Adaptive Projected Subgradient Method and its Acceleration Techniques

Isao Yamada, Nobuhiko Ogura, Masahiro Yukawa

Tokyo Institute of Technology

ALCOSP2004 @ Yokohama, August 31, 2004

Adaptive Projected Subgradient Method

A Paradigm Shift ?

Asymptotic Minimization of

Sequence of

unsmooth convex functions.

$\Downarrow$

1. Unified View of Adaptive Algorithms.

   Nonstationary adaptive filtering problems

2. Ideas for acceleration of convergence.

We consider

Problem

Minimize Asymptotically

$\underline{\Theta_n : \mathbb{R}^N \to [0,\infty), n \in \mathbb{N}}$, **sequence of continuous convex functions**

Subject to $\boldsymbol{h} \in K \left(\subset \mathbb{R}^N\right)$ closed convex set

where $\inf_{\boldsymbol{h} \in K} \Theta_n(\boldsymbol{h}) = 0$ fixed target value.

Note: Θ_n is **not necessarily a sample of random variable** of which mean value is widely used as an ideal cost function in standard adaptive filtering.

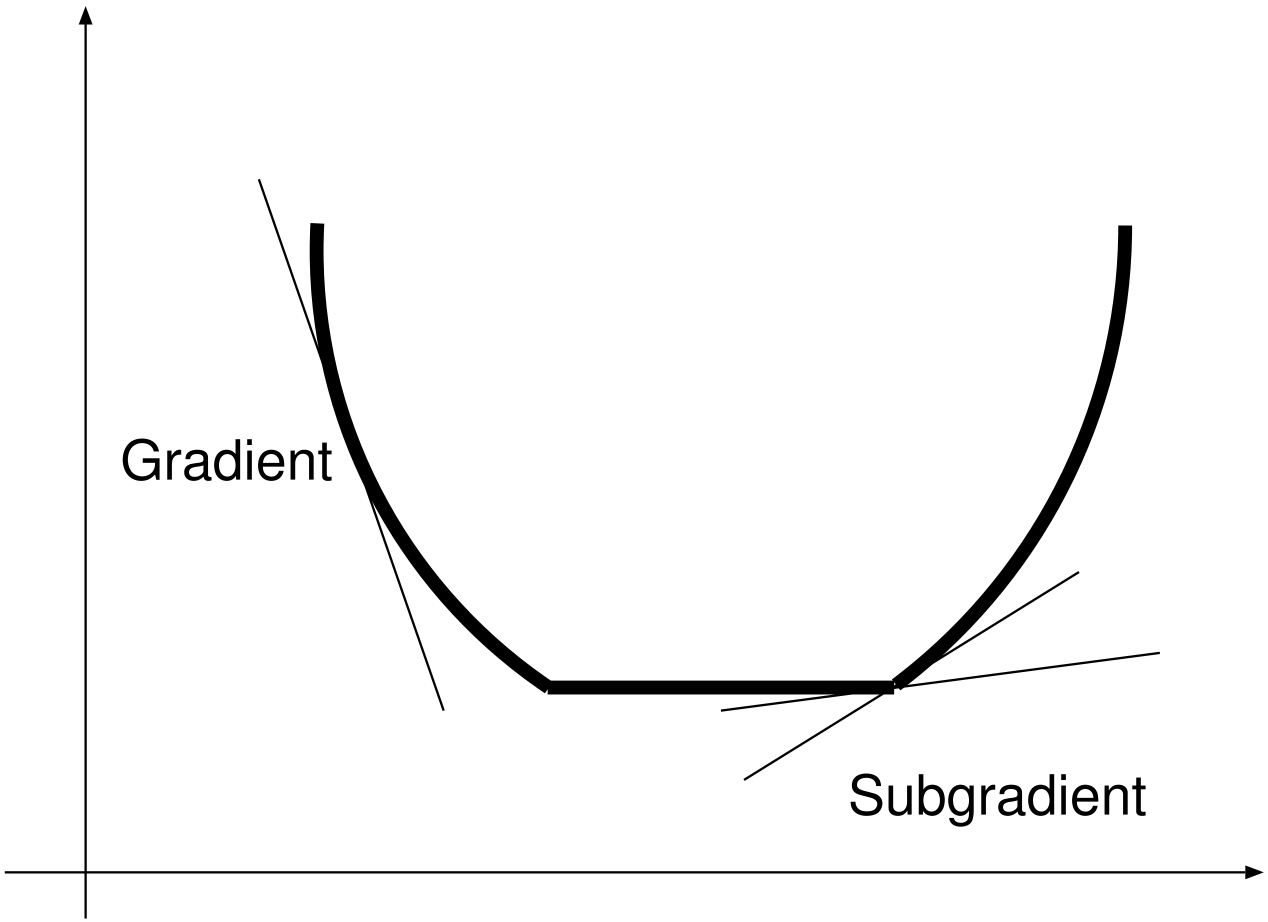
We propose

Adaptive Projected Subgradient Method

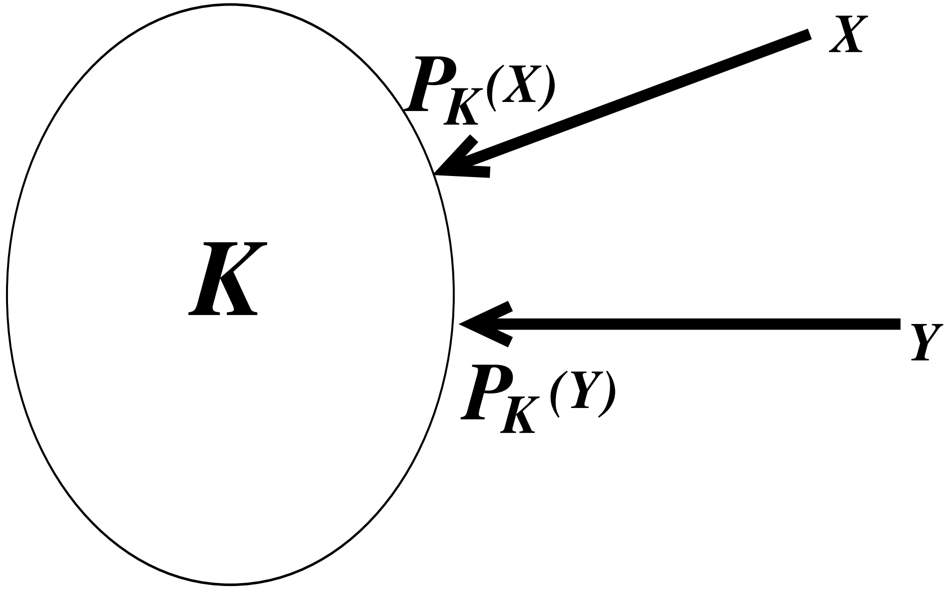
$$h_{n+1} := P_K \left(h_n - \alpha_n \frac{\Theta_n(h_n)}{\|\Theta_n'(h_n)\|^2} \Theta_n'(h_n) \right)$$

1. If $\Theta_n \Rightarrow \Theta$, it reproduces Polyak('69). Similar to the Newton's method but does not require any matrix inversion.
2. $\Theta_n : \mathbb{R}^N \to [0, \infty)$ must be continuous convex but not necessarily smooth.
3. Constraint $h_n \in K$ can be imposed by Convex Projection P_K.

Subgradient: a generalization of Gradient



What is Convex Projection ?



What can we say by

Adaptive Projected Subgradient Method

$$h_{n+1} := P_K\left(h_n - \alpha_n \frac{\Theta_n(h_n)}{\|\Theta_n'(h_n)\|^2}\Theta_n'(h_n)\right)$$

?

Theorem Adaptive Projected Subgradient Method

(a) (**Monotone approximation**) Suppose

$$\boldsymbol{h}_n \notin \Omega_n := \left\{ \boldsymbol{h} \in K \mid \Theta_n(\boldsymbol{h}) = \Theta_n^* := \min_{\boldsymbol{u} \in K} \Theta_n(\boldsymbol{u}) \right\}.$$

$$\Rightarrow \text{ By using } \forall \alpha_n \in \left(0, 2\left(1 - \frac{\Theta_n^*}{\Theta_n(\boldsymbol{h}_n)}\right)\right),$$

$$\left(\forall \boldsymbol{h}^{*(n)} \in \Omega_n\right) \quad \|\boldsymbol{h}_{n+1} - \boldsymbol{h}^{*(n)}\| < \|\boldsymbol{h}_n - \boldsymbol{h}^{*(n)}\|.$$

(b) (**Asymptotic minimization**)

Suppose $\Theta_n^* = 0$, $\forall n \geq N_0$ and $\Omega := \bigcap_{n \geq N_0} \Omega_n \neq \emptyset$.

$$\Rightarrow \text{ By } \alpha_n \in [\varepsilon_1, 2 - \varepsilon_2],$$

$$\lim_{n \to \infty} \Theta_n(\boldsymbol{h}_n) = 0 \text{ if } \left(\Theta_n'(\boldsymbol{h}_n)\right)_{n \in \mathbb{N}} \text{: bdd.}$$

(c) (**Convergence 1**) Under conditions in (b), if Ω has a relative interior w.r.t. some hyperplane,

$$\Rightarrow \lim_{n \to \infty} \boldsymbol{h}_n = \exists \widehat{\boldsymbol{h}} \in K.$$

Theorem (Contd.)

(d) (**Convergence 2**) In addition to conditions in (b), if Ω has an interior $\tilde{\boldsymbol{h}}$ satisfying

$$(\forall \varepsilon > 0, \forall r > 0, \exists \delta > 0) \quad \inf_{\substack{d(h_n, \mathrm{lev}_{\leq 0}\,\Theta_n) \geq \varepsilon, \\ \|\tilde{h} - h_n\| \leq r, \\ n \geq N_0}} \Theta_n(h_n) \geq \delta,$$

$$\Downarrow$$

$$\lim_{n\to\infty} \boldsymbol{h}_n = \exists \hat{\boldsymbol{h}} \in \overline{\liminf_{n\to\infty} \Omega_n},$$

where

$$\liminf_{n\to\infty} \Omega_n := \bigcup_{n \geq 0} \bigcap_{k \geq n} \Omega_k.$$

How can we design

Suitable Θ_n for

Adaptive Filtering ?

$$\Downarrow$$

Imagine Ideal Situations !

Example 1

Suppose $S_\iota^{(n)} \ni \boldsymbol{h}^*$ ($\iota \in \mathcal{J}_n$, $n \in \mathbb{N}$) with high probability.
Let $\Omega_n := K \cap \bigcap_{\iota \in \mathcal{J}_n} S_\iota^{(n)}$.

Define

$$\Theta_n(\boldsymbol{h}) := \max_{\iota \in \mathcal{J}_n} d(\boldsymbol{h}, S_\iota^{(n)}) \quad (\Leftarrow \text{unsmooth})$$

$$\boldsymbol{h}_{n+1} := P_K \left(\boldsymbol{h}_n + \mu_n \left(\sum_{\iota \in \mathcal{I}_n} \omega_\iota^{(n)} P_{S_\iota^{(n)}}(\boldsymbol{h}_n) - \boldsymbol{h}_n \right) \right),$$

$$\forall \mu_n \in [0, 2\mathcal{M}_n^{(0)}],$$

$$\mathcal{I}_n := \left\{ \iota \in \mathcal{J}_n : d(\boldsymbol{h}_n, S_\iota^{(n)}) \geq d(\boldsymbol{h}_n, S_k^{(n)}), \forall k \in \mathcal{J}_n \right\}.$$

$$\mathcal{M}_n^{(0)} := \begin{cases} \dfrac{\max_{\iota \in \mathcal{J}_n} \left\| P_{S_\iota^{(n)}}(\boldsymbol{h}_n) - \boldsymbol{h}_n \right\|^2}{\left\| \sum_{\iota \in \mathcal{J}_n} \omega_\iota^{(n)} P_{S_\iota^{(n)}}(\boldsymbol{h}_n) - \boldsymbol{h}_n \right\|^2}, & \boldsymbol{h}_n \notin \bigcap_{\iota \in \mathcal{J}_n} S_\iota^{(n)}, \\ 1, & \text{otherwise.} \end{cases}$$

Example 2

Suppose $S_\iota^{(n)} \ni \boldsymbol{h}^*$ $(\iota \in \mathcal{J}_n,\ n \in \mathbb{N})$ with high probability.

Let $\Omega_n := K \cap \bigcap_{\iota \in \mathcal{J}_n} S_\iota^{(n)}$.

Define

$$\Theta_n(\boldsymbol{h}) := \frac{1}{L} \sum_{\iota \in \mathcal{J}_n} \omega_\iota^{(n)} d(\boldsymbol{h}, S_\iota^{(n)})^2 \quad (\Leftarrow \textbf{smooth})$$

$$\boldsymbol{h}_{n+1} := P_K \left(\boldsymbol{h}_n + \mu_n \left(\sum_{\iota \in \mathcal{J}_n} \omega_\iota^{(n)} P_{S_\iota^{(n)}}(\boldsymbol{h}_n) - \boldsymbol{h}_n \right) \right),$$

$\forall \mu_n \in [0, 1\mathcal{M}_n^{(1)}]$,

$$\mathcal{M}_n^{(1)} := \begin{cases} \dfrac{\sum_{\iota \in \mathcal{J}_n} \omega_\iota^{(n)} \left\| P_{S_\iota^{(n)}}(\boldsymbol{h}_n) - \boldsymbol{h}_n \right\|^2}{\left\| \sum_{\iota \in \mathcal{J}_n} \omega_\iota^{(n)} P_{S_\iota^{(n)}}(\boldsymbol{h}_n) - \boldsymbol{h}_n \right\|^2}, & \boldsymbol{h}_n \notin \bigcap_{\iota \in \mathcal{J}_n} S_\iota^{(n)}, \\ 1, & \text{otherwise.} \end{cases}$$

Example 3

Suppose $S_\iota^{(n)} \ni \boldsymbol{h}^*$ ($\iota \in \mathcal{J}_n$, $n \in \mathbb{N}$) with high probability.

Let $\mathcal{I}_n := \{\iota \in \mathcal{J}_n : \boldsymbol{h}_n \notin S_\iota^{(n)}\}$ and $\Omega_n := K \cap \bigcap_{\iota \in \mathcal{I}_n} S_\iota^{(n)}$.

Define

$$\Theta_n(\boldsymbol{h}) := \frac{1}{L} \sum_{\iota \in \mathcal{I}_n} \omega_\iota^{(n)} d(\boldsymbol{h}_n, S_\iota^{(n)}) d(\boldsymbol{h}, S_\iota^{(n)}) \quad (\Leftarrow \text{Unsmooth})$$

$$\boldsymbol{h}_{n+1} := P_K\left(\boldsymbol{h}_n + \mu_n\left(\sum_{\iota \in \mathcal{J}_n} \omega_\iota^{(n)} P_{S_\iota^{(n)}}(\boldsymbol{h}_n) - \boldsymbol{h}_n\right)\right),$$

$\forall \mu_n \in [0, \mathbf{2}\mathcal{M}_n^{(1)}]$

$$\mathcal{M}_n^{(1)} := \begin{cases} \frac{\sum_{\iota \in \mathcal{J}_n} \omega_\iota^{(n)} \left\| P_{S_\iota^{(n)}}(\boldsymbol{h}_n) - \boldsymbol{h}_n \right\|^2}{\left\| \sum_{\iota \in \mathcal{J}_n} \omega_\iota^{(n)} P_{S_\iota^{(n)}}(\boldsymbol{h}_n) - \boldsymbol{h}_n \right\|^2}, & \boldsymbol{h}_n \notin \bigcap_{\iota \in \mathcal{J}_n} S_\iota^{(n)}, \\ 1, & \text{otherwise.} \end{cases}$$

$\Rightarrow$ **Adaptive Parallel Subgradient Projection**

(Yamada, Slavakis, Yamada 2002)

Let's focus on

Adaptive Parallel Subgradient Projection (Adaptive PSP)

$$h_{n+1} := h_n + \mu_n \left(\sum_{\iota \in \mathcal{J}_n} \omega_\iota^{(n)} P_{S_\iota^{(n)}}(h_n) - h_n \right)$$

Consider **Strategic Weight Design** for Acceleration of Convergence.

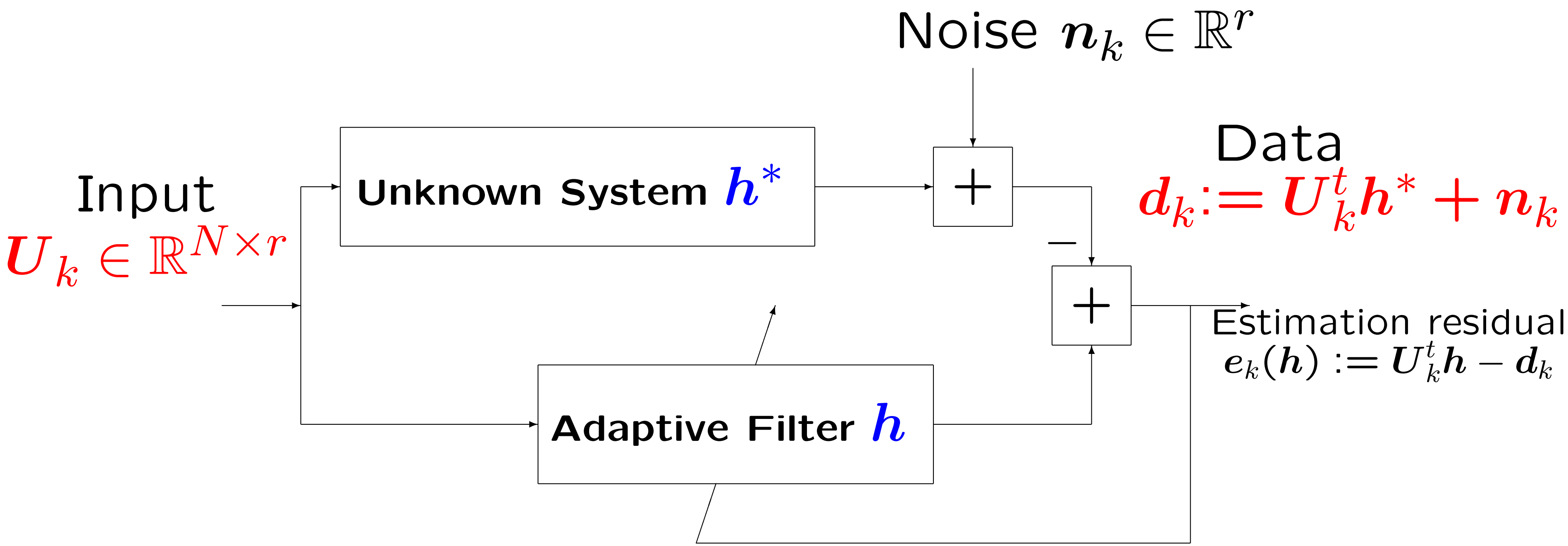
Think simple case where

$S_\iota^{(n)} = H_\iota^-$ $(\iota \in \mathcal{J}_n)$ are halfspaces.

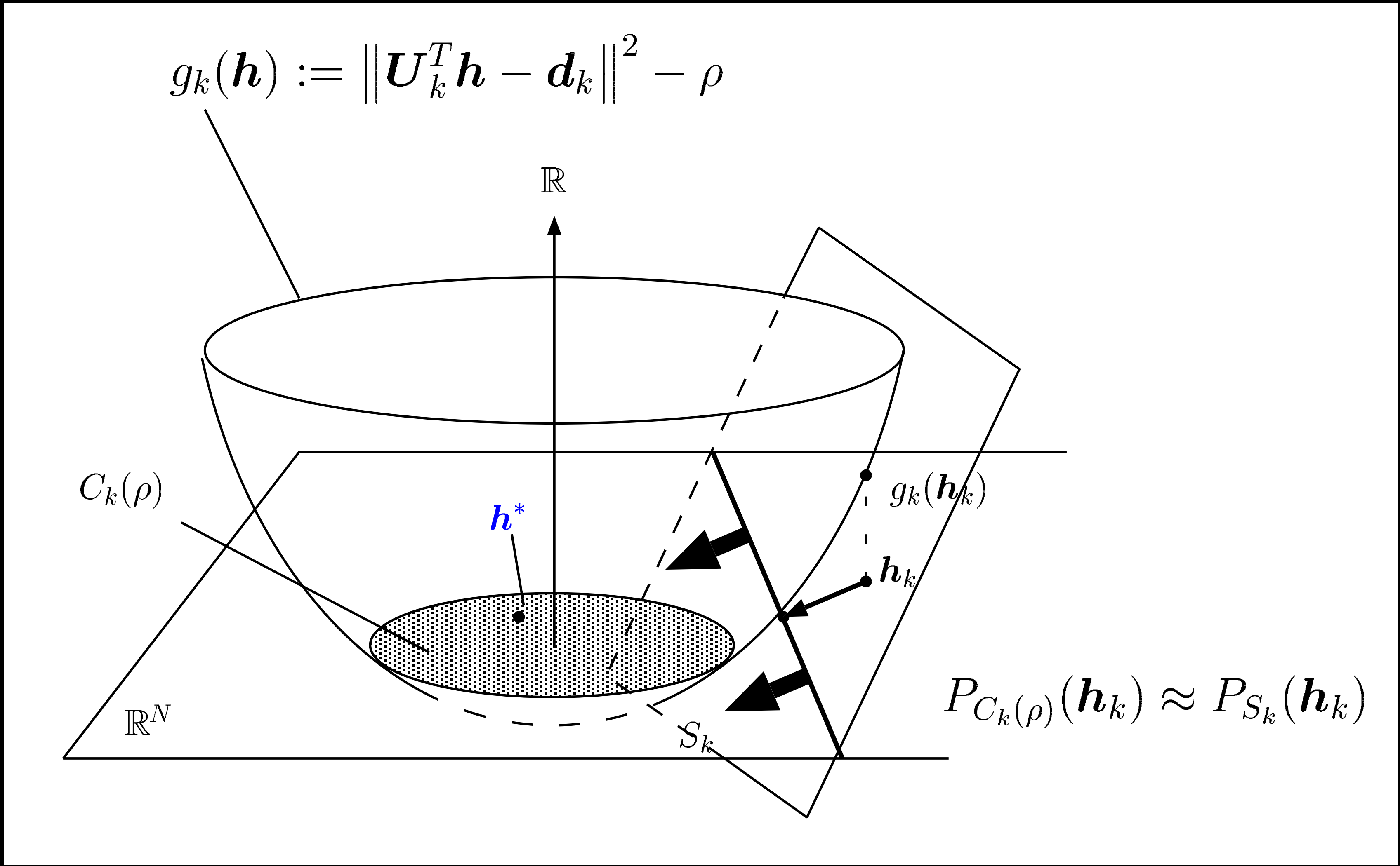
Adaptive Filtering Problem

Approximate $\boldsymbol{h} \in \mathbb{R}^N$ **to** $\boldsymbol{h}^* \in \mathbb{R}^N$

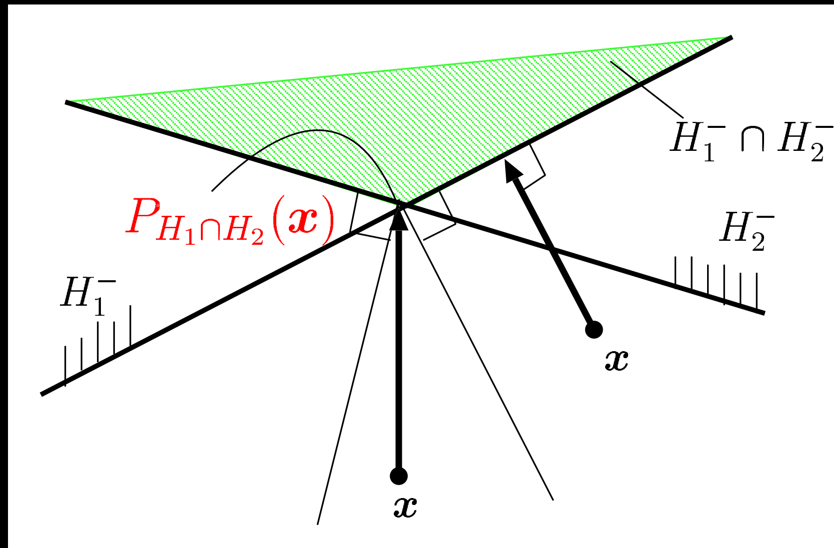
by using $(\boldsymbol{U}_k)_{k\in\mathbb{Z}}$ and $(\boldsymbol{d}_k)_{k\in\mathbb{Z}}$.



Subgradient Projection



POWER-PSP designs Weights by using inductively Projection onto Intersection of Two Half-Spaces

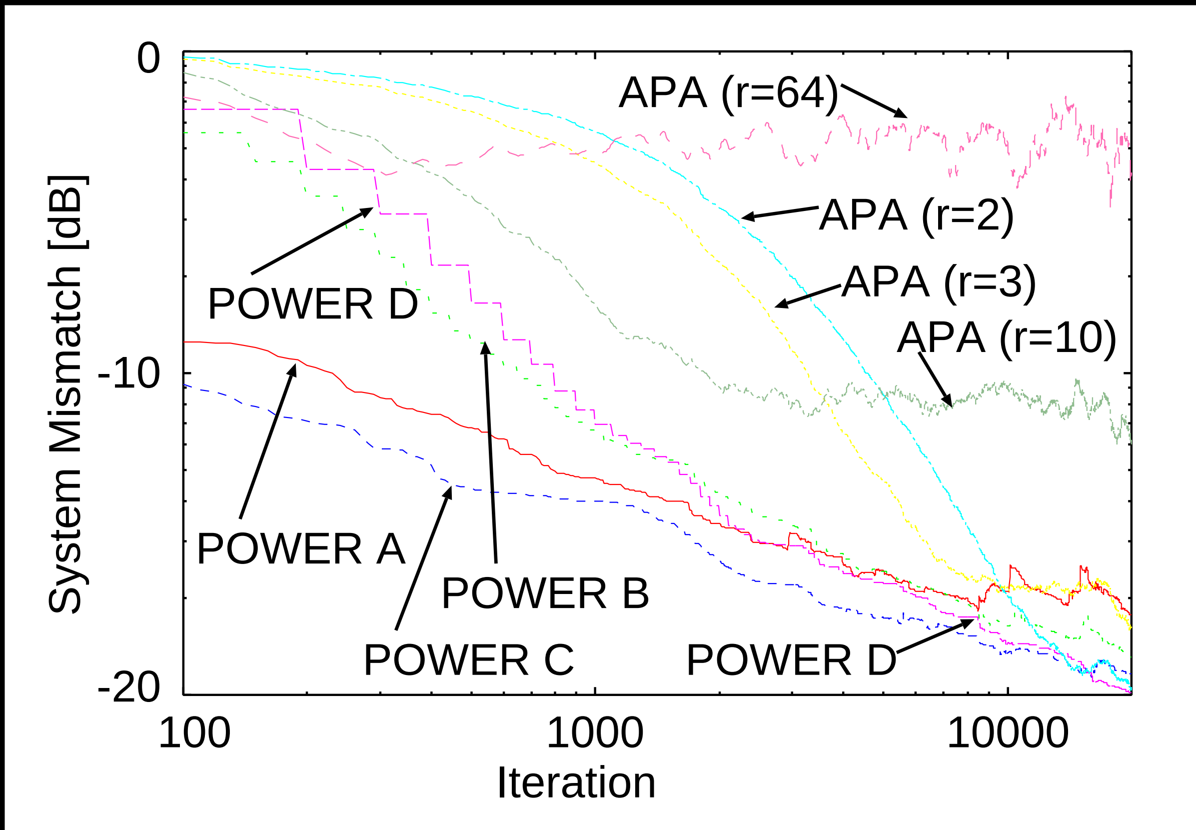


Proposition 1

For any $x \in \mathcal{H}$ and closed half-spaces $H_1^-, H_2^- \subset \mathcal{H}$,

$$P_{H_1^- \cap H_2^-}(\boldsymbol{x}) = \begin{cases} P_{H_1^-}(\boldsymbol{x}) & \text{if } P_{H_1^-}(\boldsymbol{x}) \in H_2^- \\ P_{H_2^-}(\boldsymbol{x}) & \text{if } P_{H_2^-}(\boldsymbol{x}) \in H_1^- \\ P_{H_1 \cap H_2}(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

POWER-PSP versus APA



$$\text{System Mismatch}(k) := 10\log_{10}\frac{\|\boldsymbol{h}^* - \boldsymbol{h}_k\|^2}{\|\boldsymbol{h}^*\|^2}$$

Conditions

- SNR=10dB

Parameters

- $q = 8$
- $\mu_k = \mathcal{M}_k$
- $\rho = \rho_1$
- $(r, R) =$

A:(64,1)

B:(64,50)

C:(200,10)

D:(200,100)

Conclusion

1. Adaptive Projected Subgradient Method

$$h_{n+1} := P_K\left(h_n - \alpha_n \frac{\Theta_n(h_n)}{\|\Theta'_n(h_n)\|^2}\Theta'_n(h_n)\right)$$

minimizes asymptotically

Θ_n, $n = 1, 2, \ldots$, **over** K**,**

2. This scheme unifies a wide range of adaptive algorithms.

EX. **NLMS, APA, Adaptive-PSP, Adaptive Min-max Projection** and **their Embedded Constraint versions: Constrained-NLMS, Constrained-APA**

3. Acceleration of Convergence

EX. **POWER-PSP**:

Pairwise Optimal WEighting Realization

1. I.Yamada, "Adaptive projected subgradient method — A unified view of projection based adaptive algorithms," (in Japanese), Journal of IEICE, **86**, 2003.
2. I.Yamada and N.Ogura "Adaptive projected subgradient method and its applications to set-theoretic adaptive filtering," Proc. 37th Asilomar Conf., Nov. 2003.
3. I.Yamada, K.Slavakis, and K. Yamada, "An efficient robust filtering algorithm based on parallel subgradient projection techniques," IEEE Trans. Signal Processing, vol.50, no.5, 2002.
4. K.Slavakis, I.Yamada and N.Ogura "Adaptive projected subgradient method and set-theoretic adaptive filtering with multiple convex constraints," Proc. 38th Asilomar Conf., 2004, to be presented.