ECE 830 Spring 2013 Statistical Signal Processing

instructors: K. Jamieson and R. Nowak

Lecture: Adaptive Filtering

Adaptive filters are commonly used for online filtering of signals. The goal is to estimate a signal y from a signal x. An adaptive filter is an adjustable filter that processes in time x. The output of the filter is the estimator $\widehat{y}$ of y. The filter is adjusted after each time step to improve the estimation, as depicted in the block diagram below.

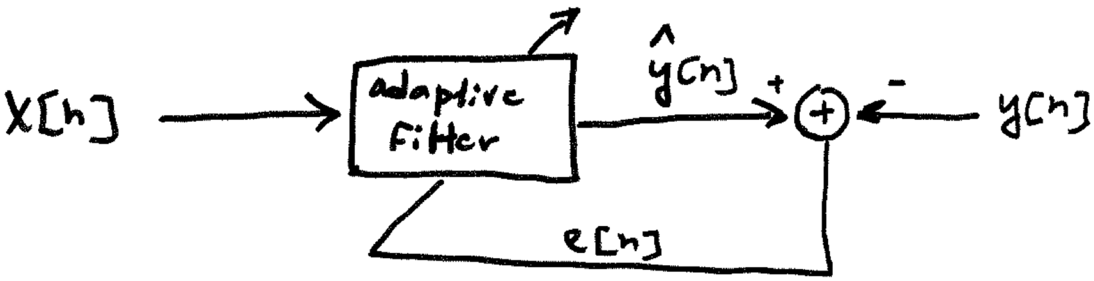


Figure 1: Adaptive filtering block diagram

Wireless channel equalization is a situation where adaptive filters are commonly used. Figure 2 shows how an FIR filter can represent multipath effects in a wireless channel. Consider the traditional adaptive filtering setup where at each time $t = 1, 2, 3, \ldots$ we have a sampled input signal x_n that is sent through an unknown channel and we observe a sampled output signal y_t. The goal is to learn the parameters of the channel so that when we don't know x_t but only observe y_t, we can estimate x_t by removing the effect of the channel. A prototypical example is that of improving the reliability of cell-service or wi-fi. The channel from the cell tower or hotspot to your cell-phone is described by multipath, additive noise, and a frequency selective channel. In order to guarantee high-quality communication performance between the cell tower and your cell phone, the cell tower constantly sends test signals x_t that are known to the phone. The phone then tries to learn the parameters of the channel by comparing the observed signal y_t to the known signal x_t that was transmitted. For this to work on a cell phone, the task of learning the parameters of the channel must be very inexpensive computationally and not require too much wall-clock time. In this note we analyze the the least mean squares (LMS) algorithm from the perspective of online convex optimization via gradient descent. Fetal heart monitoring is another good example, depicted in Fig. 3.

1 Steepest Descent

A bit more formally, suppose that we would like to design an FIR filter to estimate a signal y_t from another signal x_t. The estimator has the form

$$\widehat{y}_t \;=\; \sum_{\tau=0}^{N-1} w_\tau \; x_{t-\tau} \;=\; \mathbf{w}^T \mathbf{x}_t \tag{1}$$

where $\mathbf{x}_t = (x_t, x_{t-1}, \ldots, x_{t-N+1})^T$ and $\mathbf{w} = (w_0, \ldots, w_{N-1})$ is a vector of the FIR filter weights. Throughout, scalars will be roman, vectors will be bold-face, and the dimension of $\mathbf{w}$ will be equal to N. For some given time horizon T define $\mathbf{w}^* \in \mathbb{R}^N$ such that

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{T}\sum_{t=1}^{T}(y_t - \mathbf{w}^T\mathbf{x}_t)^2. \tag{2}$$

If we stack the y_t into a vector $\mathbf{y}$ and the $\mathbf{x}_t$s into a matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)^T$, we also have

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^N} ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\ . \tag{3}$$

Since the squared error is quadratic (and hence convex) in $\mathbf{w}$, we have a simple linear-algebraic solution. It would seem as if we are done. Unfortunately, computing inverses of matrices can be computationally very hard and is impractical for real-time environments on a cell-phone. So an alternative to the matrix-inverse approach is to minimize the squared error using steepest descent. This requires computing the gradient of the squared error, which is $-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w})$. Note that the gradient is zero at the optimal solution, so the optimal $\mathbf{w}^*$ is the solution to the equations $\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}$. Computing the gradient requires all the data, and so gradient descent isn't suitable for an online adaptive filtering algorithm that adjusts the filter as each new sample is obtained. Also, in many situations you need an estimate for $\mathbf{w}^*$ on a timescale much shorter than it takes to perform the inverse and you are willing to accept a poor estimate at first that improves over time and eventually converges to $\mathbf{w}^*$. Finally, there are often situations where the $\mathbf{w}^*$ is not a constant but changes over time and you would like an estimate for $\mathbf{w}^*$ that changes over time. We will study methods that iteratively solve for $\mathbf{w}^*$ and show that these iterates converge to $\mathbf{w}^*$ as T gets big.

To gain a little insight, let us consider the steepest descent algorithm. Steepest descent is an iterative algorithm following these steps:

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} - \frac{1}{2}\mu\frac{\partial\|\mathbf{y} - \mathbf{X}\|_2^2}{\partial\mathbf{w}}\big|_{\mathbf{w}=\mathbf{w}_{t-1}} \\ &= \mathbf{w}_{t-1} + \mu\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}_{t-1})\end{aligned}$$

where $\mu > 0$ is a step-size. Note that the algorithm takes a step in the negative gradient direction (i.e., 'downhill'). The choice of the step size is crucial. If the steps are too large, then the algorithm may diverge. If they are too small, then convergence may take a long time. We can understand the effect of the step size as follows. Note that we can write the iterates as

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} + \mu(\mathbf{X}^T y - \mathbf{X}^T\mathbf{X}\mathbf{w}_{t-1}) \\ &= \mathbf{w}_{t-1} + \mu\mathbf{X}^T\mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y - \mathbf{w}_{t-1}) \\ &= \mathbf{w}_{t-1} - \mu\mathbf{X}^T\mathbf{X}(\mathbf{w}_{t-1} - \mathbf{w}^*)\end{aligned}$$

Subtracting $\mathbf{w}^*$ from both sides gives us

$$\mathbf{v}_t = \mathbf{v}_{t-1} - \mu\mathbf{X}^T\mathbf{X}\mathbf{v}_{t-1}$$

where $\mathbf{v}_t = \mathbf{w}_t - \mathbf{w}^*$, $t = 1, 2, \ldots$ So we have

$$\begin{aligned}\mathbf{v}_t &= (\mathbf{I} - \mu\mathbf{X}^T\mathbf{X})\mathbf{v}_{t-1} \\ &= (\mathbf{I} - \mu\mathbf{X}^T\mathbf{X})^{t-1}\mathbf{v}_1\end{aligned}$$

Thus the sequence $\mathbf{v}_t \to 0$ if all the eigenvalues of $(\mathbf{I} - \mu\mathbf{X}^T\mathbf{X})$ are less than 1. This holds if $\mu < \lambda_{\mathtt{max}}^{-1}(\mathbf{X}^T\mathbf{X})$. We will see that the eigenvalues of $\mathbf{X}^T\mathbf{X}$ play a key role in adaptive filtering algorithms.

2 The LMS Algorithm

The Least Mean Square (LMS) algorithm is an online variant of steepest descent. One can think of the LMS algorithm as considering each term in the sum of (2) individually in order. The LMS iterates are

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} - \frac{1}{2}\mu\frac{(y_t - \mathbf{x}_t^T\mathbf{w})^2}{\partial \mathbf{w}}\Big|_{\mathbf{w}=\mathbf{w}_{t-1}} \\ &= \mathbf{w}_{t-1} - \mu(y_t - \mathbf{w}_{t-1}^T\mathbf{x}_t)\mathbf{x}_t\end{aligned}$$

The full gradients are simply replaced by instantaneous gradients. Geometrically, for $T > N$ the complete sum of (2) tends to look like a convex, quadratic bowl while each individual term is described by a degenerate quadratic in the sense that in all but 1 of the N orthogonal directions, the function is flat. This concept is illustrated in Figure 4 with f_t equal to $(y_t - \mathbf{x}_t^T\mathbf{w})^2$. Intuitively, each individual function f_t only tells us about at most one dimension of the total N so we should expect $T \gg N$.

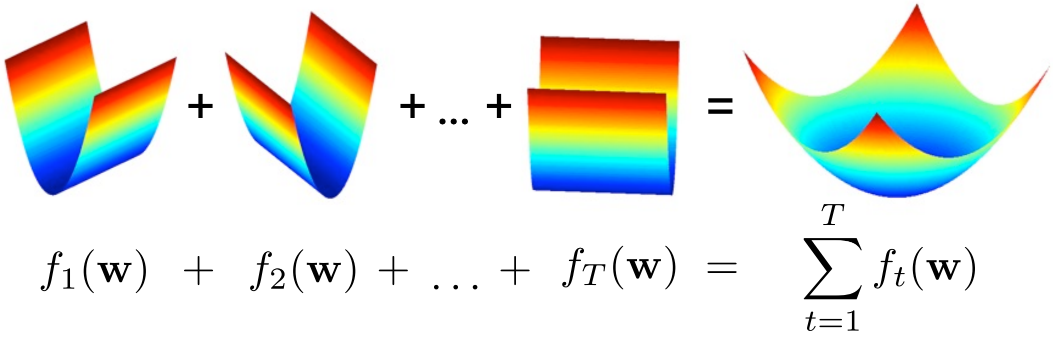

+
+ ... +
=
$f_1(\mathbf{w}) \ + \ f_2(\mathbf{w}) + \ \ldots + \ f_T(\mathbf{w}) \ = \ \sum_{t=1}^{T} f_t(\mathbf{w})$

Figure 4: The LMS algorithm can be thought of as considering each of the T terms of (2) individually. Because each term is "flat" in all but 1 of the total N directions, this implies that each term is convex but not *strongly* convex (see Figure 5). However, if $T > N$ we typically have that the complete sum *is* strongly convex which can be exploited to achieve faster rates of convergence.

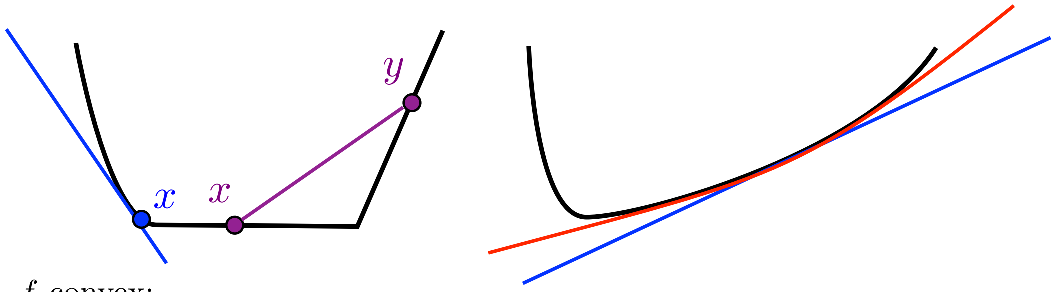
To analyze this algorithm we will study the slightly more general problem

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^N} \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{w}) \tag{4}$$

where each of the $f_t : \mathbb{R}^N \to \mathbb{R}$ are general convex functions (see Figure 5. In the context of LMS, $f_t(\mathbf{w}) := (y_t - \mathbf{x}_t^T\mathbf{w})^2$, which is quadratic and hence convex in $\mathbf{w}$. The problem of (4) is known as an unconstrained online convex optimization program [1]. A very popular approach to solving these problems is gradient descent: for each time t we have an estimate for the best estimate $\mathbf{w}^*$ denoted as $\mathbf{w}_t$ and we set

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t) \tag{5}$$

where η_t is a positive, non-increasing sequence of step sizes and the algorithm is initialized with some arbitrary $\mathbf{w}_1 \in \mathbb{R}^N$. The following theorem characterizes the performance of this algorithm.



f convex:

$$f\left(\lambda x + (1-\lambda) y\right) \le \lambda f(x) + (1-\lambda) f(y) \qquad \forall x, y, \lambda \in [0,1]$$

$$f(y) \ge f(x) + \nabla f(x)^T (y - x) \qquad \forall x, y$$

f ℓ-strongly convex:

$$f(y) \ge f(x) + \nabla f(x)^T (y - x) + \tfrac{\ell}{2} ||y - x||_2^2 \qquad \forall x, y$$

$$\nabla^2 f(x) \succ \ell I \qquad \forall x$$

Figure 5: A function f is said to be convex if for all $\lambda \in [0,1]$ and x, y we have $f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$. If f is differentiable then an equivalent definition is that $f(y) \ge f(x) + \nabla f(x)^T (y - x)$. A function f is said to be ℓ-strongly convex if $f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{\ell}{2} ||y - x||_2^2$ for all x, y. An equivalent definition is that $\nabla^2 f(x) \succ \ell I$ for all x.

Theorem 1. *[1] Let f_t be convex and $||\nabla f_t(\mathbf{w}_t)||_2 \le G$ for all t,$\mathbf{w}_t$ and let $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^N} \sum_{t=1}^T f_t(\mathbf{w})$. Using the algorithm of (5) with $\eta_t = 1/\sqrt{T}$ and arbitrary $\mathbf{w}_1 \in \mathbb{R}^N$ we have*

$$\frac{1}{T} \sum_{t=1}^{T} \left(f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)\right) \le \frac{||\mathbf{w}_1 - \mathbf{w}^*||_2^2 + G^2}{2\sqrt{T}}$$

for all T.

Begin proving the theorem, note that this is a very strong result. It only uses the fact that the f_t functions are convex and that the gradients of f_t are bounded. In particular, it assumes *nothing* about how the f_t functions relate to each other from time to time. Moreover, it requires no unknown parameters to set the step size. In fact, the step size η_t is a *constant* and because the theorem holds for *any* T we can easily turn this into a statement about how well this algorithm tracks a $\mathbf{w}^*$ that changes over time.

Proof. We begin by observing that

$$\begin{aligned} ||\mathbf{w}_{t+1} - \mathbf{w}^*||_2^2 &= ||\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t) - \mathbf{w}^*||_2^2 \\ &= ||\mathbf{w}_t - \eta_t \nabla f_t(\mathbf{w}_t) - \mathbf{w}^*||_2^2 \\ &= ||\mathbf{w}_t - \mathbf{w}^*||_2^2 - 2\eta_t \nabla f_t(\mathbf{w}_t)^T (\mathbf{w}_t - \mathbf{w}^*) + \eta_t^2 ||\nabla f_t(\mathbf{w}_t)||_2^2 \end{aligned}$$

and after rearranging we have that

$$\nabla f_t(\mathbf{w}_t)^T (\mathbf{w}_t - \mathbf{w}^*) \le \frac{||\mathbf{w}_t - \mathbf{w}^*||_2^2 - ||\mathbf{w}_{t+1} - \mathbf{w}^*||_2^2}{2\eta_t} + \frac{\eta_t}{2} G^2. \qquad (6)$$

By the convexity of f_t for all t, $\mathbf{w}_t$ we have $f_t(\mathbf{w}^*) - f_t(\mathbf{w}_t) \ge \nabla f_t(\mathbf{w}_t)^T (\mathbf{w}^* - \mathbf{w}_t)$. Thus, summing both sides of this equation from $t = 1$ to T and plugging in $\eta_t = 1/\sqrt{T}$ we have

$$\sum_{t=1}^{T} f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*) \leq \left(\frac{||\mathbf{w}_1 - \mathbf{w}^*||_2^2}{2\eta_1} + \frac{1}{2}\sum_{t=2}^{T} ||\mathbf{w}_t - \mathbf{w}^*||_2^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \right) + \frac{G^2}{2}\sum_{t=1}^{T} \eta_t$$
$$= \left(||\mathbf{w}_1 - \mathbf{w}^*||_2^2 + G^2 \right) \frac{\sqrt{T}}{2} \ .$$

□

This agnostic approach to the functions f_t allow us to apply the theorem to analyzing LMS for adaptive filters where there is lots of feedback and dependencies between iterations. If we plugin $f_t(\mathbf{w}) = ||y_t - \mathbf{w}^T\mathbf{x}_t||_2$ so that $\nabla f_t(\mathbf{w}) = -2(y_t - \mathbf{w}^T\mathbf{x}_t)\mathbf{x}_t$ we see that $G^2 \leq 4N \max_t x_t^2 \left(\max_t y_t^2 + N \max_t x_t^2 ||\mathbf{w}_t||_2^2 \right) \approx N^2 \max_t x_t^4 ||\mathbf{w}_t||_2^2$. The takeaway here is that if we assume *nothing* about the input signal x_t we can do about as well as $\mathbf{w}^*$ at a rate proportional to $1/\sqrt{T}$. With some additional assumptions on x_t, however, we can achieve a $1/T$ rate.

To gain more intuition for how the algorithm actually performs in practice, suppose x_t was a zero-mean, stationary random process and the model of (1) was correct: $y_t = \mathbf{x}_t^T\mathbf{w}^* + e_t$ for some $\mathbf{w}^* \in \mathbb{R}^N$. The errors e_t are assumed to be uncorrelated with x_t and generated by a zero-mean stationary noise process with variance $\mathbb{E}[e_t^2] \leq \sigma^2$. Then for any $\mathbf{w} \in \mathbb{R}^N$

$$\mathbb{E}[f_t(\mathbf{w})] = \mathbb{E}\left[||y_t - \mathbf{w}^T\mathbf{x}_t||_2^2\right] = \mathbb{E}\left[||(\mathbf{w}^* - \mathbf{w})^T\mathbf{x}_t||_2^2\right] + \sigma^2 = (\mathbf{w} - \mathbf{w}^*)^T\mathbb{E}\left[\mathbf{x}_t\mathbf{x}_t^T\right](\mathbf{w} - \mathbf{w}^*) + \sigma^2 \ .$$

Because x_t is stationary, define R_{xx} be the autocorrelation matrix for x such that

$$(R_{xx})_{i,j} = \mathbb{E}[x_{t-i+1}x_{t-j+1}] = \mathbb{E}[x_0 x_{i-j}]$$

for all t. It follows that $f(\mathbf{w}) := \mathbb{E}[f_t(\mathbf{w})] = (\mathbf{w}-\mathbf{w}^*)^T R_{xx}(\mathbf{w}-\mathbf{w}^*)+\sigma^2$ and $\mathbb{E}[\nabla f_t(\mathbf{w})] = \nabla f(\mathbf{w}) = 2R_{xx}(\mathbf{w}-\mathbf{w}^*)$. The following theorem refines are convergence analysis of the algorithm under these assumptions. In the theorem, $F(\mathbf{w}, \xi)$ will play the role of $f_t(\mathbf{w})$, which is a function of x_t and e_t, which can be identified with ξ in the context of the theorem.

Theorem 2. *Let $F(\mathbf{w}, \xi)$ be a function that takes as input $\mathbf{w} \in \mathbb{R}^N$ and a random vector $\xi \in \Xi$ drawn from some probability distribution. Let $f(\mathbf{w}) = \mathbb{E}_\xi[F(\mathbf{w}, \xi)]$ be ℓ-strongly convex and for any $t = 1, \dots, T$ let $\nabla f_t(\mathbf{w}) := G(\mathbf{w}, \xi)$ be an unbiased estimator of $\nabla f(\mathbf{w})$ with respect to ξ with $\mathbb{E}[||G(\mathbf{w}_t, \xi_t)||_2^2] \leq M^2$ for all t, $\mathbf{w}_t$. If $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^N} f(\mathbf{w})$ and $\mathbf{w}_2, \dots, \mathbf{w}_T$ are a sequence of iterates generated from equation* (5) *with $\eta_t = \frac{1}{\ell t}$ and arbitrary $\mathbf{w}_1 \in \mathbb{R}^N$ then*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)] \leq \frac{||\mathbf{w}_1 - \mathbf{w}^*||_2^2 + M^2\log(eT)}{2\ell T} \ .$$

and

$$\mathbb{E}[||\mathbf{w}_T - \mathbf{w}^*||_2^2] \leq \frac{\max\{||\mathbf{w}_1 - \mathbf{w}^*||_2^2, M^2/\ell^2\}}{T}$$

for all T.

Before proving the theorem, we note that both results are known to be minimax optimal [2,3]. Unfortunately, unlike the previous theorem, to set the step size we need to know ℓ which in general is usually unknown. However, for adaptive filtering we will show that we essentially get to pick ℓ. The following proof is based on the analyses of [2,4].

Proof. We begin by taking the expectation of (6) on both sides:

$$\mathbb{E}[G(\mathbf{w}_t,\xi_t)^T(\mathbf{w}_t-\mathbf{w}^*)] \le \frac{\mathbb{E}[||\mathbf{w}_t-\mathbf{w}^*||_2^2]-\mathbb{E}[||\mathbf{w}_{t+1}-\mathbf{w}^*||_2^2]}{2\eta_t}+\frac{\eta_t}{2}M^2 \ . \tag{7}$$

Define $\xi_{[k]}=\{\xi_1,\dots,\xi_k\}$ so that under the assumption of the theorem, we have

$$\mathbb{E}_{\xi_{[t]}}[(\mathbf{w}_t-\mathbf{w}^*)^T G(\mathbf{w}_t,\xi_t)] = \mathbb{E}_{\xi_{[t-1]}}\left[\mathbb{E}_{\xi_t}\left[(\mathbf{w}_t-\mathbf{w}^*)^T G(\mathbf{w}_t,\xi_t)|\xi_{[t-1]}\right]\right] = \mathbb{E}_{\xi_{[t-1]}}\left[(\mathbf{w}_t-\mathbf{w}^*)^T\nabla f(\mathbf{w}_t)\right]$$

for any $t=1,\dots,T$. By the strong convexity of f we have

$$(\mathbf{w}_t-\mathbf{w}^*)^T\nabla f(\mathbf{w}_t) \ge f(\mathbf{w}_t)-f(\mathbf{w}^*)+\frac{\ell}{2}||\mathbf{w}_t-\mathbf{w}^*||_2^2 \tag{8}$$

Note that strong convexity also implies

$$\begin{aligned} f(\mathbf{w}_t)-f(\mathbf{w}^*) &\ge \nabla f(\mathbf{w}^*)^T(\mathbf{w}_t-\mathbf{w}^*)+\ell/2||\mathbf{w}_t-\mathbf{w}^*||^2 \\ &\ge \ell/2||\mathbf{w}_t-\mathbf{w}^*||^2 \end{aligned}$$

since by definition $\nabla f(\mathbf{w}^*)=0$. So we also see that

$$(\mathbf{w}_t-\mathbf{w}^*)^T\nabla f(\mathbf{w}_t) \ge \ell||\mathbf{w}_t-\mathbf{w}^*||_2^2. \tag{9}$$

Thus, applying (8) and summing both sides of (7) from $t=1$ to T and plugging in $\eta_t=\frac{1}{\ell t}$ we have

$$\begin{aligned}\sum_{t=1}^T \mathbb{E}[f(\mathbf{w}_t)-f_t(\mathbf{w}^*)] &\le \left(\frac{||\mathbf{w}_1-\mathbf{w}^*||_2^2}{2\eta_1}+\frac{1}{2}\sum_{t=2}^T\mathbb{E}[||\mathbf{w}_t-\mathbf{w}^*||_2^2]\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-\ell\right)\right)+\frac{M^2}{2}\sum_{t=1}^T\eta_t \\ &= \left(\frac{||\mathbf{w}_1-\mathbf{w}^*||_2^2}{2\ell}+0\right)+\frac{M^2}{2}\sum_{t=1}^T\frac{1}{\ell t} \le \frac{||\mathbf{w}_1-\mathbf{w}^*||_2^2}{2\ell}+\frac{M^2}{2\ell}\left(1+\log(T)\right) \ .\end{aligned}$$

Returning to (7) and applying (9) after rearranging we also have

$$\mathbb{E}[||\mathbf{w}_{t+1}-\mathbf{w}^*||_2^2] \le (1-2\eta_t\ell)\mathbb{E}[||\mathbf{w}_t-\mathbf{w}^*||_2^2]+\eta_t^2M^2 \ .$$

We'll now show by induction that with $\eta_t=\frac{1}{\ell t}$ we have $\mathbb{E}[||\mathbf{w}_t-\mathbf{w}^*||_2^2]\le\frac{Q}{t}$ for $Q=\max\{||\mathbf{w}_1-\mathbf{w}^*||_2^2,\frac{M^2}{\ell^2}\}$. It is obviously true for $t=1$ so assume it holds for some time t. Plugging these values into the above recursion assuming $\mathbb{E}[||\mathbf{w}_t-\mathbf{w}^*||_2^2]\le\frac{Q}{t}$ we have

$$\left(1-\frac{2}{t}\right)\frac{Q}{t}+\frac{M^2}{\ell^2t^2}\le Q\left(\frac{1}{t}-\frac{1}{t^2}\right)=Q\left(\frac{t+1}{t(t+1)}-\frac{t+1}{t^2(t+1)}\right)\le Q\left(\frac{t+1}{t(t+1)}-\frac{t}{t^2(t+1)}\right)=\frac{Q}{t+1} \ .$$

□

To apply the theorem to adaptive filtering, recall that $f_t(\mathbf{w})$ is represented by $F(\mathbf{w},\xi_t)$ with $\xi_t:=(x_t,e_t)$. We also need to define M and ℓ. For M we have

$$\begin{aligned}\mathbb{E}[||\nabla f_t(\mathbf{w}_t)||_2^2] &= \mathbb{E}[||2(y_t-\mathbf{w}_t^T\mathbf{x}_t)\mathbf{x}_t||_2^2] \\ &= \mathbb{E}[||2(\mathbf{w}^*-\mathbf{w}_t)^T\mathbf{x}_t\mathbf{x}_t+2e_t\mathbf{x}_t||_2^2] \\ &\le 4||\mathbf{w}^*-\mathbf{w}_1||_2^2\ \mathbb{E}[||\mathbf{x}_t||_2^4]+8\sigma^2\ \mathbb{E}[||\mathbf{x}_t||_2^2] =: M^2\end{aligned}$$

The strong convexity parameter ℓ is equal to the minimum eigenvalue of R_{xx}: $\ell=\lambda_{\min}(R_{xx})$. The larger $\lambda_{\min}$ the greater the curvature of bowl that we are trying to minimize in the least squares problem and the faster the convergence rate.

Example: Let $x_t\sim\mathcal{N}(0,1)$, $e_t\sim\mathcal{N}(0,\sigma^2)$, and $y_t=\mathbf{x}_t^T\mathbf{w}^*+e_t$. Then $R_{xx}=I$ and $\ell=1$ and $M^2=4||\mathbf{w}^*-\mathbf{w}_1||_2^2(N+4)^2+8\sigma^2N$. Thus, $\mathbb{E}[||\mathbf{w}_T-\mathbf{w}^*||_2^2]\le\frac{4||\mathbf{w}^*-\mathbf{w}_1||_2^2(N+4)^2+8\sigma^2N}{T}$.

References

[1] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. 2003.

[2] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[3] Maxim Raginsky and Alexander Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 803–510. IEEE, 2009.

[4] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

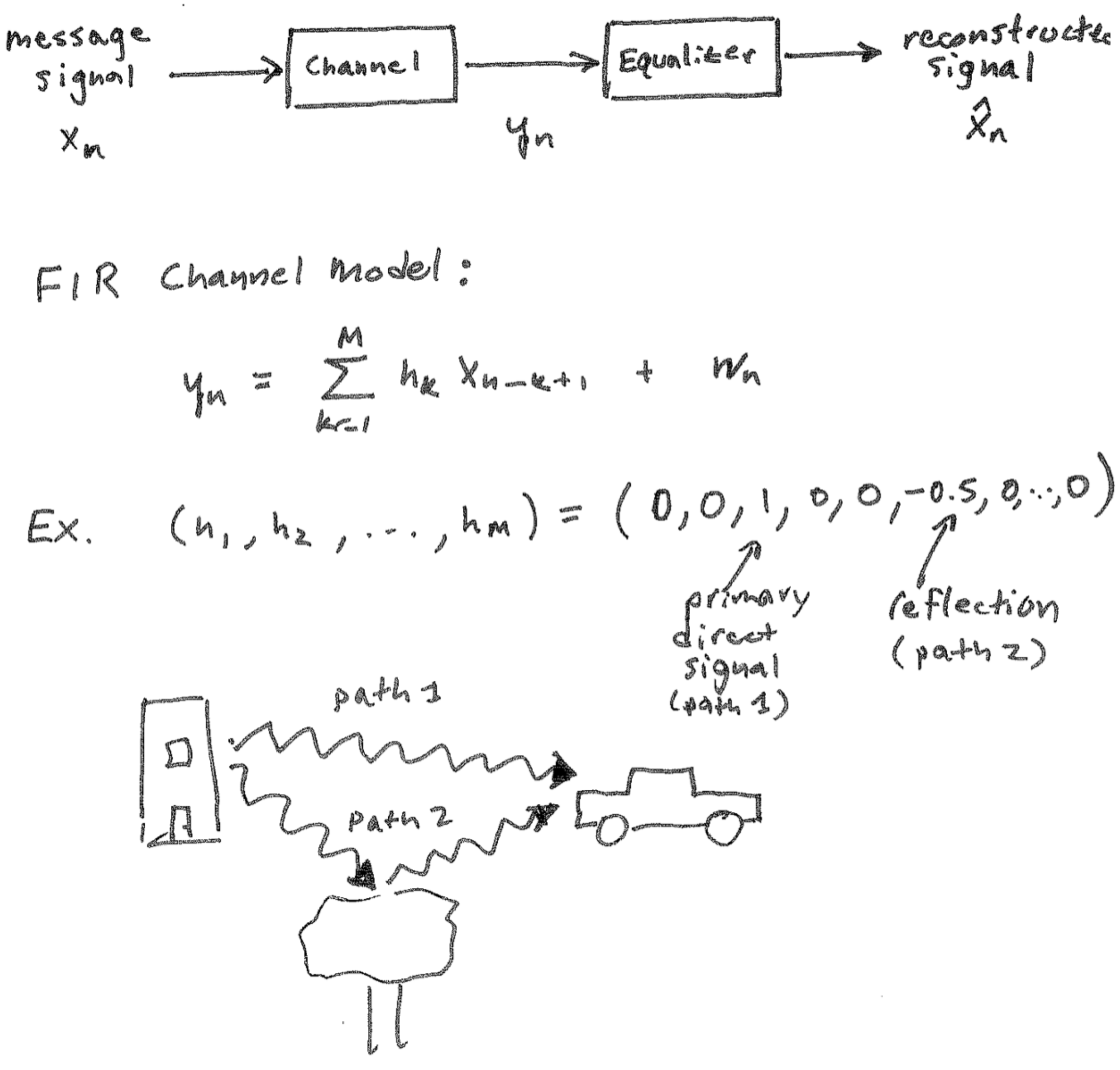

message signal
x_n
Channel
y_n
Equalizer
reconstructe
signal
$\hat{x}_n$
FIR Channel model:
$y_n = \sum_{k=1}^{M} h_k x_{n-k+1} + w_n$
Ex. $(h_1, h_2, \ldots, h_M) = (0, 0, 1, 0, 0, -0.5, 0, \ldots, 0)$
primary direct signal (path 1)
reflection (path 2)
path 1
path 2

Figure 2: Channel equalization

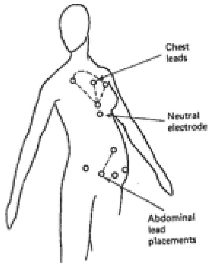
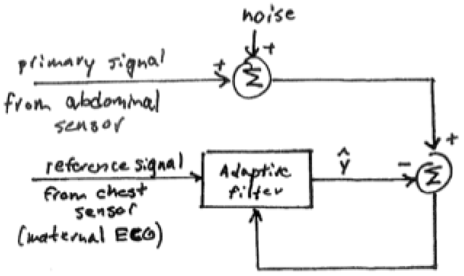
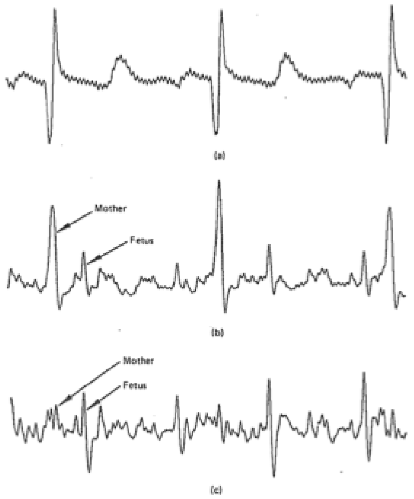


Figure 1.1 Cancelling maternal heartbeat in fetal electrocardiography (ECG): position of leads. From B. Widrow et al, *Adaptive Noise Cancelling: Principles and applications* [W3], © December 1975, IEEE.



$\hat{y}$ is a estimate of the maternal ECG signal present in abdominal signal



(a)

(b)

(c)

Figure 1.3 Results of fetal ECG experiment (bandwidth, 3–35 Hz; sampling rate, 256 Hz): (a) reference input (chest lead); (b) primary input (abdominal lead); (c) noise-canceller output. From B. Widrow et al, *Adaptive Noise Canceling: Principles and applications* [W3], © December 1975, IEEE.

Figure 3: Fetal heart monitoring