COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Maryam Bahrani

Lecture # 18
April 11, 2018

In this lecture, we introduce the online linear regression problem. We give an algorithm for this problem due to Widrow and Hoff, along with its analysis. Finally, as part of a new topic, we explore the relationship between Online Learning and Batch learning, giving a possible reduction from the latter to the former.

1 Online Linear Regression

The goal of online linear regression is to minimize the square loss of a linear function in an online setting, according to the following framework:

- Initialize $\mathbf{w}_1 = \mathbf{0}$
- For each round $t = 1, \ldots, T$:
  - Get $\mathbf{x}_t \in \mathbb{R}^n$
  - Predict $\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t \in \mathbb{R}$
  - Observe $y_t \in \mathbb{R}$
  - Update $\mathbf{w}_t$.

We have the following notions of "loss" for this algorithm. The square loss in round t is given by $(\hat{y}_t - y_t)^2$. The cumulative loss of an algorithm A, denoted by L_A, is the sum of the losses in individual rounds, *i.e.* $L_A = \sum_{t=1}^{T} (\hat{y}_t - y_t)^2$. The loss of a specific weight vector $\mathbf{u}$ is given by $L_{\mathbf{u}} = \sum_{t=1}^{T} (\mathbf{u} \cdot \mathbf{x}_t - y_t)^2$. For an algorithm A to have good performance, we would like it to satisfy an inequality of the form

$$L_A \leqslant \min_{\mathbf{u}} L_{\mathbf{u}} + \text{small number}.$$

The vector $\mathbf{u}$ that minimizes $L_{\mathbf{u}}$ describes the best weight vector that we could have picked if we knew all the data points *offline*. We can think of this smallest achievable loss as the clairvoyant loss. This inequality ensures that the cumulative loss of our online algorithm does not exceed the clairvoyant loss by more than a small amount.

One possible instantiation of this framework is the following algorithm:

- Initialize $\mathbf{w}_1 = \mathbf{0}$
- Choose parameter $\eta > 0$.
- For each round $t = 1, \ldots, T$:
  - Get $\mathbf{x}_t \in \mathbb{R}^n$
  - Predict $\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t \in \mathbb{R}$
  - Observe $y_t \in \mathbb{R}$
  - Update $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t)\mathbf{x}_t$.

The blue lines are what was changed from the template above. This algorithm is called *Widrow-Hoff (WH)* after its inventors, as well as *Least Mean Squares (LMS)*.

1.1 Motivation for Update Function

Before analyzing this algorithm, we will motivate why the particular weight update function $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t)\mathbf{x}_t$ was used. We provide two motivations.

Motivation 1. The first motivation is related to performing Gradient Descent on the loss function. Remember that the loss of a weight vector $\mathbf{w}$ on one example $(\mathbf{x}, y)$ is given by $L(\mathbf{w}, \mathbf{x}, y) = (\mathbf{w} \cdot \mathbf{x} - y)^2$. At a minimum, we expect a good update rule to decrease the loss on the most recently seen example (*i.e.* $(\mathbf{x}_t, y_t)$ at round t). We know that the gradient of a continuous, differentiable function points in the direction of fastest *increase*. Therefore, it is a natural idea to *decrease* a function by taking a small step in the opposite direction of the gradient. The gradient of the loss function is given by

$$\nabla_{\mathbf{w}} L(\mathbf{x}, y) = \begin{pmatrix} \partial L/\partial w_1 \\ \partial L/\partial w_2 \\ \vdots \\ \partial L/\partial w_n \end{pmatrix} = 2(\mathbf{w} \cdot \mathbf{x} - y)\mathbf{x}$$

where n is the number of dimensions.

A good update rule is thus

$$\mathbf{w}_{t+1} = \mathbf{w}_t - (\text{constant}) \cdot \nabla_{\mathbf{w}} L(\mathbf{x}, y) = \mathbf{w}_t - (\text{constant}) \cdot (\mathbf{w} \cdot \mathbf{x} - y)\mathbf{x},$$

which matches the update rule of the algorithm.

Motivation 2. Ideally, we want the new weight vector $\mathbf{w}_{t+1}$ to

1. achieve smaller loss on the current example $(\mathbf{x}_t, y_t)$ — so we want a small $L(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t)$;
2. stay close to $\mathbf{w}_t$, since $\mathbf{w}_t$ embodies all the training examples we have seen so far and we don't want to throw away all the progress we have made — so we want a small $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2$.

Since we want to minimize the above two things simultaneously, it is natural to minimize their weighted sum

$$\eta(\mathbf{w}_{t+1} \cdot \mathbf{x} - y_t)^2 + \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2 . \tag{1}$$

Solving the above optimization for $\mathbf{w}_{t+1}$, we get

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

This is almost the same as the update rule in the algorithm, except that on the right hand side we have a $\mathbf{w}_{t+1}$ instead of a $\mathbf{w}_t$. However, since we want $\mathbf{w}_{t+1}$ and $\mathbf{w}_t$ to be close to each other any way, we can use $\mathbf{w}_t$ as an approximation for $\mathbf{w}_{t+1}$ on the right-hand-side, giving us exactly the update rule used in the algorithm.

1.2 Analysis

In this section, we provide an analysis for the online linear regression algorithm described above.

Theorem 1. *If $\|\mathbf{x}_t\|_2 \leqslant 1$ for all rounds t, then the following bound holds for L_{WH}, the cumulative loss of the Widrow-Hoff algorithm:*

$$L_{WH} \leqslant \min_{\mathbf{u}\in\mathbb{R}^n} \left(\frac{L_{\mathbf{u}}}{1-\eta} + \frac{\|\mathbf{u}\|_2^2}{\eta} \right).$$

Before going into the proof, we list a few remarks.

- The dependence on the length of $\mathbf{u}$ is needed because it is sometimes possible to make $L_{\mathbf{u}}$ arbitrarily small by making $\mathbf{u}$ larger, which is not desirable. An alternative way of addressing the dependence on length is to take the minimum over all vectors $\mathbf{u}$ with bounded length.
- One can roughly think of $\mathbf{u}$ as the best vector that could have been chosen even if all the data points were known in advance (except there is an additional term that depends on the length of $\mathbf{u}$ so this analogy is not completely accurate).
- If we show the bound is true for all $\mathbf{u}$, it follows that it must be true for the "best" $\mathbf{u}$ (*i.e.* the one minimizing the right-hand side). Fix some $\mathbf{u} \in \mathbb{R}^n$, and divide both sides of the inequality by T to get

  $$\frac{L_{WH}}{T} \leqslant \frac{L_{\mathbf{u}}}{T}\frac{1}{1-\eta} + \frac{\|\mathbf{u}\|_2^2}{\eta T}.$$

  For sufficiently small η, $\frac{1}{1-\eta}$ is close to 1. Furthermore, as $T \to \infty$, $\frac{\|\mathbf{u}\|_2^2}{\eta T} \to 0$. It follows that as $T \to \infty$, the *rate* of loss of Widrow-Hoff (L_{WH}/T) approaches the *rate* of loss of the "best" vector $\mathbf{u}$ ($L_{\mathbf{u}}/T$).

Proof. The proof uses a potential function argument. We need to establish some notation before giving the proof.

- As we argued above, it is enough to show the bound holds for all vectors $\mathbf{u}$ to conclude that the bound holds for the best such vector as well. For the rest of the proof, we fix some $\mathbf{u} \in \mathbb{R}^n$.
- Define $\Phi_t = \|\mathbf{w}_t - \mathbf{u}\|_2^2$ to be the potential at round t. The closer $\mathbf{w}_t$ and $\mathbf{u}$, the lower the potential.
- Let $\ell_t = \mathbf{w}_t \cdot \mathbf{x}_t - y_t = \hat{y}_t - y_t$. With this notation, ℓ_t^2 denotes the loss of Widrow-Hoff at round t.
- Let $g_t = \mathbf{u} \cdot \mathbf{x}_t - y_t$. Similarly, g_t^2 denotes the loss of the weight vector $\mathbf{u}$ at round t.
- Let $\boldsymbol{\Delta}_t = \eta(\hat{y}_t - y_t)\mathbf{x}_t = \eta\ell_t\mathbf{x}_t$. It follows that $\mathbf{w}_{t+1} = \mathbf{w}_t - \boldsymbol{\Delta}_t$, so $\boldsymbol{\Delta}_t$ measures the change in the weight vector from round t to round $t+1$.

Equipped with the notation above, we can make the following claim:

Claim. $\Phi_{t+1} - \Phi_t \leqslant -\eta \cdot \ell_t^2 + \frac{\eta}{1-\eta} g_t^2$.
Roughly speaking, the first term $-\eta \cdot \ell_t^2$ corresponds to the loss of the Widrow-Hoff learner — as the learner suffers loss, the potential goes down. The second term $\frac{\eta}{1-\eta} g_t^2$ corresponds to the loss of $\mathbf{u}$ — as $\mathbf{u}$ suffers loss, the potential goes up. The potential can thus be thought of as a budget of how much loss the learner is allowed to suffer to not fall behind $\mathbf{u}$ too much.

Proof of Claim. We have

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \|\mathbf{w}_{t+1} - \mathbf{u}\|_2^2 - \|\mathbf{w}_t - \mathbf{u}\|_2^2 && (2)\\
&= \|\mathbf{w}_t - \mathbf{u} - \boldsymbol{\Delta}_t\|_2^2 - \|\mathbf{w}_t - \mathbf{u}\|_2^2 && (3)\\
&= \|\mathbf{w}_t - \mathbf{u}\|_2^2 - 2(\mathbf{w}_t - \mathbf{u}) \cdot \boldsymbol{\Delta}_t + \|\boldsymbol{\Delta}_t\|_2^2 - \|\mathbf{w}_t - \mathbf{u}\|_2^2 && (4)\\
&= -2(\mathbf{w}_t - \mathbf{u}) \cdot \boldsymbol{\Delta}_t + \|\boldsymbol{\Delta}_t\|_2^2 && (5)\\
&= -2\eta\ell_t(\mathbf{w}_t \cdot \mathbf{x}_t - \mathbf{u} \cdot \mathbf{x}_t) + \eta^2 \ell_t^2 \|\mathbf{x}_t\|_2^2 && (6)\\
&\leqslant -2\eta\ell_t(\mathbf{w}_t \cdot \mathbf{x}_t - y_t + y_t - \mathbf{u} \cdot \mathbf{x}_t) + \eta^2\ell_t^2 && (7)\\
&= -2\eta\ell_t(\ell_t - g_t) + \eta^2\ell_t^2 && (8)\\
&= \eta^2\ell_t^2 - 2\eta\ell_t^2 + 2\eta\ell_t g_t && (9)\\
&\leqslant \eta^2\ell_t^2 - 2\eta\ell_t^2 + \eta\left(\frac{g_t^2}{1-\eta} + \ell_t^2(1-\eta)\right) && (10)\\
&= -\eta \cdot \ell_t^2 + \frac{\eta}{1-\eta} g_t^2. && (11)
\end{aligned}$$

Equality (2) follows from the definition of Φ_t. Equality (3) holds by definition of $\boldsymbol{\Delta}_t$. Equality (4) is obtained by expanding the first squared term. Equality (5) is due to the cancellation of the first and last terms. Equality (6) is obtained by replacing $\boldsymbol{\Delta}_t$ with $\eta\ell_t\mathbf{x}_t$. Equality (6) results from multiplying out the first product. Inequality (7), we have added and subtracted y_t to the first term, and used the assumption that $\|\mathbf{x}_t\|_2 \leqslant 1$. Inequality (8), we have used the definitions of ℓ_t and g_t. Equality (9) is simple algebra. Inequality (10) uses the following trick: $ab \leqslant \frac{a^2+b^2}{2}$ for all a, b, so it specifically holds for $a = \frac{g_t}{\sqrt{1-\eta}}$ and $b = \ell_t \cdot \sqrt{1-\eta}$. Lastly, Equality (11) is a result of a few simple cancellations, and gives us exactly the desired bound in the claim.

The Theorem follows from the claim because

$$
\begin{aligned}
-\|\mathbf{u}\|_2^2 &= -\Phi_1 && (12)\\
&\leqslant \Phi_{T+1} - \Phi_1 && (13)\\
&= \Phi_{T+1} - \Phi_T + \Phi_T - \Phi_{T-1} + \Phi_{T-1} - \ldots + \Phi_2 - \Phi_1 && (14)\\
&= \sum_{t=1}^{T} (\Phi_{t+1} - \Phi_t) && (15)\\
&\leqslant \sum_{t=1}^{T} \left(-\eta \cdot \ell_t^2 + \frac{\eta}{1-\eta} g_t^2 \right) && (16)\\
&= -\eta \sum_{t=1}^{T} \ell_t^2 + \frac{\eta}{1-\eta} \sum_{t=1}^{T} g_t^2 && (17)\\
&= -\eta L_{WH} + \frac{\eta}{1-\eta} L_{\mathbf{u}}. && (18)
\end{aligned}
$$

Equality (12) holds because $\mathbf{w}_1 = \mathbf{0}$ and therefore $\Phi_1 = \|\mathbf{w}_1 - \mathbf{u}\|_2^2 = \|\mathbf{0} - \mathbf{u}\|_2^2 = \|\mathbf{u}\|_2^2$. Equality (13) holds because Φ_{T+1} is a norm and therefore non-negative. Equality (14) holds because each new term is added once and subtracted once. Inequality (16) holds by the claim shown above. Equality (17) is obtained by distributing the sum. Equality (18) follows from the observation that the cumulative loss of Widrow-Hoff (respectively $\mathbf{u}$) is the sum of the losses at every round, *i.e.* $L_{WH} = \sum_{t=1}^{T} \ell_t^2$ and $L_{\mathbf{u}} = \sum_{t=1}^{T} g_t^2$.

Solving for L_{WH}, the above inequality gives exactly the statement of the theorem. □

1.3 Generalizing the Motivation

As explained in Motivation 2, the goal of online regression was to minimize

$$\eta \cdot (\textit{loss of } \mathbf{w} \text{ on } (\mathbf{x}_t, y_t)) + (\text{“distance” between } \mathbf{w}_{t+1} \text{ and } \mathbf{w}_t).$$

In particular, we chose the *square loss* function and the *Euclidean distance squared*. However, there are many different notions of loss and distance that can be used in the above measure.

For example, for an arbitrary loss function $L(\mathbf{w}, \mathbf{x}, y)$ and the Euclidean norm squared as a measure of distance, we get the following update rule:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla_{\mathbf{w}} L(\mathbf{w}_t, \mathbf{x}_t, y_t).$$

Note that this update rule is simply performing gradient descent on the loss function.

Another possibility is an arbitrary loss function $L(\mathbf{w}, \mathbf{x}, y)$ along with *relative entropy* as a measure of distance. More formally, we can restrict the $\mathbf{w}$'s to be probability distributions, and try to minimize $\mathrm{RE}(\mathbf{w}_t || \mathbf{w}_{t+1})$ in every step. The weight update function then becomes

$$w_{t+1,i} = \frac{\mathbf{w}_{t,i} \cdot e^{-\eta \frac{\partial L}{\partial w_i}(\mathbf{w}_t, \mathbf{x}_t, y_t)}}{Z_t}$$

where Z_t is a normalization factor to make sure $\mathbf{w}_{t+1}$ is a probability distribution. This update rule is referred to as *Exponentiated Gradient (EG)*.

Note that in the case of Euclidean norm squared as a distance measure, the update function is additive, whereas in the case of Relative Entropy, the update rule is multiplicative. The following table summarizes the update rules we have seen so far.

Additive	Multiplicative
Support Vector Machine (SVM)	AdaBoost
Perceptron	Winnow / Weighted Majority Algorithm (WMA)
Gradient Descent (GD)	Exponentiated Gradient (EG)

2 Relating Batch Learning and Online Learning

So far in the class, we have discussed the following two learning models:

- **Batch learning**, including the PAC model, where we are given a set of random examples offline, and our goal is to minimize the generalization error.
- **Online Learning**, where we are given a stream of possibly adversarial examples, and our goal is to minimize cumulative loss (so in a sense training and testing are mixed together).

It is natural to ask whether these two models are related. Intuitively, the online setting is stronger, since it requires no randomness assumption. So we can ask whether batch learning can be reduced to online learning, *i.e.* given an online learning algorithm, could we use it to learn in the batch setting? Besides theoretical interest, this reduction turns out to have practical applications, as online algorithms are often used for offline learning tasks.

More formally, we get a set of examples $S = \langle(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\rangle$ drawn *i.i.d* from a distribution D as training data. We then get a test example $(\mathbf{x}, y)$. Define the *risk* (expected loss) of vector $\mathbf{v}$ to be $R_{\mathbf{v}} = \mathbb{E}_{(\mathbf{x},y)\sim D}\left[(\mathbf{v}\cdot\mathbf{x} - y)^2\right]$. The goal is to use an online learning algorithm to find $\mathbf{v}$ with small risk.

It turns out that Widrow-Hoff and its analysis yield both an efficient algorithm for this problem, and an immediate means of bounding the risk of the vector that it finds. In particular, we propose the following algorithm for solving this problem:

- Run Widrow-Hoff for $T = m$ rounds on $S = \langle(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2)\ldots, (\mathbf{x}_m, y_m)\rangle$, in the random order they were given to us.
- Windrow-Hoff produces a sequence of weight vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m$
- We output the average of these vectors $\mathbf{v} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{w}_i$, (instead of the last one).

The following theorem shows that the *expected* risk of the output vector is low. One can also show that the risk is low with high probability, but we do not do so in today's lecture.

Theorem 2.

$$\mathbb{E}_S[R_{\mathbf{v}}] \leqslant \min_{\mathbf{u}}\left(\frac{R_{\mathbf{u}}}{1-\eta} + \frac{\|\mathbf{u}\|_2^2}{\eta m}\right),$$

where $\mathbb{E}_S[\cdot]$ denotes expectation with respect to the random choice of the sample S.

Proof. Fix any vector $\mathbf{u} \in \mathbb{R}^n$, and let $(\mathbf{x}, y)$ be a random test point drawn from the distribution D. We write $\mathbb{E}[\cdot]$, with no subscript, to denote expectation with respect to both the random sample S and the random test point $(\mathbf{x}, y)$.

The following three observations will be needed:

Observation 1. $(\mathbf{v}\cdot\mathbf{x}-y)^2 \leqslant \frac{1}{m}\sum_{t=1}^{m}(\mathbf{w_t}\cdot\mathbf{x}-y)^2$.
This is due to Jensen's inequality, since $f(x)=x^2$ is a convex function:

$$(\mathbf{v}\cdot\mathbf{x}-y)^2 = \left(\frac{1}{m}\sum_{t=1}^{m}(\mathbf{w_t}\cdot\mathbf{x}-y)\right)^2 \leqslant \frac{1}{m}\sum_{t=1}^{m}(\mathbf{w_t}\cdot\mathbf{x}-y)^2.$$

Observation 2. $\mathbb{E}\left[(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2\right] = \mathbb{E}\left[(\mathbf{u}\cdot\mathbf{x}-y)^2\right]$.
This is because $(\mathbf{x}_t,y_t)$ and $(\mathbf{x},y)$ come from the same distribution.

Observation 3. $\mathbb{E}\left[(\mathbf{w}_t\cdot\mathbf{x}_t-y_t)^2\right] = \mathbb{E}\left[(\mathbf{w}_t\cdot\mathbf{x}-y)^2\right]$.
This is because $(\mathbf{x}_t,y_t)$ and $(\mathbf{x},y)$ come from the same distribution, and $\mathbf{w}_t$ is independent of both $(\mathbf{x}_t,y_t)$ and $(\mathbf{x},y)$, since it was trained only on $(\mathbf{x}_1,y_1),\ldots(\mathbf{x}_{t-1},y_{t-1})$ and thus depends only on the randomness of the first $t-1$ examples.

Putting these observations together, we have

$$\begin{aligned}
\mathbb{E}_S[R_{\mathbf{v}}] &= \mathbb{E}\left[(\mathbf{v}\cdot\mathbf{x}-y)^2\right] && (19)\\
&\leqslant \mathbb{E}\left[\frac{1}{m}\sum_{t=1}^{m}(\mathbf{w_t}\cdot\mathbf{x}-y)^2\right] && (20)\\
&= \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left[(\mathbf{w_t}\cdot\mathbf{x}-y)^2\right] && (21)\\
&= \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left[(\mathbf{w_t}\cdot\mathbf{x_t}-y_t)^2\right] && (22)\\
&= \frac{1}{m}\mathbb{E}\left[\sum_{t=1}^{m}(\mathbf{w_t}\cdot\mathbf{x_t}-y_t)^2\right] && (23)\\
&\leqslant \frac{1}{m}\mathbb{E}\left[\frac{\sum_{t=1}^{m}(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2}{1-\eta}+\frac{\|\mathbf{u}\|_2^2}{\eta}\right] && (24)\\
&= \frac{1}{m}\frac{\sum_{t=1}^{m}\mathbb{E}\left[(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2\right]}{1-\eta}+\frac{\|\mathbf{u}\|_2^2}{\eta m} && (25)\\
&= \frac{1}{m}\frac{\sum_{t=1}^{m}\mathbb{E}\left[(\mathbf{u}\cdot\mathbf{x}-y)^2\right]}{1-\eta}+\frac{\|\mathbf{u}\|_2^2}{\eta m} && (26)\\
&= \frac{R_{\mathbf{u}}}{1-\eta}+\frac{\|\mathbf{u}\|_2^2}{\eta m}, && (27)
\end{aligned}$$

with all the expectations taken over the random sample S and the random test point $(\mathbf{x},y)$.

Equality (19) holds by definition of risk of $\mathbf{v}$. Inequality (20) is an application of Observation 1. Equality (21) is by linearity of expectation. Equality (22) follows from Observation 3. Equality (23) is by linearity of expectation. Inequality (24) holds because the term inside the expectation is exactly the cumulative loss of Widrow-Hoff, to which we can apply the bound from the first part of lecture. Equality (25) holds by linearity of expectation and pulling out the constant terms. Equality (26) holds by Observation 2, and Equality (27) is true by definition of risk of $\mathbf{u}$. □