lambda b^2\left(\sum_{i=1}^{N}\mathrm{soft}\left(h_i^{(k)}; \frac{1}{b^2}\right)\boldsymbol{e}_i - \boldsymbol{h}_k\right) \tag{9}$$

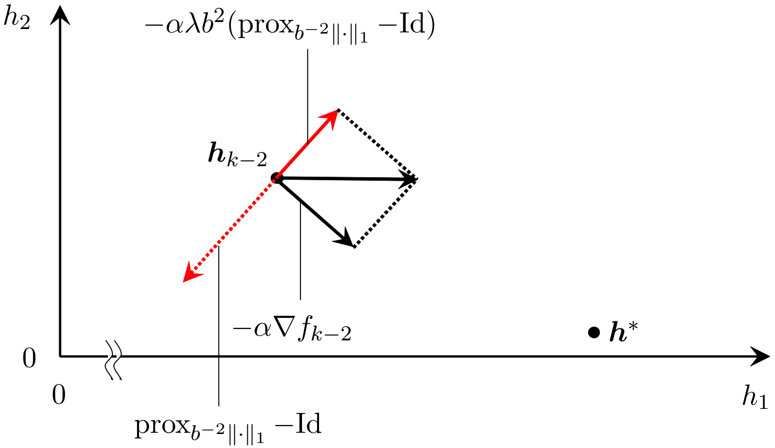
Geometric Interpretation of TS-NLMS



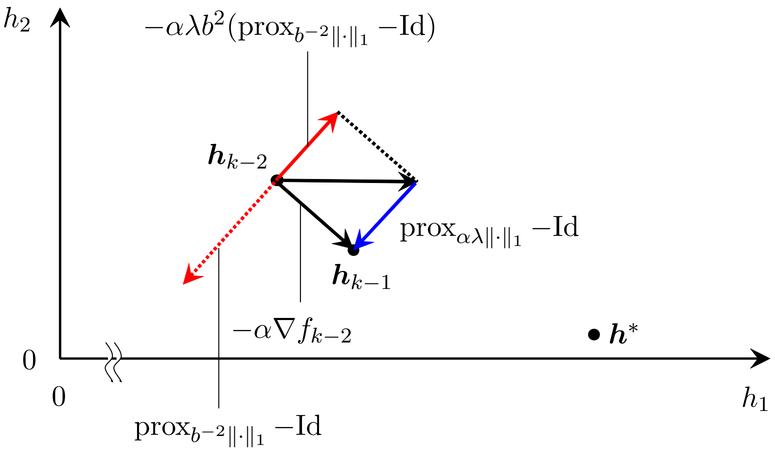
Geometric Interpretation of TS-NLMS



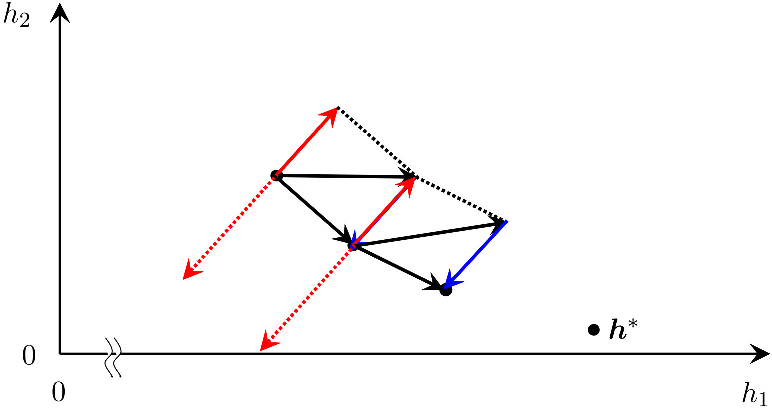
Geometric Interpretation of TS-NLMS



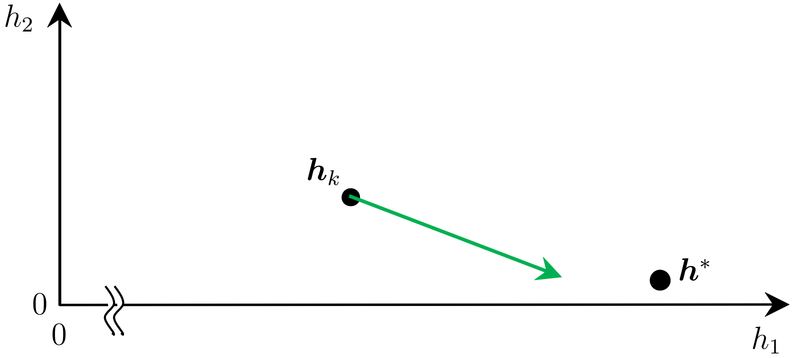
Geometric Interpretation of TS-NLMS



Geometric Interpretation of TS-NLMS

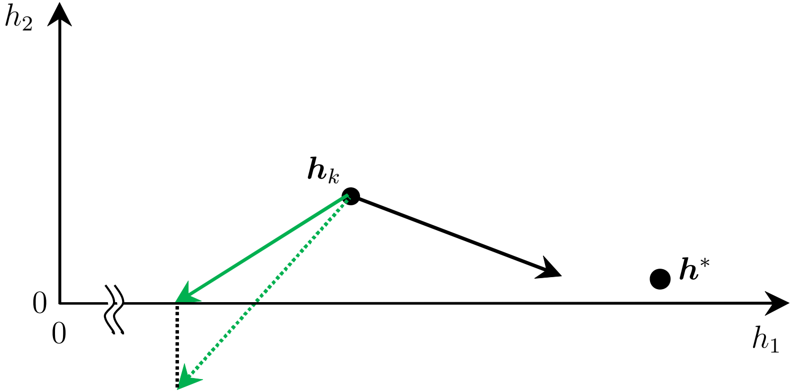


Step-by-step illustrations of the following iterations



$$\boldsymbol{h}_{k+1} := \sum_{i=1}^{2} \mathrm{soft}\left(h_i^{(k)} - \alpha[\nabla f_k(\boldsymbol{h}_k)]_i - \alpha\lambda b^2 \left(\mathrm{soft}\left(h_i^{(k)}; \frac{1}{b^2}\right) - h_i^{(k)}\right); \alpha\lambda\right)\boldsymbol{e}_i$$

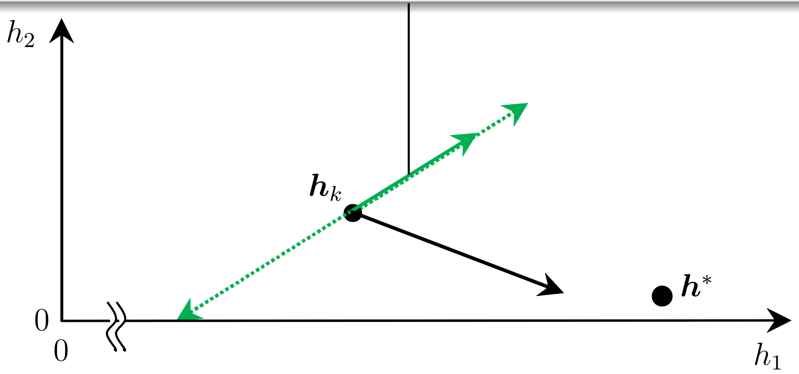
Step-by-step illustrations of the following iterations



$$\boldsymbol{h}_{k+1} := \sum_{i=1}^{2} \mathrm{soft}\left(h_i^{(k)} - \alpha[\nabla f_k(\boldsymbol{h}_k)]_i - \alpha\lambda b^2 \left(\mathrm{soft}\left(h_i^{(k)}; \frac{1}{b^2}\right) - h_i^{(k)}\right); \alpha\lambda\right)\boldsymbol{e}_i$$

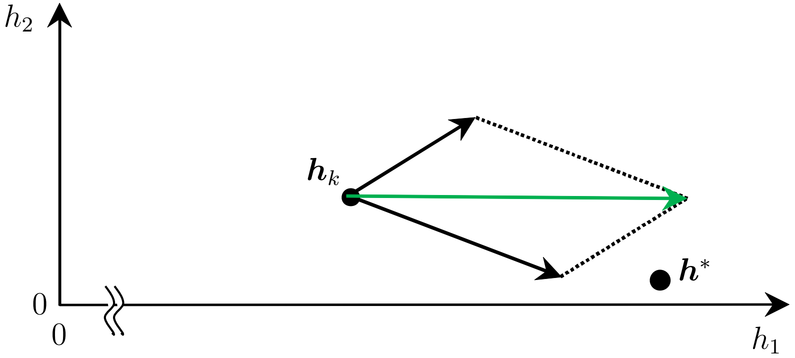
Geometric Interpretation of TS-NLMS

The vector inclines to the h_1 axis. $\rightarrow$ Reduce the biases by compensating the shrinkage in the previous iteration more for the large components



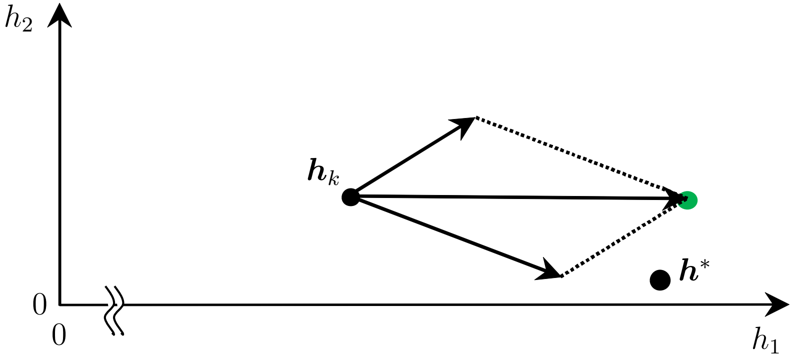
$$\boldsymbol{h}_{k+1} := \sum_{i=1}^{2} \mathrm{soft}\left(h_i^{(k)} - \alpha[\nabla f_k(\boldsymbol{h}_k)]_i - \alpha\lambda b^2\left(\mathrm{soft}\left(h_i^{(k)}; \frac{1}{b^2}\right) - h_i^{(k)}\right); \alpha\lambda\right)\boldsymbol{e}_i$$

Step-by-step illustrations of the following iterations



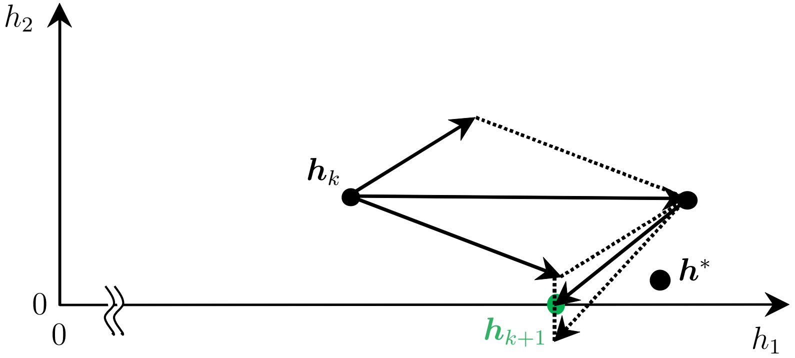
$$\boldsymbol{h}_{k+1} := \sum_{i=1}^{2} \text{soft}\left(h_i^{(k)} - \alpha[\nabla f_k(\boldsymbol{h}_k)]_i - \alpha\lambda b^2 \left(\text{soft}\left(h_i^{(k)}; \frac{1}{b^2}\right) - h_i^{(k)}\right); \alpha\lambda\right) \boldsymbol{e}_i$$

Step-by-step illustrations of the following iterations



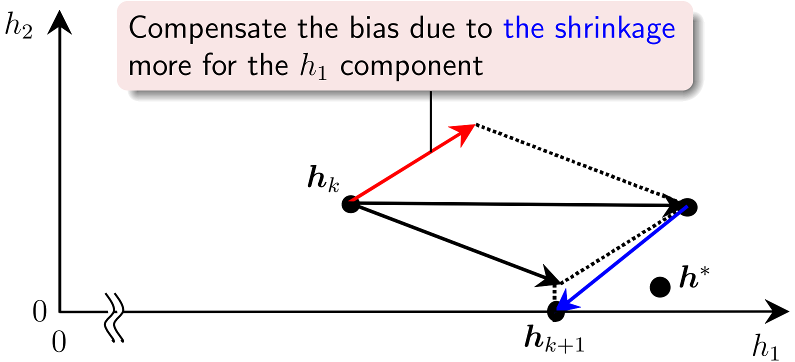
$$\boldsymbol{h}_{k+1} := \sum_{i=1}^{2} \mathrm{soft}\left(h_i^{(k)} - \alpha[\nabla f_k(\boldsymbol{h}_k)]_i - \alpha\lambda b^2\left(\mathrm{soft}\left(h_i^{(k)}; \frac{1}{b^2}\right) - h_i^{(k)}\right); \alpha\lambda\right)\boldsymbol{e}_i$$

Step-by-step illustrations of the following iterations

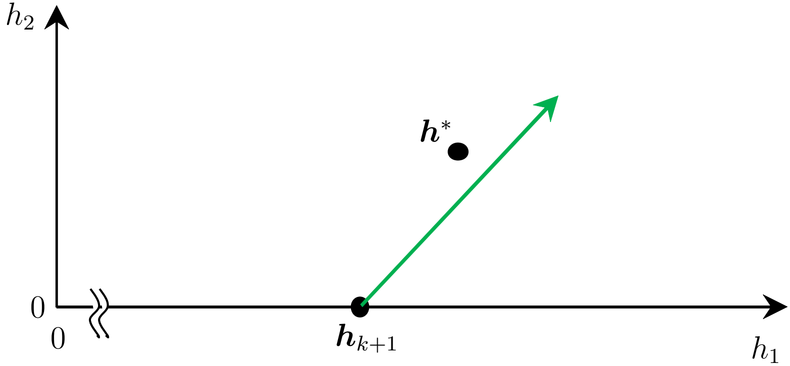


$$\boldsymbol{h}_{k+1} := \sum_{i=1}^{2} \mathrm{soft}\left(h_i^{(k)} - \alpha[\nabla f_k(\boldsymbol{h}_k)]_i - \alpha\lambda b^2 \left(\mathrm{soft}\left(h_i^{(k)}; \frac{1}{b^2}\right) - h_i^{(k)}\right); \alpha\lambda\right)\boldsymbol{e}_i$$

Step-by-step illustrations of the following iterations

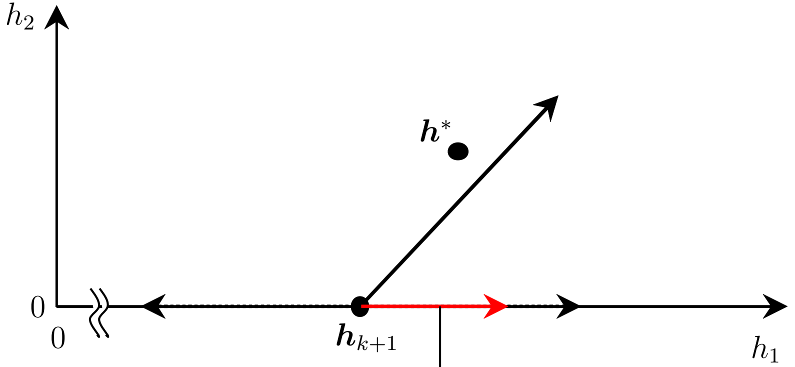


Step-by-step illustrations of the following iterations



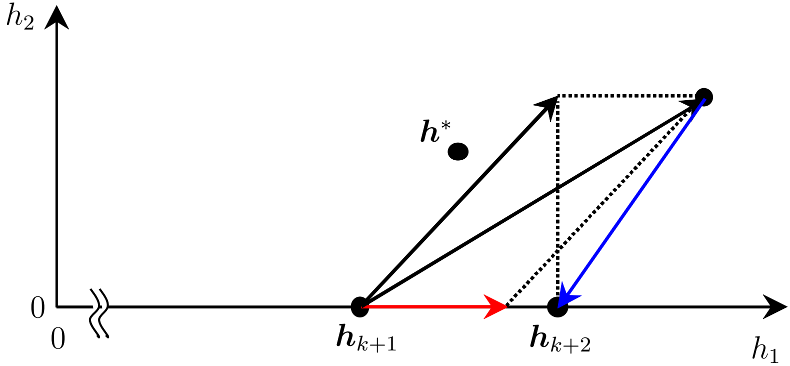
$$\boldsymbol{h}_{k+2} := \sum_{i=1}^{2} \mathrm{soft}\left(h_i^{(k+1)} - \alpha [\nabla f_{k+1}(\boldsymbol{h}_{k+1})]_i - \alpha\lambda b^2 \left(\mathrm{soft}\left(h_i^{(k+1)}; \frac{1}{b^2} \right) - h_i^{(k+1)} \right); \alpha\lambda \right) \boldsymbol{e}_i$$

Step-by-step illustrations of the following iterations

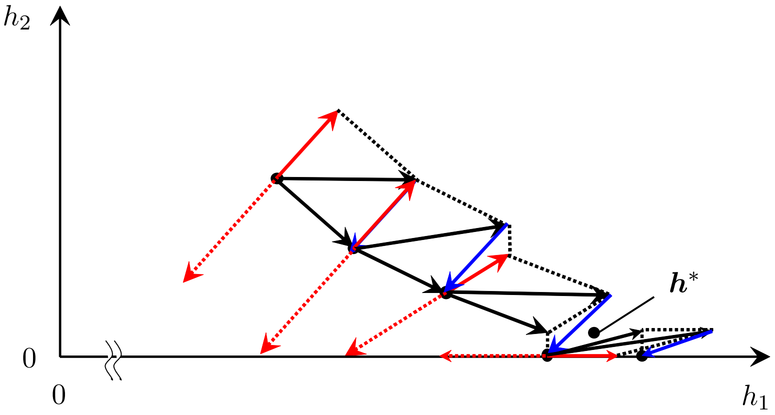


- Debias $h_1 \to$ small bias
- Preserve $h_2 = 0 \to$ high sparsity

Step-by-step illustrations of the following iterations



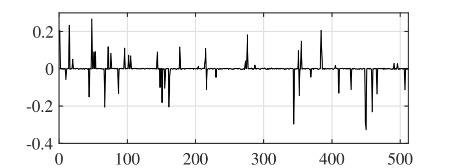
A big picture of the whole iterations



Outline

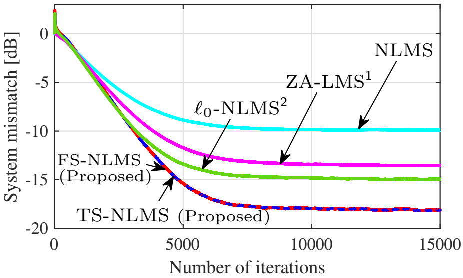
Simulation Setup

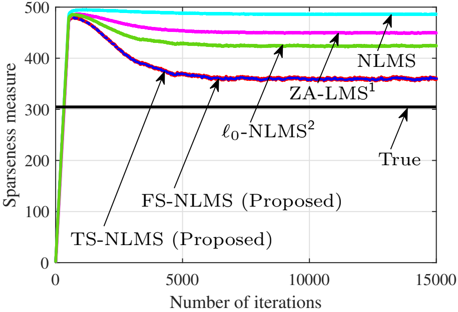
- $\boldsymbol{h}^*$: Unit weakly-sparse vector with 10% large-amplitude components ($N = 512$)
  $\boldsymbol{h}^* := \tilde{\boldsymbol{h}}^* / \|\tilde{\boldsymbol{h}}^*\|_2$, $\tilde{\boldsymbol{h}}^* := \boldsymbol{x} + \epsilon \boldsymbol{y}$, $\boldsymbol{x}$: Sparse vector, $\boldsymbol{y} \sim N(0,1)$, $\epsilon := 0.01$



- $\boldsymbol{u}_k := [u_k, u_{k-1}, \cdots, u_{k-N+1}]^{\mathsf{T}}$
- $\{u_k\}_{k=1}^{1.5\times 10^4}$: AR signal normalized to unit variance
  $u_k := \hat{u}_k / \operatorname{var}(\hat{u}_k)$, $\hat{u}_k := \gamma \hat{u}_{k-1} + v_k$, $\gamma := 0.8$, v_k: i.i.d.
- SNR 10 dB
- λ: Use the best parameter in terms of system mismatch
- Averaged over 300 simulation runs

Simulation Results





System Mismatch

$$\eta(\boldsymbol{h}_k) := 10\log_{10}\frac{\|\boldsymbol{h}^* - \boldsymbol{h}_k\|_2^2}{\|\boldsymbol{h}^*\|_2^2}$$

Sparseness

$$\xi(\boldsymbol{h}_k) := \sum_{i=1}^{N}\left(1 - \exp(-1000|h_i^{(k)}|)\right)$$

Remark 2

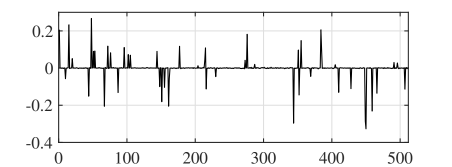
FS-NLMS and TS-NLMS yield sparser and less biased solutions.

[1] Y. Chen, Y. Gu, and A. O. Hero, "Sparse LMS for System Identification," in Proc. IEEE ICASSP, 2009.

[2] Y. Gu, J. Jin, and S. Mei, "ℓ_0 Norm Constraint LMS Algorithm for Sparse System Identification," IEEE Signal Processing Letters, 2009.

Simulation Setup

- $\boldsymbol{h}^*$: Unit weakly-sparse vector with 10% large-amplitude components $(N = 512)$
  $\boldsymbol{h}^* := \tilde{\boldsymbol{h}}^*/\|\tilde{\boldsymbol{h}}^*\|_2$, $\tilde{\boldsymbol{h}}^* := \boldsymbol{x} + \epsilon\boldsymbol{y}$, $\boldsymbol{x}$: Sparse vector, $\boldsymbol{y} \sim N(0,1)$, $\epsilon := 0.01$



- $\boldsymbol{u}_k := [u_k, u_{k-1}, \cdots, u_{k-N+1}]^\mathsf{T}$
- $\{u_k\}_{k=1}^{1.5\times10^4}$: AR signal normalized to unit variance
  $u_k := \hat{u}_k/\operatorname{var}(\hat{u}_k)$, $\hat{u}_k := \gamma\hat{u}_{k-1} + v_k$, $\gamma := 0.8$, v_k: i.i.d.
- SNR 10 dB
- λ: Use the best parameter in terms of system mismatch
- Averaged over 300 simulation runs

Outline

Conclusion

- We presented the FS-NLMS and TS-NLMS algorithms, which were sparsity-aware adaptive filtering algorithms based on the MC penalty.
- The whole cost function was convex under a proper choice of the parameters, while the instantaneous cost function was always nonconvex. This problem has not been studied so far to the best of our knowledge.
- The proposed algorithms outperformed the existing sparsity-aware adaptive filtering algorithms in system mismatch and sparseness of the solution.