Linear/Additive Non-Parametric Estimation
10716, Spring 2020
Pradeep Ravikumar (amending notes from Larry Wasserman)

1 Introduction

Consider the linear regression model:

$$Y = \sum_{j=1}^{d} \beta_j X_j + \epsilon,$$

where $\mathbb{E}(\epsilon) = 0$, so that the regression function $m(X) := \mathbb{E}(Y|X)$ is a linear function of X. More generally, suppose $m \in \mathcal{H}$, so that the regression function lies in a Hilbert space of functions. Let $\mathcal{D}$ be a **dictionary** which is any set of functions from $\mathcal{H}$ whose linear span is $\mathcal{H}$. The elements of $\mathcal{D}$ are called **atoms**, or simply, dictionary elements. We do not necessarily assume that the dictionary elements are orthogonal. Though we will typically assume that the atoms are normalized, that is, $\|\psi\| = 1$ for all $\psi \in \mathcal{D}$. We will also assume that $\mathcal{D}$ is finite or countable so we can write $\mathcal{D} = \{\psi_j\}_{j \in \mathbb{N}}$.

In such a case, we can then write $m(X) = \sum_j \beta_j \Psi_j(X_j)$, so that it is linear given the dictionary $\mathcal{D}$.

This could thus be viewed as infinite-dimensional analogue of linear regression. In the sequel, in places, we might overload notation, and assume that X itself is infinite-dimensional, without loss of generality.

We then wish to estimate this regression function given observations $D := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where $X_i = (X_i(1), \ldots, X_i(d)) \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$.

We would be interested in the *conditional prediction risk*

$$r(\widehat{m}) = \mathbb{E}[(Y - \widehat{m}(X))^2 | D] = \int (y - \widehat{m}(x))^2 dP(x, y)$$

and the *prediction risk* of $\widehat{m}$ is

$$R(\widehat{m}) = \mathbb{E}(Y - \widehat{m}(X))^2 = \mathbb{E}[r(\widehat{m})]$$

where the expected value is over all random variables.

2 Review: Low Dimensional Linear Regression

Let us first review some classical results for finite low-dimensional linear regression.

Recall that the ordinary least squares (OLS) estimate of the best linear model

$$\widehat{\beta} \in \arg\min_{\beta} \frac{1}{n} \sum_{i} (Y_i - X_i^T \beta)^2$$

is given by

$$\widehat{\beta} = \widehat{\Sigma}^{-1} \widehat{\alpha}$$

where $\widehat{\Sigma} = n^{-1} \sum_{i=1}^{n} X_i X_i^T$ and $\widehat{\alpha} = n^{-1} \sum_{i=1}^{n} Y_i X_i$.

Theorem 1 (Theorem 11.3 of Gyorfi, Kohler, Krzyzak and Walk, 2002) *Let* $\sigma^2 = \sup_x \mathrm{Var}(Y|X = x) < \infty$. *Assume that all the random variables are bounded by* $L < \infty$. *Then*

$$\mathbb{E} \int |\widehat{\beta}^T x - m(x)|^2 dP(x) \leq 8 \inf_{\beta} \int |\beta^T x - m(x)|^2 dP(x) + \frac{Cd(\log(n) + 1)}{n}.$$

The proof is straightforward but is very long. The strategy is to first bound $n^{-1} \sum_i (\widehat{\beta}^T X_i - m(X_i))^2$ using the properties of least squares. Then, using concentration of measure one can relate $n^{-1} \sum_i f^2(X_i)$ to $\int f^2(x) dP(x)$.

We can further tighten the result above via the following concentration result:

Theorem 2 (Hsu, Kakade and Zhang 2014) *Let* $m(x) = \mathbb{E}[Y|X = x]$ *and* $\epsilon = Y - m(X)$. *Suppose there exists* $\sigma \geq 0$ *such that*

$$\mathbb{E}[e^{t\epsilon}|X = x] \leq e^{t^2\sigma^2/2}$$

for all x *and all* $t \in \mathbb{R}$. *Let* $\beta^T x$ *be the best linear apprximation to* $m(x)$. *With probability at least* $1 - 3e^{-t}$,

$$r(\widehat{\beta}) - r(\beta) \leq \frac{2A}{n}(1 + \sqrt{8t})^2 + \frac{\sigma^2(d + 2\sqrt{dt} + 2t)}{n} + o(1/n)$$

where $A = \mathbb{E}[||\Sigma^{-1/2} X (m(X) - \beta^T X)||^2]$.

Thus, the excess prediction risk scales as $\frac{d \log n}{n}$, which clearly does not allow for high-dimensional d, let alone infinite-dimensions.

2.1 Ridge Regression

An alternative classical approach is to minimize

$$\frac{1}{n}\sum_i (Y_i - X_i^T\beta)^2 + \lambda||\beta||^2$$

where $\lambda \geq 0$. The minimizer is

$$\widehat{\beta} = (\widehat{\Sigma} + \lambda I)^{-1}\widehat{\alpha}.$$

Theorem 3 (Hsu, Kakade and Zhang 2014) *Suppose that* $||X_i|| \leq r$. *Let* $\beta^T x$ *be the best linear apprximation to* $m(x)$. *Then, with probability at least* $1 - 4e^{-t}$,

$$r(\widehat{\beta}) - r(\beta) \leq \mathbb{E}[(Y - \widehat{\beta}^T X)^2] - \mathbb{E}[(Y - \beta^T X)^2] \leq \left(1 + O\left(\frac{1 + \frac{r^2}{\lambda}}{n}\right)\right)\frac{\lambda||\beta||^2}{2} + \frac{\sigma^2}{n}\frac{\mathrm{tr}(\Sigma)}{2\lambda}.$$

Thus, provided the effective dimensionality for ridge regression depends on the spectrum of the population covariance Σ, namely on $\mathrm{tr}(\Sigma)$. The analysis could be tightened, but is nonetheless not suited for infinite-dimensional contexts in the traditional ridge regression form.

3 Orthogonal Greedy Algorithm

Let Σ_N denote all linear combinations of elements of $\mathcal{D}$ with at most N terms. Define the best N-term approximation error

$$\sigma_N(f) = \inf_{|\Lambda| \leq N} \inf_{g \in \mathrm{Span}(\Lambda)} \|f - g\| \tag{1}$$

where Λ denotes a subset of $\mathcal{D}$ and $\mathrm{Span}(\Lambda)$ is the set of linear combinations of functions in Λ.

We will assume that the regresstion function f is in the span of the dictionary. The function may then have more than one expansion of the form $f = \sum_j \beta_j \psi_j$. We define the norm

$$\|f\|_{\mathcal{L}_p} = \inf \|\beta\|_p$$

where the infimum is over all expansions of f.

We will first consider the population or noiseless variant of the regression problem, where we are given the true function f and we aim to approximate it via a linear combination of dictionary elements. In this population or functional setting, we present an algorithm

```
1. Input: $f$.
2. Initialize: $r_0 = f$, $f_0 = 0$, $V = \emptyset$.
3. Repeat: At step $N$ define
       $g_N = \text{argmax}_{\psi \in \mathcal{D}} |\langle r_{N-1}, \psi \rangle|$
   and set $V_N = V_{N-1} \cup \{g_N\}$. Let $f_N$ be the projection of $r_{N-1}$ onto $\text{Span}(V_N)$.
   Let $r_N = f - f_N$.
```

Figure 1: The Orthogonal Greedy Algorithm.

called **Orthogonal Greedy Algorithm** (OGA), also known as **Orthogonal Matching Pursuit**. The algorithm is given in Figure 1.

The algorithm produces a series of approximations f_N with corresponding residuals r_N. We have the following two theorems from Barron et al (2008), the first dating back to DeVore and Temlyakov (1996).

Theorem 4 *For all $f \in \mathcal{L}_1$, the residual r_N after N steps of OGA satsifies*

$$\|r_N\| \leq \frac{\|f\|_{\mathcal{L}_1}}{\sqrt{N+1}} \tag{2}$$

for all $N \geq 1$.

Proof. Note that f_N is the best approximation to f from $\text{Span}(V_N)$. On the other hand, the best approximation from the set $\{a\, g_N : \ a \in \mathbb{R}\}$ is $\langle f, g_N \rangle g_N$. The error of the former must be smaller than the error of the latter. In other words, $||f - f_N||^2 \leq ||f - f_{N-1} - \langle r_{N-1}, g_N \rangle g_N||^2$. Thus,

$$\begin{aligned} \|r_N\|^2 &\leq \|r_{N-1} - \langle r_{N-1}, g_N \rangle g_N\|^2 \\ &= \|r_{N-1}\|^2 + |\langle r_{N-1}, g_N \rangle|^2 \underbrace{\|g_N\|^2}_{=1} - 2|\langle r_{N-1}, g_N \rangle|^2 \\ &= \|r_{N-1}\|^2 - |\langle r_{N-1}, g_N \rangle|^2. \end{aligned} \tag{3}$$

Now, $f = f_{N-1} + r_{N-1}$ and $\langle f_{N-1}, r_{N-1} \rangle = 0$. So,

$$\begin{aligned} \|r_{N-1}\|^2 &= \langle r_{N-1}, r_{N-1} \rangle = \langle r_{N-1}, f - f_{N-1} \rangle = \langle r_{N-1}, f \rangle - \underbrace{\langle r_{N-1}, f_{N-1} \rangle}_{=0} \\ &= \langle r_{N-1}, f \rangle = \sum_j \beta_j \langle r_{N-1}, \psi_j \rangle \leq \sup_{\psi \in \mathcal{D}} |\langle r_{N-1}, \psi \rangle| \sum_j |\beta_j| \\ &= \sup_{\psi \in \mathcal{D}} |\langle r_{N-1}, \psi \rangle| \, \|f\|_{\mathcal{L}_1} = |\langle r_{N-1}, g_N \rangle| \, \|f\|_{\mathcal{L}_1}. \end{aligned}$$

Continuing from equation (3), we have

$$\begin{aligned}\|r_N\|^2 &\leq \|r_{N-1}\|^2 - |\langle r_{N-1}, g_N\rangle|^2 = \|r_{N-1}\|^2 \left(1 - \frac{\|r_{N-1}\|^2 |\langle r_{N-1}, g_N\rangle|^2}{\|r_{N-1}\|^4}\right) \\ &\leq \|r_{N-1}\|^2 \left(1 - \frac{\|r_{N-1}\|^2 |\langle r_{N-1}, g_N\rangle|^2}{|\langle r_{N-1}, g_N\rangle|^2 \, \|f\|^2_{\mathcal{L}_1}}\right) = \|r_{N-1}\|^2 \left(1 - \frac{\|r_{N-1}\|^2}{\|f\|^2_{\mathcal{L}_1}}\right).\end{aligned}$$

If $a_0 \geq a_1 \geq a_2 \geq \cdots$ are nonnegative numbers such that $a_0 \leq M$ and $a_N \leq a_{N-1}(1 - a_{N-1}/M)$ then it follows from induction that $a_N \leq M/(N+1)$. The result follows by setting $a_N = \|r_N\|^2$ and $M = \|f\|^2_{\mathcal{L}_1}$. □

If f is not in $\mathcal{L}_1$, it is still possible to bound the error as follows.

Theorem 5 *For all $f \in \mathcal{H}$ and $h \in \mathcal{L}_1$,*

$$\|r_N\|^2 \leq \|f - h\|^2 + \frac{4\|h\|^2_{\mathcal{L}_1}}{N}. \tag{4}$$

Proof. Choose any $h \in \mathcal{L}_1$ and write $h = \sum_j \beta_j \psi_j$ where $\|h\|_{\mathcal{L}_1} = \sum_j |\beta_j|$. Write $f = f_{N-1} + f - f_{N-1} = f_{N-1} + r_{N-1}$ and note that r_{N-1} is orthogonal to f_{N-1}. Hence, $\|r_{N-1}\|^2 = \langle r_{N-1}, f\rangle$ and so

$$\begin{aligned}\|r_{N-1}\|^2 &= \langle r_{N-1}, f\rangle = \langle r_{N-1}, h + f - h\rangle = \langle r_{N-1}, h\rangle + \langle r_{N-1}, f - h\rangle \\ &\leq \langle r_{N-1}, h\rangle + \|r_{N-1}\| \, \|f - h\| \\ &= \sum_j \beta_j \langle r_{N-1}, \psi_j\rangle + \|r_{N-1}\| \, \|f - h\| \\ &\leq \sum_j |\beta_j| \, |\langle r_{N-1}, \psi_j\rangle| + \|r_{N-1}\| \, \|f - h\| \\ &\leq \max_j |\langle r_{N-1}, \psi_j\rangle| \sum_j |\beta_j| + \|r_{N-1}\| \, \|f - h\| \\ &= |\langle r_{N-1}, g_k\rangle| \, \|h\|_{\mathcal{L}_1} + \|r_{N-1}\| \, \|f - h\| \\ &\leq |\langle r_{N-1}, g_k\rangle| \, \|h\|_{\mathcal{L}_1} + \frac{1}{2}(\|r_{N-1}\|^2 + \|f - h\|^2).\end{aligned}$$

Hence,

$$|\langle r_{N-1}, g_k\rangle|^2 \geq \frac{(\|r_{N-1}\|^2 - \|f - h\|^2)^2}{4\|h\|^2_{\mathcal{L}_1}}.$$

Thus,

$$a_N \leq a_{N-1}\left(1 - \frac{a_{N-1}}{4\|h\|^2_{\mathcal{L}_1}}\right)$$

where $a_N = \|r_N\|^2 - \|f - h\|^2$. By induction, the last displayed inequality implies that $a_N \leq 4\|h\|^2_{\mathcal{L}_1}/k$ and the result follows. □

```
1. Input: $Y \in \mathbb{R}^n$.
2. Initialize: $r_0 = Y$, $\widehat{f}_0 = 0$, $V = \emptyset$.
3. Repeat: At step $N$ define
       $$g_N = \text{argmax}_{\psi \in \mathcal{D}} |\langle r_{N-1}, \psi \rangle_n|$$
   where $\langle a, b \rangle_n = n^{-1} \sum_{i=1}^n a_i b_i$. Set $V_N = V_{N-1} \cup \{g_N\}$. Let $f_N$ be the projection
   of $r_{N-1}$ onto Span$(V_N)$. Let $r_N = Y - f_N$.
```

Figure 2: Orthogonal Greedy (Forward Stepwise) Regression (Dictionary Version)

Corollary 6 *For each N,*

$$\|r_N\|^2 \leq \sigma_N^2 + \frac{4\theta_N^2}{N}$$

where θ_N is the $\mathcal{L}_1$ norm of the best N-atom approximation.

In Figure 2, we present the noisy variant of the algorithm, where we are provided noisy samples $Y = (Y_1, \ldots, Y_n)^T$ of the regression function at inputs $X = (X_1, \ldots, X_n)$. The corresponding algorithm is also called forward stepwise regression. We will also use the empirical norm defined by

$$\|h\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^n h^2(X_i)}.$$

We assume that the dictionary is normalized in this empirical norm.

By combining the previous results with concentration of measure arguments we get the following result, due to Barron, Cohen, Dahmen and DeVore (2008).

Theorem 7 *Let $h_n = \text{argmin}_{h \in \mathcal{F}_N} \|f_0 - h\|^2$. Suppose that $\limsup_{n \to \infty} \|h_n\|_{\mathcal{L}_{1,n}} < \infty$. Let $N \sim \sqrt{n}$. Then, for every $\gamma > 0$, there exist $C > 0$ such that*

$$\|f - \widehat{f}_N\|^2 \leq 4\sigma_N^2 + \frac{C \log n}{n^{1/2}}$$

except on a set of probability $n^{-\gamma}$.

The rate $n^{-1/2}$ is in fact optimal. A fascinating facet of the bound above is the rate is independent of the dimension.

4 Weak Greedy Algorithm

We will next present an algorithm, known as the **weak greedy algorithm**, which is as follows. We will again first consider the population or functional setting, where we aim to approximate the given regression function as a linear combination of dictionary elements. Let $R_0(f) = f$, $F_0 = 0$. At step k, find $g_k \in \mathcal{D}$ so that

$$|\langle R_{k-1}(f), g_k\rangle| \geq t_k \sup_{h\in\mathcal{D}} |\langle R_{k-1}(f), h\rangle|$$

for some $0 < t_k \leq 1$. In the weak greedy algorithm we take $F_k = F_{k-1} + \langle f, g_k\rangle g_k$. In the weak orthogonal greedy algorithm we take F_k to be the projection of $R_{k-1}(f)$ onto $\{g_1, \ldots, g_k\}$. Finally set $R_k(f) = f - F_k$.

Theorem 8 (Temlyakov 2000) *Let $f(x) = \sum_j \beta_j g_j(x)$ where $g_j \in \mathcal{D}$ and $\sum_{j=1}^{\infty} |\beta_j| \leq B < \infty$. Then, for the weak orthogonal greedy algorithm*

$$\|R_k(f)\| \leq \frac{B}{\left(1 + \sum_{j=1}^{k} t_j^2\right)^{1/2}} \tag{5}$$

and for the weak greedy algorithm

$$\|R_k(f)\| \leq \frac{B}{\left(1 + \sum_{j=1}^{k} t_j^2\right)^{t_k/(2(2+t_k))}}. \tag{6}$$

The noisy sample variants of the above algorithms simply replace $\langle f, X_j\rangle$ with $\langle Y, X_j\rangle_n = n^{-1}\sum_i Y_i X_{ij}$.

We already covered the sample counterparts of orthogonal greedy or matching pursuit. The weak version translates similarly. The sample counterpart of (weak) greedy is also called matching pursuit. In slight variant, called L_2 boosting, at step k, find $g_k \in \mathcal{D}$ as

$$\arg\min_{h\in\mathcal{D}} \min_{\alpha\in\mathbb{R}} \|R_{k-1}(f) - \alpha h\|_2^2.$$

Theorem 9 *The sample variant of the matching pursuit estimator is linear. In particular,*

$$\widehat{Y}^{(k)} = B_k Y \tag{7}$$

where $\widehat{Y}^{(k)} = (\widehat{m}^{(k)}(X_1), \ldots, \widehat{m}^{(k)}(X_n))^T$,

$$B_k = I - (I - H_k)(I - H_{k-1})\cdots(I - H_1), \tag{8}$$

and

$$H_j = \frac{\bar{\Psi}_j \bar{\Psi}_j^T}{\|\mathbb{X}_j\|^2}, \tag{9}$$

where $\bar{\Psi}_j = (\Psi_j(X_1), \ldots, \Psi_j(X_n))$.

Theorem 10 (Bühlmann 2005) *Let $m_n(x) = \sum_{j=1}^{d_n} \beta_{j,n}\psi_j(x)$ be the best linear approximation based on d_n terms. Suppose that:*

(A1 Growth) $d_n \leq C_0 e^{C_1 n^{1-\xi}}$ for some $C_0, C_1 > 0$ and some $0 < \xi \leq 1$.

(A2 Sparsity) $\sup_n \sum_{j=1}^{d_n} |\beta_{j,n}| < \infty$.

(A3 Bounded Covariates) $\sup_n \max_{1\leq j\leq d_n} \max_i |\Psi_j(X_i| < \infty$ with probability 1.

(A4 Moments) $\mathbb{E}|\epsilon|^s < \infty$ for some $s > 4/\xi$.

Then there exists $k_n \to \infty$ such that

$$\mathbb{E}_X|\widehat{m}_n(X) - m_n(x)|^2 \to 0 \tag{10}$$

as $n \to 0$.

As a sketch of the proof, recall that the noisy sample variant of L_2 boosting replaces $\langle f, X_j\rangle$ with $\langle Y, X_j\rangle_n = n^{-1}\sum_i Y_i X_{ij}$.

Note that $\langle Y, X_j\rangle_n$ has mean $\langle f, X_j\rangle$. The main burden of the proof is to show that $\langle Y, X_j\rangle_n$ is close to $\langle f, X_j\rangle$ with high probability and then apply Temlyakov's result. For this we use Bernstein's inequality. Recall that if $|Z_j|$ are bounded by M and Z_j has variance σ^2 then

$$\mathbb{P}(|\overline{Z} - \mathbb{E}(Z_j)| > \epsilon) \leq 2\exp\left\{-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + M\epsilon/3}\right\}. \tag{11}$$

Hence, the probability that any empirical inner products differ from their functional counterparts is no more than

$$d_n^2 \exp\left\{-\frac{1}{2}\frac{n\epsilon^2}{\sigma^2 + M\epsilon/3}\right\} \to 0 \tag{12}$$

because of the growth condition.

5 Additive Models for Classification: Boosting

One could use such greedy estimation of additive models for general losses, beyond just the squared loss.

Boosting (at least originally in the context of machine learning) refers to a class of methods that build classifiers in a greedy, iterative way. The original boosting algorithm is called *AdaBoost* and is due to Freund and Schapire (1996). See Figure 3.

The algorithm seems mysterious and there is still room to understand why (and when) it works. Perhaps the most compelling explanation is due to Friedman, Hastie and Tibshirani

```
1. Input: $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $Y_i \in \{-1, +1\}$.
2. Set $w_i = 1/n$ for $i = 1, \ldots, n$.
3. Repeat for $m = 1, \ldots, M$.
    (a) Compute the weighted error $\epsilon(h) = \sum_{i=1}^n w_i I(Y_i \neq h(X_i)$ and find $h_m$ to minimize $\epsilon(h)$.
    (b) Let $\alpha_m = (1/2)\log((1-\epsilon)/\epsilon)$.
    (c) Update the weights:
            $$w_i \leftarrow \frac{w_i e^{-\alpha_m Y_i h_m(X_i)}}{Z}$$
        where $Z$ is chosen so that the weights sum to 1.
4. The final classifier is
            $$h(x) = \text{sign}\left(\sum_{m=1}^M \alpha_m h_m(x)\right).$$
```

Figure 3: AdaBoost

(2000), which essentially derives boosting as *greedy function approximation*, along the lines of previous sections.

In this section, we assume that $Y_i \in \{-1, +1\}$. Many classifiers then have the form

$$h(x) = \text{sign}(H(x))$$

for some function $H(x)$. For example, a linear classifier corresponds to $H(x) = \beta^T x$. The risk can then be written as

$$R(h) = \mathbb{P}(Y \neq h(X)) = \mathbb{P}(YH(X) < 0) = \mathbb{E}(L(A))$$

where $A = YH(X)$ and $L(a) = I(a < 0)$. As a function of a, the loss $L(a)$ is discontinuous which makes it difficult to work with. Friedman, Hastie and Tibshirani (2000) show that AdaBoost corresponds to using a surrogate loss, namely, $L(a) = e^{-a} = e^{-yH(x)}$. Consider finding a classifier of the form $\sum_m \alpha_m h_m(x)$ by minimizing the exponential loss $\sum_i e^{-Y_i H(X_i)}$. If we do this iteratively, adding one function at a time, this leads precisely to AdaBoost. Typically, the classifiers h_j in the sum $\sum_m \alpha_m h_m(x)$ are taken to be very simple classifiers such as small classification trees.

The argument in Friedman, Hastie and Tibshirani (2000) is as follows. Consider minimizing the expected loss $J(F) = \mathbb{E}(e^{-YF(X)})$. Suppose our current estimate is F and consider updating to an improved estimate $F(x) + cf(x)$.

$$\begin{aligned} J(F + cf) &= \mathbb{E}(e^{-Y(F(X)+cf(X))}) \approx \mathbb{E}(e^{-YF(X)}(1 - cYf(X) + c^2Y^2f^2(X)/2)) \\ &= \mathbb{E}(e^{-YF(X)}(1 - cYf(X) + c^2/2)) \end{aligned}$$

since $Y^2 = f^2(X) = 1$. Now consider minimizing the latter expression a fixed $X = x$. If we minimize over $f(x) \in \{-1, +1\}$ we get $f(x) = 1$ if $E_w(y|x) > 0$ and $f(x) = -1$ if $E_w(y|x) < 0$ where $E_w(y|x) = E(w(x,y)y|x)/E(w(x,y)|x)$ and $w(x,y) = e^{-yF(x)}$. In other words, the optimal f is simply the Bayes classifier with respect to the weights. This is exactly the first step in AdaBoost. If we fix now fix $f(x)$ and minimize over c we get

$$c = \frac{1}{2}\log\left(\frac{1-\epsilon}{\epsilon}\right)$$

where $\epsilon = E_w(I(Y \neq f(x)))$. Thus the updated $F(x)$ is

$$F(x) \leftarrow F(x) + cf(x)$$

as in AdaBoost. When we update F this way, we change the weights to

$$w(x,y) \leftarrow w(x,y)e^{-cf(x)y} = w(x,y)\exp\left(\log\left(\frac{1-\epsilon}{\epsilon}\right)I(Y \neq f(x))\right)$$

which again is the same as AdaBoost.

Seen in this light, boosting is really greedy function approximation, given a surrogate to the zero-one loss (namely, the exponential loss).