The p-norm generalization of the LMS algorithm for adaptive filtering

Jyrki Kivinen
University of Helsinki

Manfred Warmuth
University of California, Santa Cruz

Babak Hassibi
California Institute of Technology

Least Mean Squares (LMS) update

Pick learning rate $\eta > 0$. Initialize $\boldsymbol{w}_0 = \mathbf{0} \in \mathbf{R^n}$

At time t, for $t = 1, \ldots, T$, the algorithm

- observes input $\boldsymbol{x}_t \in \mathbf{R^n}$
- makes prediction $\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t \in \mathbf{R}$
- observes feedback $y_t \in \mathbf{R}$, and
- updates its hypothesis as

$$\boldsymbol{w}_t = \boldsymbol{w}_{t-1} - \eta(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t - y_t)\boldsymbol{x}_t$$

Main Results

- Techniques from machine learning lead to generalizations of LMS
- H^∞-optimal filtering in signal processing is similar to relative on-line loss bounds in machine learning

Motivation

- Non-Gaussian modeling
- Get away from rotation invariant algorithms
- Develop algorithms that work well when instances orthogonal and target weight vectors "sparse"

Expected Bounds for LMS

- Assume $y_t = \boldsymbol{u} \cdot \boldsymbol{x}_t + \nu_t$, where ν_t iid with $\mathsf{E}[\nu_t^2] = \varepsilon$. Then

$$\mathsf{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{u}\cdot\boldsymbol{x}_t - \boldsymbol{w}_{t-1}\cdot\boldsymbol{x}_t)^2\right] \leq \varepsilon + \frac{1}{T}X_2^2||\boldsymbol{u}||_2^2$$

Better algorithms exist for probabilistic setting
However our goal is to weaken the assumptions

H^∞ bound for LMS [HSK96]

Assume $||\boldsymbol{x}_t||_2 \leq X_2$ for all t. Choose $\eta = 1/X_2^2$
For any $\boldsymbol{u} \in \mathbf{R^n}$

$$\frac{\sum_{t=1}^{T}(\boldsymbol{u}\cdot\boldsymbol{x}_t - \boldsymbol{w}_{t-1}\cdot\boldsymbol{x}_t)^2}{\sum_{t=1}^{T}(\boldsymbol{u}\cdot\boldsymbol{x}_t - y_t)^2 + X_2^2||\boldsymbol{u}||_2^2} \leq 1$$

- If some $\boldsymbol{u}$ with small norm is good predictor then LMS must approximate predictions of $\boldsymbol{u}$
- Bound holds for any $\boldsymbol{u}$ and $(\boldsymbol{x}_t, y_t)$
- No probabilistic assumptions
- LMS is H^∞-optimal:
  No algorithm can achieve ratio < 1
  $\forall \boldsymbol{u}$ and $(\boldsymbol{x}_t, y_t)$

Two related problems

A priori filtering: Control Theory

Try to match $\boldsymbol{u} \cdot \boldsymbol{x}_t$

$$\sum_{t=1}^{T} (\boldsymbol{u} \cdot \boldsymbol{x}_t - \boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t)^2$$

Prediction: On-line Learning

Try to match y_t

$$\sum_{t=1}^{T} (y_t - \boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t)^2$$

Comparison of known LMS-related bounds

- For $\eta = \alpha/X_2^2$,

$$\sum_{t=1}^{T}(\boldsymbol{u}\cdot\boldsymbol{x}_t - \boldsymbol{w}_{t-1}\cdot\boldsymbol{x}_t)^2 \;\leq\; \sum_{t=1}^{T}(\boldsymbol{u}\cdot\boldsymbol{x}_t - y_t)^2 + X_2^2||\boldsymbol{u}||_2^2$$

- For $\eta = \alpha/X_2^2$ $(0 < \alpha < 1)$

$$\sum_{t=1}^{T}(y_t - \boldsymbol{w}_{t-1}\cdot\boldsymbol{x}_t)^2 \leq \frac{1}{1-\alpha}\underbrace{\sum_{t=1}^{T}(\boldsymbol{u}\cdot\boldsymbol{x}_t - y_t)^2}_{Loss_{\boldsymbol{u}}} + \frac{1}{\alpha}X_2^2||\boldsymbol{u}||_2^2$$

$$\overset{\text{tuned } \alpha}{\leq} Loss_{\boldsymbol{u}} + 2\sqrt{Loss_{\boldsymbol{u}}}X_2||\boldsymbol{u}||_2 + X_2^2||\boldsymbol{u}||_2^2$$

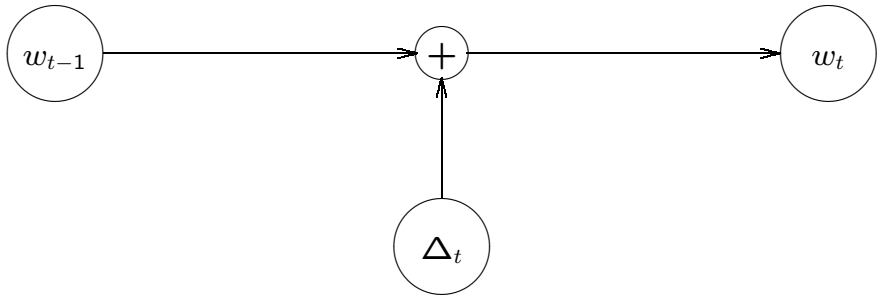
[CBLW96]

Generalizing the LMS bound

- Replace $||\boldsymbol{x}||_2||\boldsymbol{u}||_2$ by $||\boldsymbol{x}||_p||\boldsymbol{u}||_q$ where $1/p + 1/q = 1$ and $||\boldsymbol{x}||_p = (\sum_i |x_i|^p)^{1/p}$
- Instead of comparing predictions to $\boldsymbol{u} \cdot \boldsymbol{x}_t$ for a fixed target $\boldsymbol{u}$ compare to $\boldsymbol{u}_t \cdot \boldsymbol{x}_t$ where $\boldsymbol{u}_t$ may change
- Replace $(\cdot)^2$ by more general loss

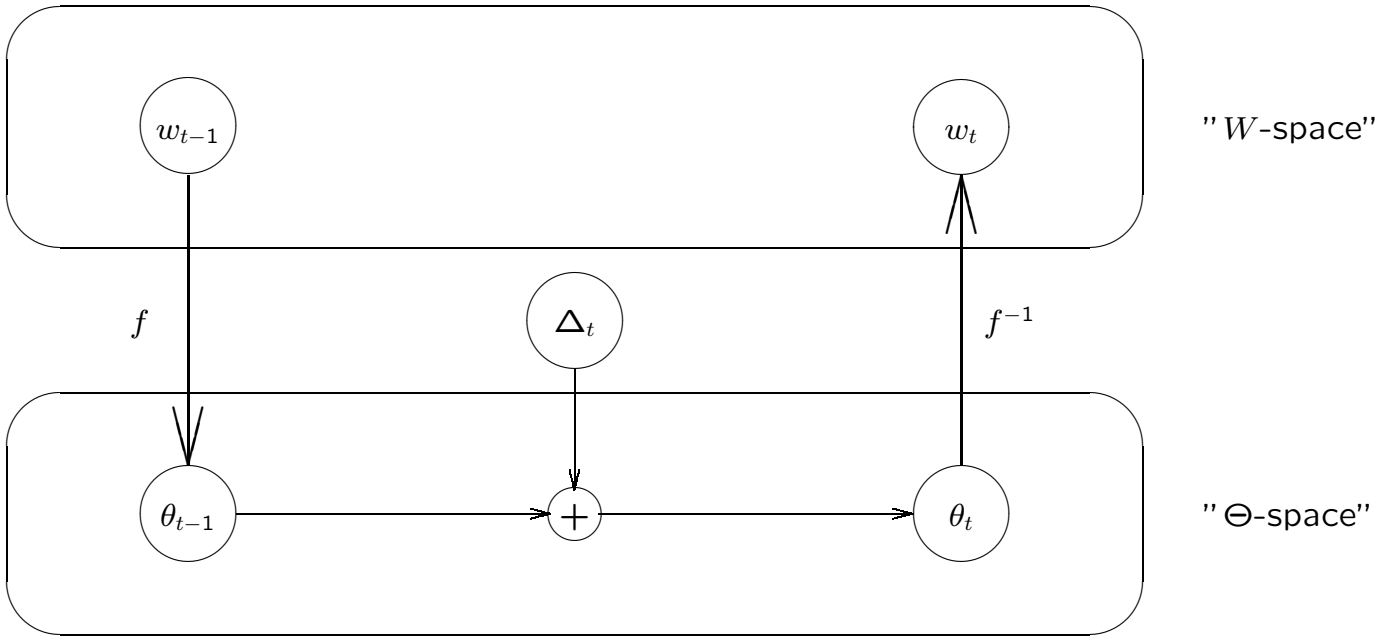
Basic LMS

$$\Delta_t = -\eta(w_{t-1} \cdot x_t - y_t)x_t$$



p-norm LMS

Write $\theta_t = f(w_t)$



p-norm LMS based on [GLS01]

$$\boldsymbol{w}_t = \boldsymbol{f}^{-1}\left(\boldsymbol{f}(\boldsymbol{w}_{t-1}) - \eta(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x}_t - y_t)\boldsymbol{x}_t\right)$$

where $\boldsymbol{f}\colon \mathbf{R^n} \to \mathbf{R^n}$ given by

$$f_i(\boldsymbol{w}) = \frac{\text{sign}(w_i)|w_i|^{q-1}}{||\boldsymbol{w}||_q^{q-2}} \qquad \text{and} \qquad f_i^{-1}(\boldsymbol{\theta}) = \frac{\text{sign}(\theta_i)|\theta_i|^{p-1}}{||\boldsymbol{\theta}||_p^{p-2}}$$

When $p = q = 2$, then $\boldsymbol{f}(\boldsymbol{w}) = \boldsymbol{w}$: LMS

For large p, $\boldsymbol{f}^{-1}$ emphasizes differences in components

A priori filtering bound

Theorem Assume $||\boldsymbol{x}_t||_p \leq X_p$ for all t, and let $\eta = 1/((p-1)X_p^2)$
Then for any $\boldsymbol{u}$ the p-norm algorithm satisfies

$$\sum_{t=1}^{T}(\boldsymbol{u}\cdot\boldsymbol{x}_t - \boldsymbol{w}_{t-1}\cdot\boldsymbol{x}_t)^2 \leq \sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 + (p-1)X_p^2||\boldsymbol{u}||_q^2$$

- $1/p + 1/q = 1$ and $2 \leq p < \infty$, $1 < q \leq 2$
- How do we get the dual norm pair $(\infty, 1)$ (where $||\boldsymbol{x}||_\infty = \max_i |x_i|$)?

For $p = 2\ln n$, $(p-1)||\boldsymbol{x}||_p^2||\boldsymbol{u}||_q^2 \leq (2e\ln n)||\boldsymbol{x}||_\infty^2||\boldsymbol{u}||_1^2$

Comparison with basic LMS

New bounds incomparable with old ones because for $p > 2$ and $q < 2$

$$||\boldsymbol{x}||_p < ||\boldsymbol{x}||_2 \qquad \text{and} \qquad ||\boldsymbol{u}||_q > ||\boldsymbol{u}||_2$$

Compare $p = 2$ and $p = O(\log n)$ in two extreme cases:

Sparse target, dense instances: Let $\boldsymbol{u} = (1, 0, \ldots, 0)$ and $\boldsymbol{x} = (1, \ldots, 1)$.

- $||\boldsymbol{x}||_2^2 ||\boldsymbol{u}||_2^2 = n^2$
- $(\log n)||\boldsymbol{x}||_\infty^2 ||\boldsymbol{u}||_1^2 = \log n$
- Thus large p better

Dense target, sparse instances: Let $\boldsymbol{u} = (1, 1, \ldots, 1)$ and $\boldsymbol{x} = (1, 0, \ldots, 0)$.

- $||\boldsymbol{x}||_2^2 ||\boldsymbol{u}||_2^2 = n^2$
- $(\log n)||\boldsymbol{x}||_\infty^2 ||\boldsymbol{u}||_1^2 = n^2 \log n$
- Thus $p = 2$ better

The p-norm LMS can behave like EG

Hadamard Matrix:

$$
\begin{array}{rcccc}
 & \rightarrow & +1 & +1 & +1 & +1 \\
instances & \rightarrow & +1 & -1 & +1 & -1 \\
 & \rightarrow & +1 & +1 & -1 & -1 \\
 & \rightarrow & +1 & -1 & -1 & +1 \\
 & & \uparrow & \uparrow & \uparrow & \uparrow \\
 & & & targets & &
\end{array}
$$

- Instances are orthogonal
- Target weight vectors are units
- LMS: error $\geq 1 - \frac{k}{n}$
- p-norm LMS with $p = O(\log n)$: error $\geq \frac{\ln n}{k}$

Time-varying target (following [HW01])

Up to now, model has been $y_t = \boldsymbol{u} \cdot \boldsymbol{x}_t + \text{noise}$ where target $\boldsymbol{u}$ is fixed

Generalize this to $y_t = \boldsymbol{u}_t \cdot \boldsymbol{x}_t + \text{noise}$ where target $\boldsymbol{u}_t$ may vary over time

Example 1: target makes one jump Choose $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{R}^n$ and take

$$\boldsymbol{u}_t = \begin{cases} \boldsymbol{a} & \text{for } 1 \le t \le T/2 \\ \boldsymbol{b} & \text{for } T/2 < t \le T \end{cases}$$

Example 2: target moves steadily Choose $\boldsymbol{a}, \boldsymbol{b} \in \mathbf{R}^n$ and take

$$\boldsymbol{u}_t = \frac{T-t}{T-1}\boldsymbol{a} + \frac{t-1}{T-1}\boldsymbol{b}$$

Algorithms for time-varying target

Old update:

$$\boldsymbol{w}'_t = \boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{w}_{t-1}) - \eta(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t - y_t)\boldsymbol{x}_t)$$

Bounding update:

$$\boldsymbol{w}_t = \begin{cases} \boldsymbol{w}'_t & \text{if } ||\boldsymbol{w}'_t||_q \leq U_q \\ U_q \frac{\boldsymbol{w}'_t}{||\boldsymbol{w}'_t||_q} & \text{otherwise} \end{cases}$$

where $U_q > 0$ is a norm bound

We rescale the weight vector whenever q-norm larger than U_q

Bound for time-varying target

Theorem Assume $||\boldsymbol{x}_t||_p \leq X_p$ for all t, and let $\eta = 1/((p-1)X_p^2)$
Then if $||\boldsymbol{u}_t||_q \leq U_q$ for all t, the bounded p-norm LMS satisfies

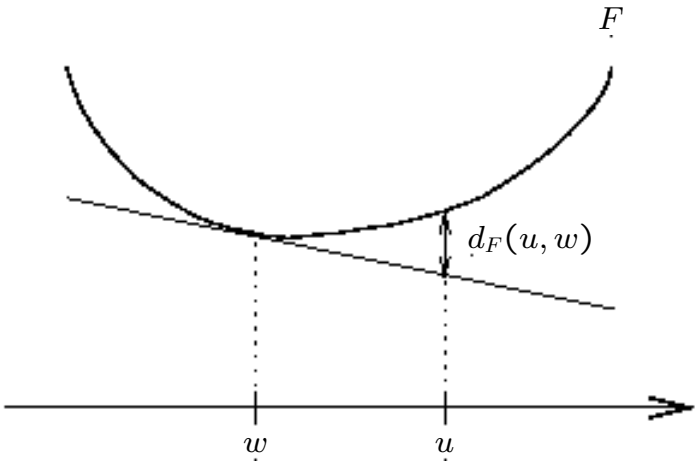
$$\sum_{t=1}^{T}(\boldsymbol{u}_t \cdot \boldsymbol{x}_t - \boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t)^2 \leq \sum_{t=1}^{T}(y_t - \boldsymbol{u}_t \cdot \boldsymbol{x}_t)^2 + (p-1)X_p^2 U_q^2 + 2(p-1)X_p^2 U_q \sum_{t=1}^{T-1} ||\boldsymbol{u}_{t+1} - \boldsymbol{u}_t||_q$$

- Only total distance $\sum_t ||\boldsymbol{u}_{t+1} - \boldsymbol{u}_t||_q$ traveled by the target matters
- Cost $2(p-1)X_p^2 U_q$ per unit target movement
- For fixed target $\boldsymbol{u}_{t+1} = \boldsymbol{u}_t$, we recover previous bound
- However U_q needs to be known in advance

Bregman divergences

- Key tool in analyzing and understanding the algorithms
- Fix strictly convex differentiable $F: \mathbf{R^n} \to \mathbf{R}$.
  Denote the gradient by $\boldsymbol{f} = \nabla F$.
- Now the Bregman divergence $d_F: \mathbf{R^n} \times \mathbf{R^n} \to \mathbf{R}$ is

$$d_F(\boldsymbol{u}, \boldsymbol{w}) = F(\boldsymbol{u}) - F(\boldsymbol{w}) - \boldsymbol{f}(\boldsymbol{w}) \cdot (\boldsymbol{u} - \boldsymbol{w})$$



$d_F(\boldsymbol{u}, \boldsymbol{w})$ is the error of first-order Taylor approximation of $F(\boldsymbol{u})$ around $\boldsymbol{w}$

Basic properties of Bregman divergences

- $d_F(\boldsymbol{u}, \boldsymbol{w}) \geq 0$, $d_F(\boldsymbol{u}, \boldsymbol{w}) = 0$ iff $\boldsymbol{u} = \boldsymbol{w}$
- not symmetrical (in general)
- does not satisfy triangle inequality
- $d_F(\boldsymbol{u}, \boldsymbol{w})$ convex in $\boldsymbol{u}$, not necessarily in $\boldsymbol{w}$

Connection to exponential families (roughly):

- F is cumulant function, $\boldsymbol{f}$ is link function
- $\boldsymbol{w}$ is expectation parameter, $\boldsymbol{f}(\boldsymbol{w})$ canonical parameter
- $d_F(\boldsymbol{u}, \boldsymbol{w})$ is the KL divergence between distributions parameterized by $\boldsymbol{u}$ and $\boldsymbol{w}$

Example: p-norm divergence [GLS01]

- $$F(\boldsymbol{w}) = \frac{1}{2}||\boldsymbol{w}||_q^2$$

- Then the gradient $\boldsymbol{f} = \nabla F$ satisfies

$$f_i(\boldsymbol{w}) = \frac{\text{sign}(w_i)|w_i|^{q-1}}{||\boldsymbol{w}||_q^{q-2}} \qquad \text{and} \qquad f_i^{-1}(\boldsymbol{\theta}) = \frac{\text{sign}(\theta_i)|\theta_i|^{p-1}}{||\boldsymbol{\theta}||_p^{p-2}}$$

- The divergence is

$$d_F(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}||\boldsymbol{u}||_q^2 - \frac{1}{2}||\boldsymbol{w}||_q^2 - \boldsymbol{f}(\boldsymbol{w}) \cdot (\boldsymbol{u} - \boldsymbol{w}).$$

- Special case $p = q = 2$ gives $d_F(\boldsymbol{u}, \boldsymbol{w}) = \frac{1}{2}||\boldsymbol{u} - \boldsymbol{v}||_2^2$

Deriving the updates

- Define a regularized instantaneous loss

$$C_t(\boldsymbol{w}) = d_F(\boldsymbol{w}, \boldsymbol{w_{t-1}}) + \frac{\eta}{2}(y_t - \boldsymbol{w} \cdot \boldsymbol{x}_t)^2$$

- Basic aim is to have

$$\boldsymbol{w_t} = \operatorname*{argmin}_{\boldsymbol{w}} C_t(\boldsymbol{w})$$

- Minimize by setting $\nabla C_t(\boldsymbol{w}_t) = \boldsymbol{0}$, obtaining the implicit update

$$\boldsymbol{f}(\boldsymbol{w}_t) = \boldsymbol{f}(\boldsymbol{w}_{t-1}) - \eta(\boldsymbol{w}_t \cdot \boldsymbol{x}_t - y_t)\boldsymbol{x}_t$$

- Approximate $\boldsymbol{w}_t \cdot \boldsymbol{x}_t \approx \boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t$ to obtain the update

$$\boldsymbol{f}(\boldsymbol{w}_t) = \boldsymbol{f}(\boldsymbol{w}_{t-1}) - \eta(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t - y_t)\boldsymbol{x}_t$$

Analyzing the update

- Measure of progress

$$d_F(\boldsymbol{u}, \boldsymbol{w}_{t-1}) - d_F(\boldsymbol{u}, \boldsymbol{w}_t) = \eta(y_t - \boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t)\boldsymbol{x}_t \cdot (\boldsymbol{u} - \boldsymbol{w}_{t-1}) - d_F(\boldsymbol{w}_{t-1}, \boldsymbol{w}_t)$$

- Massage the term $(y_t - \boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t)\boldsymbol{x}_t \cdot (\boldsymbol{u} - \boldsymbol{w}_{t-1})$ until $(\boldsymbol{u} \cdot \boldsymbol{x}_t - \boldsymbol{w}_{t-1} \cdot \boldsymbol{x}_t)^2$ and $(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ appear; throw rest away

- Estimate $d_F(\boldsymbol{w}_{t-1}, \boldsymbol{w}_t)$ in terms of $||\boldsymbol{x}_t||_p$

- Sum over $t = 1, \ldots, T$

Conclusion

- LMS and normalized LMS can be derived from an optimization problem involving a certain Bregman divergence
- Different Bregman divergences lead to different algorithms, with loss bounds in terms of different norms
- Bounds can be generalized for time-varying targets (and generalized linear models, not presented in the talk); proofs easy
- Algorithms for $p = 2$ can be kernelized, for $p > 2$ probably not

Bottom line: Machinery from on-line machine learning carries over to H^∞-optimal filtering

Where are we headed?

- Develop p-norm Kalman filter
- Prove relative loss bounds