

Online Learning

Frederic Koriche

LIRMM, Université Montpellier II

Parcours Intelligence Artificielle
Master Informatique 2ème Année
Septembre 2011

Outline

1 Online Learning

2 Convex Analysis

- Convex Sets
- Convex Functions
- Loss Functions

3 Regression Learning

- The Protocol
- The Gradient-Descent Algorithm
- The Exponentiated-Gradient Algorithm

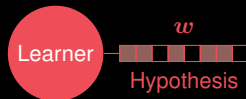
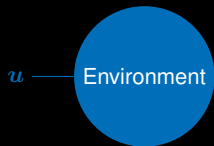
4 Classification Learning

- The Protocol
- The Perceptron Algorithm
- The Winnow Algorithm



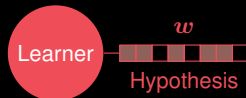
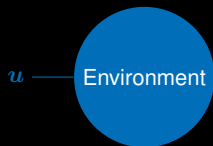
Online Learning

Repeated zero-sum game between the *learner* and its *environment*



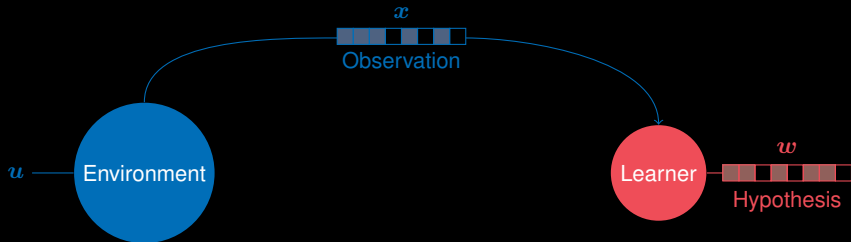
Initialization

- The environment chooses its model u
- The learner chooses its initial hypothesis w



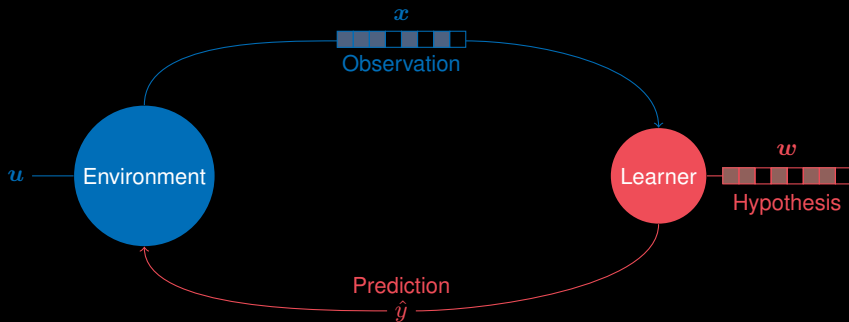
Trials (Rounds)

- *Observation*: the learner receives an observation x from its environment.
- *Decision*: the learner takes a decision \hat{y} .
- *Response*: the learner receives a reinforcement y .
- *Update*: the learner incurs a loss $\ell(\hat{y}, y)$ and updates its hypothesis w .



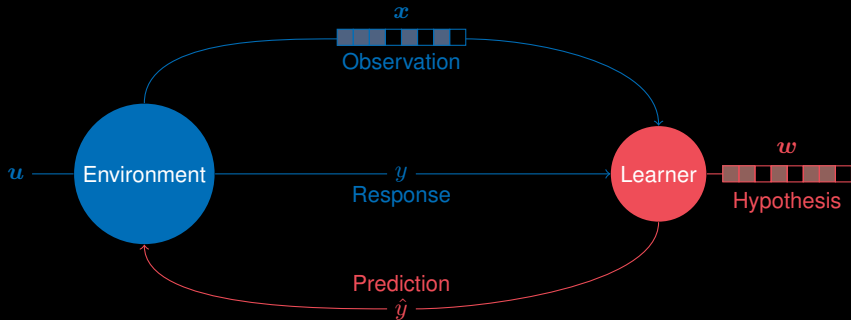
Trials (Rounds)

- *Observation*: the learner receives an observation x from its environment.
- *Decision*: the learner takes a decision \hat{y} .
- *Response*: the learner receives a reinforcement y .
- *Update*: the learner incurs a loss $\ell(\hat{y}, y)$ and updates its hypothesis w .



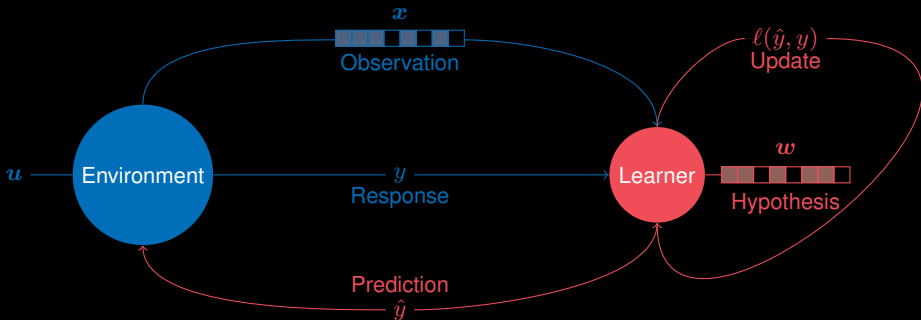
Trials (Rounds)

- *Observation*: the learner receives an observation x from its environment.
- *Decision*: the learner takes a decision \hat{y} .
- *Response*: the learner receives a reinforcement y .
- *Update*: the learner incurs a loss $\ell(\hat{y}, y)$ and updates its hypothesis w .



Trials (Rounds)

- *Observation*: the learner receives an observation x from its environment.
- *Decision*: the learner takes a decision \hat{y} .
- *Response*: the learner receives a reinforcement y .
- *Update*: the learner incurs a loss $\ell(\hat{y}, y)$ and updates its hypothesis w .



Trials (Rounds)

- *Observation*: the learner receives an observation x from its environment.
- *Decision*: the learner takes a decision \hat{y} .
- *Response*: the learner receives a reinforcement y .
- *Update*: the learner incurs a loss $\ell(\hat{y}, y)$ and updates its hypothesis w .

Outline

1 Online Learning

2 Convex Analysis

- Convex Sets
- Convex Functions
- Loss Functions

3 Regression Learning

- The Protocol
- The Gradient-Descent Algorithm
- The Exponentiated-Gradient Algorithm

4 Classification Learning

- The Protocol
- The Perceptron Algorithm
- The Winnow Algorithm

Convex Combination

A point \mathbf{x} is a *convex combination* of $\mathbf{x}_1, \dots, \mathbf{x}_k$ if there are scalars w_1, \dots, w_k such that

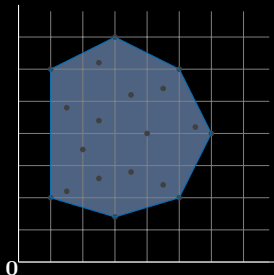
- $w_1 + \dots + w_k = 1$,
- $w_i \geq 0, i = 1, \dots, k$ and
- $\mathbf{x} = w_1 \mathbf{x}_1 + \dots + w_k \mathbf{x}_k$

Convex Hull

The space $\text{conv } S$ of all convex combinations of a set $S \subseteq \mathbb{R}^n$ is the *convex hull* of S .

Convex Set

A nonempty set $S \subseteq \mathbb{R}^n$ is *convex* if $\text{conv } S = S$.



Convex Combination

A point \mathbf{x} is a *convex combination* of $\mathbf{x}_1, \dots, \mathbf{x}_k$ if there are scalars w_1, \dots, w_k such that

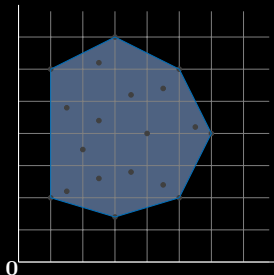
- $w_1 + \dots + w_k = 1$,
- $w_i \geq 0, i = 1, \dots, k$ and
- $\mathbf{x} = w_1 \mathbf{x}_1 + \dots + w_k \mathbf{x}_k$

Convex Hull

The space $\text{conv } S$ of all convex combinations of a set $S \subseteq \mathbb{R}^n$ is the *convex hull* of S .

Convex Set

A nonempty set $S \subseteq \mathbb{R}^n$ is *convex* if $\text{conv } S = S$.



Convex Combination

A point \mathbf{x} is a *convex combination* of $\mathbf{x}_1, \dots, \mathbf{x}_k$ if there are scalars w_1, \dots, w_k such that

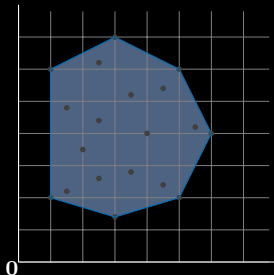
- $w_1 + \dots + w_k = 1$,
- $w_i \geq 0, i = 1, \dots, k$ and
- $\mathbf{x} = w_1 \mathbf{x}_1 + \dots + w_k \mathbf{x}_k$

Convex Hull

The space $\text{conv } S$ of all convex combinations of a set $S \subseteq \mathbb{R}^n$ is the *convex hull* of S .

Convex Set

A nonempty set $S \subseteq \mathbb{R}^n$ is *convex* if $\text{conv } S = S$.



Hyperplane

A *hyperplane* is a set in \mathbb{R}^n of the form

$$H = \{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} = b\}$$

where $\mathbf{a} \neq \mathbf{0}$ and $b \in \mathbb{R}$.

Halfspaces

A *halfspace* is a set of the form

$$S = \{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} \leq b\}$$

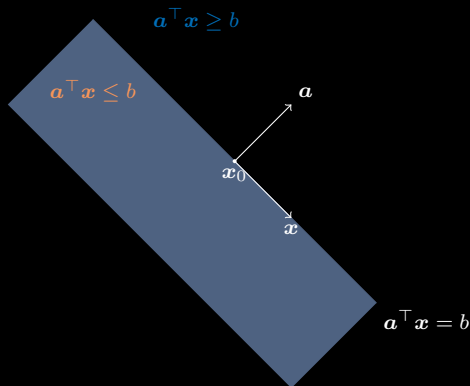
where $\mathbf{a} \neq \mathbf{0}$ and $b \in \mathbb{R}$.

Theorem of Separating Hyperplanes

Let P and N be two convex sets of points in \mathbb{R}^n which do not intersect. Then there are $\mathbf{a} \neq \mathbf{0}$ and b such that

- $\mathbf{a}^\top \mathbf{x} \leq b$ for all $\mathbf{x} \in P$, and
- $\mathbf{a}^\top \mathbf{x} \geq b$ for all $\mathbf{x} \in N$.

The set $\{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} = b\}$ is the *separating hyperplane* for P and N .



Hyperplane

A *hyperplane* is a set in \mathbb{R}^n of the form

$$H = \{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} = b\}$$

where $\mathbf{a} \neq \mathbf{0}$ and $b \in \mathbb{R}$.

Halfspaces

A *halfspace* is a set of the form

$$S = \{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} \leq b\}$$

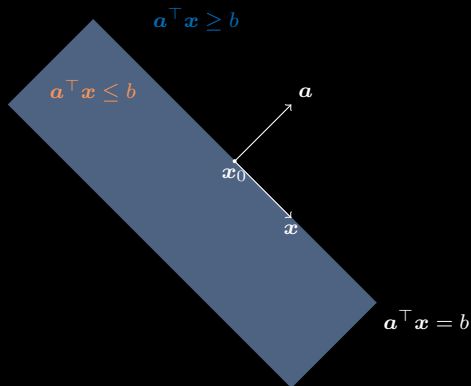
where $\mathbf{a} \neq \mathbf{0}$ and $b \in \mathbb{R}$.

Theorem of Separating Hyperplanes

Let P and N be two convex sets of points in \mathbb{R}^n which do not intersect. Then there are $\mathbf{a} \neq \mathbf{0}$ and b such that

- $\mathbf{a}^\top \mathbf{x} \leq b$ for all $\mathbf{x} \in P$, and
- $\mathbf{a}^\top \mathbf{x} \geq b$ for all $\mathbf{x} \in N$.

The set $\{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} = b\}$ is the *separating hyperplane* for P and N .



Hyperplane

A *hyperplane* is a set in \mathbb{R}^n of the form

$$H = \{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} = b\}$$

where $\mathbf{a} \neq \mathbf{0}$ and $b \in \mathbb{R}$.

Halfspaces

A *halfspace* is a set of the form

$$S = \{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} \leq b\}$$

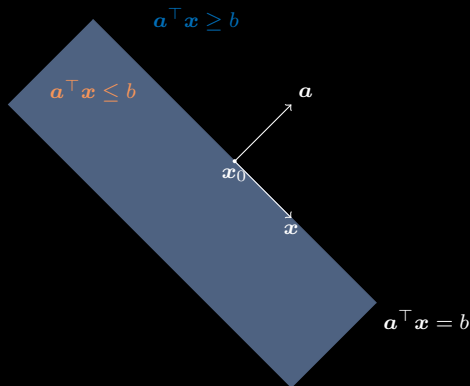
where $\mathbf{a} \neq \mathbf{0}$ and $b \in \mathbb{R}$.

Theorem of Separating Hyperplanes

Let P and N be two convex sets of points in \mathbb{R}^n which do not intersect. Then there are $\mathbf{a} \neq \mathbf{0}$ and b such that

- $\mathbf{a}^\top \mathbf{x} \leq b$ for all $\mathbf{x} \in P$, and
- $\mathbf{a}^\top \mathbf{x} \geq b$ for all $\mathbf{x} \in N$.

The set $\{\mathbf{x} \mid \mathbf{a}^\top \mathbf{x} = b\}$ is the *separating hyperplane* for P and N .



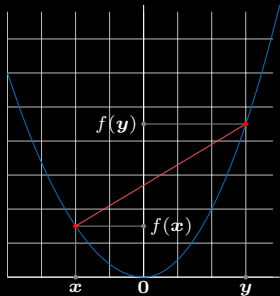
Convex Functions

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is *convex* if $\mathbf{dom} f$ is a convex set, and if for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f$ and $\theta \in [0, 1]$,

$$f(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta) f(\mathbf{y})$$

Examples

e^{px}	convex on \mathbb{R} for any $p \in \mathbb{R}$
$ x ^p$	convex on \mathbb{R} if $p \geq 1$
$x \log x$	convex on \mathbb{R}_+
$\ \mathbf{x}\ _p$	every norm on \mathbb{R}^n is convex



Gradient

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, and \mathbf{x} an interior point of $\text{dom } f$. If the partial derivatives $\frac{\partial f(\mathbf{x})}{\partial x_i}$ exist, we say that f is *differentiable* at \mathbf{x} . In this case, the *gradient* of f at \mathbf{x} is the vector

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

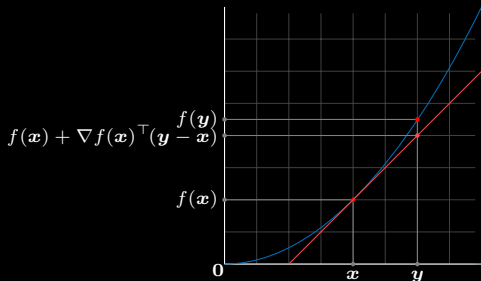
The *first-order approximation* of f at \mathbf{x} is the affine function of \mathbf{y}

$$\dot{f}(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

First-order conditions

Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable. Then f is convex if and only if $\text{dom } f$ is convex and, for all $\mathbf{x}, \mathbf{y} \in \text{dom } f$,

$$f(\mathbf{y}) \geq \dot{f}(\mathbf{y})$$



Gradient

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$, and \mathbf{x} an interior point of $\text{dom } f$. If the partial derivatives $\frac{\partial f(\mathbf{x})}{\partial x_i}$ exist, we say that f is *differentiable* at \mathbf{x} . In this case, the *gradient* of f at \mathbf{x} is the vector

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

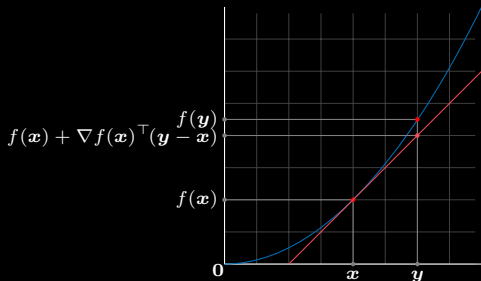
The *first-order approximation* of f at \mathbf{x} is the affine function of \mathbf{y}

$$\dot{f}(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$$

First-order conditions

Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is differentiable. Then f is convex if and only if $\text{dom } f$ is convex and, for all $\mathbf{x}, \mathbf{y} \in \text{dom } f$,

$$f(\mathbf{y}) \geq \dot{f}(\mathbf{y})$$



Loss Functions

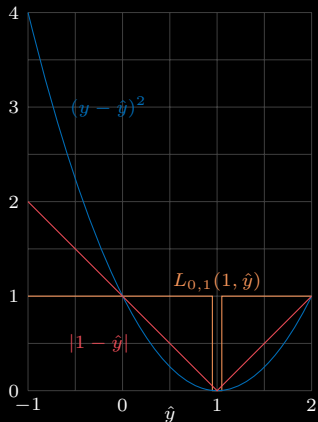
A *loss function* is a function $L : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ where $L(\hat{y}, y) = 0$ whenever $\hat{y} = y$.

Quadratic	$L_{\text{sq}}(\hat{y}, y) = (\hat{y} - y)^2$
Absolute	$L_{\text{ab}}(\hat{y}, y) = \hat{y} - y $
Zero-one	$L_{0,1}(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{otherwise} \end{cases}$

First-order conditions

If L is convex and differentiable in its first argument, then

$$L(\hat{y}, y) \leq L(\hat{z}, y) + \frac{\partial L(\hat{y}, y)}{\partial \hat{y}} (\hat{y} - \hat{z})$$



Loss Functions

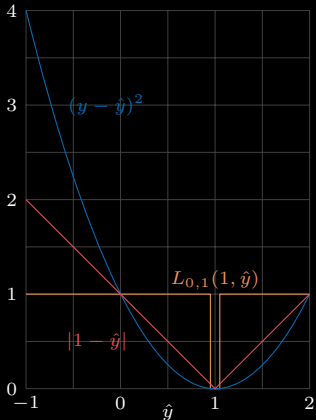
A *loss function* is a function $L : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ where $L(\hat{y}, y) = 0$ whenever $\hat{y} = y$.

Quadratic	$L_{\text{sq}}(\hat{y}, y) = (\hat{y} - y)^2$
Absolute	$L_{\text{ab}}(\hat{y}, y) = \hat{y} - y $
Zero-one	$L_{0,1}(\hat{y}, y) = \begin{cases} 0 & \text{if } \hat{y} = y \\ 1 & \text{otherwise} \end{cases}$

First-order conditions

If L is convex and differentiable in its first argument, then

$$L(\hat{y}, y) \leq L(\hat{z}, y) + \frac{\partial L(\hat{y}, y)}{\partial \hat{y}} (\hat{y} - \hat{z})$$



Outline

1 Online Learning

2 Convex Analysis

- Convex Sets
- Convex Functions
- Loss Functions

3 Regression Learning

- The Protocol
- The Gradient-Descent Algorithm
- The Exponentiated-Gradient Algorithm

4 Classification Learning

- The Protocol
- The Perceptron Algorithm
- The Winnow Algorithm

Regression Learning

Parameters: update function f .

Initialization: set \mathbf{w}_0 .

Trials: for each $t = 1, 2, \dots$

- (1) Observation: receive \mathbf{x}_t .
- (2) Prediction: output $\hat{y}_t = \mathbf{w}_{t-1}^\top \mathbf{x}_t$.
- (3) Response: get y_t and incur loss $L(\hat{y}_t, y_t)$.
- (4) Update: set $\mathbf{w}_t = f(\mathbf{w}_{t-1})$.

Best regressor

The hyperplane \mathbf{u} that minimizes

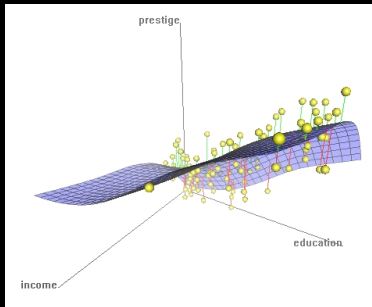
$$\mathbb{L}_{\text{sq}}(\mathbf{u}) = \sum_{t=1}^T L_{\text{sq}}(\hat{z}_t, y_t)$$

where $\hat{z}_t = \mathbf{u}^\top \mathbf{x}_t$

Goal

The goal of the learner is to minimize

$$\mathbb{L}_{\text{sq}}(\mathbf{w}) - \mathbb{L}_{\text{sq}}(\mathbf{u}) \quad \text{where} \quad \mathbb{L}_{\text{sq}}(\mathbf{w}) = \sum_{t=1}^T L(\hat{y}_t, y_t)$$



Regression Learning

Parameters: update function f .

Initialization: set \mathbf{w}_0 .

Trials: for each $t = 1, 2, \dots$

- (1) Observation: receive \mathbf{x}_t .
- (2) Prediction: output $\hat{y}_t = \mathbf{w}_{t-1}^\top \mathbf{x}_t$.
- (3) Response: get y_t and incur loss $L(\hat{y}_t, y_t)$.
- (4) Update: set $\mathbf{w}_t = f(\mathbf{w}_{t-1})$.

Best regressor

The hyperplane \mathbf{u} that minimizes

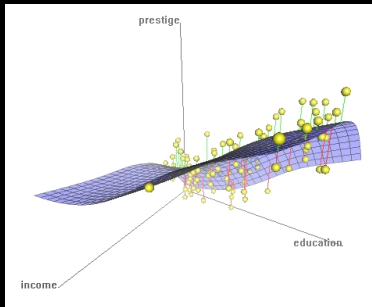
$$\mathbb{L}_{\text{sq}}(\mathbf{u}) = \sum_{t=1}^T L_{\text{sq}}(\hat{z}_t, y_t)$$

where $\hat{z}_t = \mathbf{u}^\top \mathbf{x}_t$

Goal

The goal of the learner is to minimize

$$\mathbb{L}_{\text{sq}}(\mathbf{w}) - \mathbb{L}_{\text{sq}}(\mathbf{u}) \quad \text{where} \quad \mathbb{L}_{\text{sq}}(\mathbf{w}) = \sum_{t=1}^T L(\hat{y}_t, y_t)$$



Regression Learning

Parameters: update function f .

Initialization: set \mathbf{w}_0 .

Trials: for each $t = 1, 2, \dots$

- (1) Observation: receive \mathbf{x}_t .
- (2) Prediction: output $\hat{y}_t = \mathbf{w}_{t-1}^\top \mathbf{x}_t$.
- (3) Response: get y_t and incur loss $L(\hat{y}_t, y_t)$.
- (4) Update: set $\mathbf{w}_t = f(\mathbf{w}_{t-1})$.

Best regressor

The hyperplane \mathbf{u} that minimizes

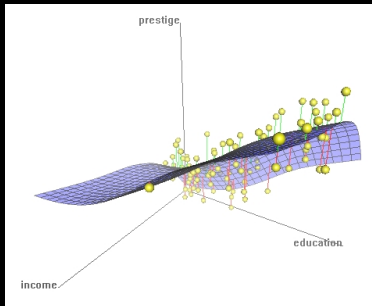
$$\mathbb{L}_{\text{sq}}(\mathbf{u}) = \sum_{t=1}^T L_{\text{sq}}(\hat{z}_t, y_t)$$

where $\hat{z}_t = \mathbf{u}^\top \mathbf{x}_t$

Goal

The goal of the learner is to minimize

$$\mathbb{L}_{\text{sq}}(\mathbf{w}) - \mathbb{L}_{\text{sq}}(\mathbf{u}) \quad \text{where} \quad \mathbb{L}_{\text{sq}}(\mathbf{w}) = \sum_{t=1}^T L(\hat{y}_t, y_t)$$



Gradient-Descent

Initialization: set $\mathbf{w}_0 = \mathbf{0}$.

Trials: for each $t = 1, 2, \dots$

- (1) Observation: receive \mathbf{x}_t
- (2) Prediction: output $\hat{y}_t = \mathbf{w}_{t-1}^\top \mathbf{x}_t$
- (3) Response: get y_t and incur loss $(\hat{y}_t - y_t)^2$
- (4) Update: set $\mathbf{w}_t = \mathbf{w}_{t-1} - \eta_t (\hat{y}_t - y_t) \mathbf{x}_t$

Convergence Theorem

For any best regressor $\mathbf{u} \in \mathbb{R}^n$, the quadratic loss of the gradient descent algorithm is bounded by

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 4X_2^2 U^2$$

where $X_2 = \max_{t=1}^T \|\mathbf{x}_t\|$, $U = \|\mathbf{u}\|$, and choosing $\eta_t = \frac{1}{2\|\mathbf{x}_t\|^2}$

$$L(\hat{y}_t, y_t) \leq L(\hat{z}_t, y_t) + 2(\hat{y}_t - y_t)(\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t)$$

Lemma

$$\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t \leq \frac{1}{2r_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) + \frac{r_t}{2} \|\mathbf{x}_t\|^2 \quad \text{where } r_t = \eta_t(\hat{y}_t - y_t)$$

Telescopic Sum

$$\sum_{t=1}^T \|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 = \|\mathbf{u} - \mathbf{w}_0\|^2 - \|\mathbf{u} - \mathbf{w}_T\|^2 \leq \|\mathbf{u} - \mathbf{w}_0\|^2 = \|\mathbf{u}\|^2$$

Combination

$$\begin{aligned} L(\hat{y}_t, y_t) &\leq L(\hat{z}_t, y_t) + \frac{1}{\eta_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) + \eta_t \|\mathbf{x}_t\|^2 L(\hat{y}_t, y_t) \\ &= \frac{1}{1 - \eta_t \|\mathbf{x}_t\|^2} \left[L(\hat{z}_t, y_t) + \frac{1}{\eta_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) \right] \\ &= 2L(\hat{z}_t, y_t) + 4\|\mathbf{x}_t\|^2 \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) \quad \text{with } \eta_t = \frac{1}{2\|\mathbf{x}_t\|^2} \end{aligned}$$

Therefore,

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 4X_2^2 U^2$$

$$L(\hat{y}_t, y_t) \leq L(\hat{z}_t, y_t) + 2(\hat{y}_t - y_t)(\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t)$$

Lemma

$$\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t \leq \frac{1}{2r_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) + \frac{r_t}{2} \|\mathbf{x}_t\|^2 \quad \text{where } r_t = \eta_t(\hat{y}_t - y_t)$$

Telescopic Sum

$$\sum_{t=1}^T \|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 = \|\mathbf{u} - \mathbf{w}_0\|^2 - \|\mathbf{u} - \mathbf{w}_T\|^2 \leq \|\mathbf{u} - \mathbf{w}_0\|^2 = \|\mathbf{u}\|^2$$

Combination

$$\begin{aligned} L(\hat{y}_t, y_t) &\leq L(\hat{z}_t, y_t) + \frac{1}{\eta_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) + \eta_t \|\mathbf{x}_t\|^2 L(\hat{y}_t, y_t) \\ &= \frac{1}{1 - \eta_t \|\mathbf{x}_t\|^2} \left[L(\hat{z}_t, y_t) + \frac{1}{\eta_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) \right] \\ &= 2L(\hat{z}_t, y_t) + 4\|\mathbf{x}_t\|^2 \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) \quad \text{with } \eta_t = \frac{1}{2\|\mathbf{x}_t\|^2} \end{aligned}$$

Therefore,

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 4X_2^2 U^2$$

$$L(\hat{y}_t, y_t) \leq L(\hat{z}_t, y_t) + 2(\hat{y}_t - y_t)(\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t)$$

Lemma

$$\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t \leq \frac{1}{2r_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) + \frac{r_t}{2} \|\mathbf{x}_t\|^2 \quad \text{where } r_t = \eta_t(\hat{y}_t - y_t)$$

Telescopic Sum

$$\sum_{t=1}^T \|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 = \|\mathbf{u} - \mathbf{w}_0\|^2 - \|\mathbf{u} - \mathbf{w}_T\|^2 \leq \|\mathbf{u} - \mathbf{w}_0\|^2 = \|\mathbf{u}\|^2$$

Combination

$$\begin{aligned} L(\hat{y}_t, y_t) &\leq L(\hat{z}_t, y_t) + \frac{1}{\eta_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) + \eta_t \|\mathbf{x}_t\|^2 L(\hat{y}_t, y_t) \\ &= \frac{1}{1 - \eta_t \|\mathbf{x}_t\|^2} \left[L(\hat{z}_t, y_t) + \frac{1}{\eta_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) \right] \\ &= 2L(\hat{z}_t, y_t) + 4\|\mathbf{x}_t\|^2 \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) \quad \text{with } \eta_t = \frac{1}{2\|\mathbf{x}_t\|^2} \end{aligned}$$

Therefore,

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 4X_2^2 U^2$$

First-Order Conditions

$$L(\hat{y}_t, y_t) \leq L(\hat{z}_t, y_t) + 2(\hat{y}_t - y_t)(\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t)$$

Lemma

$$\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t \leq \frac{1}{2r_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) + \frac{r_t}{2} \|\mathbf{x}_t\|^2 \quad \text{where } r_t = \eta_t(\hat{y}_t - y_t)$$

Telescopic Sum

$$\sum_{t=1}^T \|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 = \|\mathbf{u} - \mathbf{w}_0\|^2 - \|\mathbf{u} - \mathbf{w}_T\|^2 \leq \|\mathbf{u} - \mathbf{w}_0\|^2 = \|\mathbf{u}\|^2$$

Combination

$$\begin{aligned} L(\hat{y}_t, y_t) &\leq L(\hat{z}_t, y_t) + \frac{1}{\eta_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) + \eta_t \|\mathbf{x}_t\|^2 L(\hat{y}_t, y_t) \\ &= \frac{1}{1 - \eta_t \|\mathbf{x}_t\|^2} \left[L(\hat{z}_t, y_t) + \frac{1}{\eta_t} \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) \right] \\ &= 2L(\hat{z}_t, y_t) + 4\|\mathbf{x}_t\|^2 \left(\|\mathbf{u} - \mathbf{w}_{t-1}\|^2 - \|\mathbf{u} - \mathbf{w}_t\|^2 \right) \quad \text{with } \eta_t = \frac{1}{2\|\mathbf{x}_t\|^2} \end{aligned}$$

Therefore,

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 4X_2^2 U^2$$

Exponentiated-Gradient

Initialization: set $\mathbf{w}_0 = \mathbf{1}$.

Trials: for each $t = 1, 2, \dots$

(1) Observation: receive \mathbf{x}_t .

(2) Prediction: output $\hat{y}_t = \mathbf{w}_{t-1}^\top \mathbf{x}_t$.

(3) Response: get y_t and incur loss $(\hat{y}_t - y_t)^2$.

(4) Update: set $\mathbf{w}_{t,i} = \frac{\mathbf{w}_{t-1,i} e^{-\eta(\hat{y}_t - y_t)\mathbf{x}_{t,i}}}{\sum_{j=1}^n \mathbf{w}_{t-1,j} e^{-\eta(\hat{y}_t - y_t)\mathbf{x}_{t,j}}}$.

Convergence Theorem

For any best regressor $\mathbf{u} \in [0, 1]^n$, the quadratic loss of the exponentiated gradient descent algorithm is bounded by

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 2X_\infty^2 (\ln n - H(\mathbf{u}))$$

where $X_\infty = \max_{t=1}^T \max_{i=1}^n \mathbf{x}_{t,i}$, and choosing $\eta = \frac{2}{X_\infty^2}$

First-Order Conditions

$$L(\hat{y}_t, y_t) \leq L(\hat{z}_t, y_t) + 2(\hat{y}_t - y_t)(\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t)$$

Lemma

Let $r_t = \eta(\hat{y}_t - y_t)$ and $d_{\text{re}}(\mathbf{u}, \mathbf{w}) = \sum_{i=1}^n u_i \ln \frac{u_i}{w_i}$. Then

$$\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t \leq \frac{1}{r_t} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) + \frac{r_t X_\infty^2}{8}$$

Telescopic Sum

$$\sum_{t=1}^T d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) = d_{\text{re}}(\mathbf{u}, \mathbf{w}_0) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_T) \leq d_{\text{re}}(\mathbf{u}, \mathbf{w}_0) = \ln n - H(\mathbf{u})$$

Combination

$$\begin{aligned} L(\hat{y}_t, y_t) &\leq L(\hat{z}_t, y_t) + \frac{1}{\eta} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) + \frac{\eta X_\infty^2}{4} L(\hat{y}_t, y_t) \\ &= \frac{4}{4 - \eta X_\infty^2} \left[L(\hat{z}_t, y_t) + \frac{2}{\eta} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) \right] \\ &= 2L(\hat{z}_t, y_t) + 2X_\infty^2 \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) \quad \text{with } \eta = \frac{2}{X_\infty^2} \end{aligned}$$

Therefore,

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 2X_\infty^2 (\ln n - H(\mathbf{u}))$$

$$L(\hat{y}_t, y_t) \leq L(\hat{z}_t, y_t) + 2(\hat{y}_t - y_t)(\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t)$$

Lemma

Let $r_t = \eta(\hat{y}_t - y_t)$ and $d_{\text{re}}(\mathbf{u}, \mathbf{w}) = \sum_{i=1}^n u_i \ln \frac{u_i}{w_i}$. Then

$$\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t \leq \frac{1}{r_t} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) + \frac{r_t X_\infty^2}{8}$$

Telescopic Sum

$$\sum_{t=1}^T d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) = d_{\text{re}}(\mathbf{u}, \mathbf{w}_0) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_T) \leq d_{\text{re}}(\mathbf{u}, \mathbf{w}_0) = \ln n - H(\mathbf{u})$$

Combination

$$\begin{aligned} L(\hat{y}_t, y_t) &\leq L(\hat{z}_t, y_t) + \frac{1}{\eta} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) + \frac{\eta X_\infty^2}{4} L(\hat{y}_t, y_t) \\ &= \frac{4}{4 - \eta X_\infty^2} \left[L(\hat{z}_t, y_t) + \frac{2}{\eta} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) \right] \\ &= 2L(\hat{z}_t, y_t) + 2X_\infty^2 \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) \quad \text{with } \eta = \frac{2}{X_\infty^2} \end{aligned}$$

Therefore,

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 2X_\infty^2 (\ln n - H(\mathbf{u}))$$

First-Order Conditions

$$L(\hat{y}_t, y_t) \leq L(\hat{z}_t, y_t) + 2(\hat{y}_t - y_t)(\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t)$$

Lemma

Let $r_t = \eta(\hat{y}_t - y_t)$ and $d_{\text{re}}(\mathbf{u}, \mathbf{w}) = \sum_{i=1}^n u_i \ln \frac{u_i}{w_i}$. Then

$$\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t \leq \frac{1}{r_t} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) + \frac{r_t X_\infty^2}{8}$$

Telescopic Sum

$$\sum_{t=1}^T d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) = d_{\text{re}}(\mathbf{u}, \mathbf{w}_0) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_T) \leq d_{\text{re}}(\mathbf{u}, \mathbf{w}_0) = \ln n - H(\mathbf{u})$$

Combination

$$\begin{aligned} L(\hat{y}_t, y_t) &\leq L(\hat{z}_t, y_t) + \frac{1}{\eta} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) + \frac{\eta X_\infty^2}{4} L(\hat{y}_t, y_t) \\ &= \frac{4}{4 - \eta X_\infty^2} \left[L(\hat{z}_t, y_t) + \frac{2}{\eta} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) \right] \\ &= 2L(\hat{z}_t, y_t) + 2X_\infty^2 \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) \quad \text{with } \eta = \frac{2}{X_\infty^2} \end{aligned}$$

Therefore,

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 2X_\infty^2 (\ln n - H(\mathbf{u}))$$

First-Order Conditions

$$L(\hat{y}_t, y_t) \leq L(\hat{z}_t, y_t) + 2(\hat{y}_t - y_t)(\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t)$$

Lemma

Let $r_t = \eta(\hat{y}_t - y_t)$ and $d_{\text{re}}(\mathbf{u}, \mathbf{w}) = \sum_{i=1}^n u_i \ln \frac{u_i}{w_i}$. Then

$$\mathbf{w}_t^\top \mathbf{x}_t - \mathbf{u}^\top \mathbf{x}_t \leq \frac{1}{r_t} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) + \frac{r_t X_\infty^2}{8}$$

Telescopic Sum

$$\sum_{t=1}^T d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) = d_{\text{re}}(\mathbf{u}, \mathbf{w}_0) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_T) \leq d_{\text{re}}(\mathbf{u}, \mathbf{w}_0) = \ln n - H(\mathbf{u})$$

Combination

$$\begin{aligned} L(\hat{y}_t, y_t) &\leq L(\hat{z}_t, y_t) + \frac{1}{\eta} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) + \frac{\eta X_\infty^2}{4} L(\hat{y}_t, y_t) \\ &= \frac{4}{4 - \eta X_\infty^2} \left[L(\hat{z}_t, y_t) + \frac{2}{\eta} \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) \right] \\ &= 2L(\hat{z}_t, y_t) + 2X_\infty^2 \left(d_{\text{re}}(\mathbf{u}, \mathbf{w}_{t-1}) - d_{\text{re}}(\mathbf{u}, \mathbf{w}_t) \right) \quad \text{with } \eta = \frac{2}{X_\infty^2} \end{aligned}$$

Therefore,

$$\mathbb{L}(\mathbf{w}) \leq 2\mathbb{L}(\mathbf{u}) + 2X_\infty^2 (\ln n - H(\mathbf{u}))$$

Outline

1 Online Learning

2 Convex Analysis

- Convex Sets
- Convex Functions
- Loss Functions

3 Regression Learning

- The Protocol
- The Gradient-Descent Algorithm
- The Exponentiated-Gradient Algorithm

4 Classification Learning

- The Protocol
- The Perceptron Algorithm
- The Winnow Algorithm

Classification Learning

Parameters: update function f , threshold θ

Initialization: set \mathbf{w}_0 .

Trials: for each $t = 1, 2, \dots$

- (1) Observation: receive \mathbf{x}_t
- (2) Prediction: output $\hat{y}_t = \text{sign}(\mathbf{w}_{t-1}^\top \mathbf{x}_t - \theta)$
- (3) Response: get y_t and incur loss $L(\hat{y}_t, y_t)$
- (4) Update: set $\mathbf{w}_t = f(\mathbf{w}_{t-1})$

Best classifier

The hyperplane \mathbf{u} that minimizes

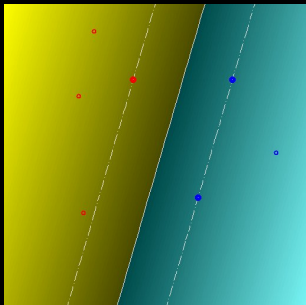
$$\mathbb{L}_{0,1}(\mathbf{u}) = \sum_{t=1}^T L_{0,1}(\hat{z}_t, y_t)$$

where $\hat{z}_t = \text{sign}(\mathbf{u}^\top \mathbf{x}_t - \theta)$

Goal

The goal of the learner is to minimize

$$\mathbb{L}_{0,1}(\mathbf{w}) - \mathbb{L}(\mathbf{u}) \quad \text{where} \quad \mathbb{L}_{0,1}(\mathbf{w}) = \sum_{t=1}^T L_{0,1}(\hat{y}_t, y_t)$$



Classification Learning

Parameters: update function f , threshold θ

Initialization: set \mathbf{w}_0 .

Trials: for each $t = 1, 2, \dots$

- (1) Observation: receive \mathbf{x}_t
- (2) Prediction: output $\hat{y}_t = \text{sign}(\mathbf{w}_{t-1}^\top \mathbf{x}_t - \theta)$
- (3) Response: get y_t and incur loss $L(\hat{y}_t, y_t)$
- (4) Update: set $\mathbf{w}_t = f(\mathbf{w}_{t-1})$

Best classifier

The hyperplane \mathbf{u} that minimizes

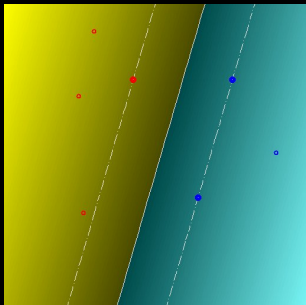
$$\mathbb{L}_{0,1}(\mathbf{u}) = \sum_{t=1}^T L_{0,1}(\hat{z}_t, y_t)$$

where $\hat{z}_t = \text{sign}(\mathbf{u}^\top \mathbf{x}_t - \theta)$

Goal

The goal of the learner is to minimize

$$\mathbb{L}_{0,1}(\mathbf{w}) - \mathbb{L}(\mathbf{u}) \quad \text{where} \quad \mathbb{L}_{0,1}(\mathbf{w}) = \sum_{t=1}^T L_{0,1}(\hat{y}_t, y_t)$$



Classification Learning

Parameters: update function f , threshold θ

Initialization: set \mathbf{w}_0 .

Trials: for each $t = 1, 2, \dots$

- (1) Observation: receive \mathbf{x}_t
- (2) Prediction: output $\hat{y}_t = \text{sign}(\mathbf{w}_{t-1}^\top \mathbf{x}_t - \theta)$
- (3) Response: get y_t and incur loss $L(\hat{y}_t, y_t)$
- (4) Update: set $\mathbf{w}_t = f(\mathbf{w}_{t-1})$

Best classifier

The hyperplane \mathbf{u} that minimizes

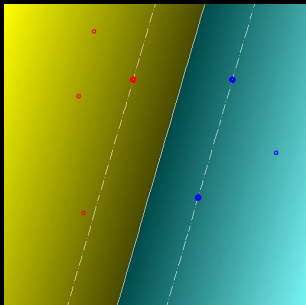
$$\mathbb{L}_{0,1}(\mathbf{u}) = \sum_{t=1}^T L_{0,1}(\hat{z}_t, y_t)$$

where $\hat{z}_t = \text{sign}(\mathbf{u}^\top \mathbf{x}_t - \theta)$

Goal

The goal of the learner is to minimize

$$\mathbb{L}_{0,1}(\mathbf{w}) - \mathbb{L}(\mathbf{u}) \quad \text{where} \quad \mathbb{L}_{0,1}(\mathbf{w}) = \sum_{t=1}^T L_{0,1}(\hat{y}_t, y_t)$$



Perceptron

Initialization: set $\mathbf{w}_0 = \mathbf{0}$.

Trials: for each $t = 1, 2, \dots$

- (1) Observation: receive \mathbf{x}_t
- (2) Prediction: output $\hat{y}_t = \text{sign}(\mathbf{w}_{t-1}^\top \mathbf{x}_t)$
- (3) Response: get y_t and incur one mistake if $y_t \neq \hat{y}_t$
- (4) Update: set $\mathbf{w}_t = \mathbf{w}_{t-1} - \frac{\eta_t}{2}(\hat{y}_t - y_t)\mathbf{x}_t$

Convergence Theorem

For any best regressor $\mathbf{u} \in \mathbb{R}^n$, the quadratic loss of the gradient descent algorithm is bounded by

$$\mathbb{L}(\mathbf{w}) \leq \mathbb{L}(\mathbf{u}) + \frac{1}{4} X_2^2 U^2$$

where $X_2 = \max_{t=1}^T \|\mathbf{x}_t\|$, $U = \|\mathbf{u}\|$, and choosing $\eta_t = \frac{2}{\|\mathbf{x}_t\|^2}$

Sketch of Proof

A variant of the proof for additive regression learning that uses absolute loss bounds with a tolerance factor (Bylander).

Normalized Winnow

Initialization: set $\mathbf{w}_0 = \mathbf{1}$.

Trials: for each $t = 1, 2, \dots$

- (1) Observation: receive \mathbf{x}_t
- (2) Prediction: output $\hat{y}_t = \text{sign}(\mathbf{w}_{t-1}^\top \mathbf{x}_t - \frac{1}{2})$
- (3) Response: get y_t and incur one mistake if $y_t \neq \hat{y}_t$
- (4) Update: set $\mathbf{w}_{t,i} = \frac{\mathbf{w}_{t-1,i} e^{-\frac{\eta}{2}(\hat{y}_t - y_t)\mathbf{x}_{t,i}}}{\sum_{j=1}^n \mathbf{w}_{t-1,j} e^{-\frac{\eta}{2}(\hat{y}_t - y_t)\mathbf{x}_{t,j}}}$

Convergence Theorem

For any best regressor $\mathbf{u} \in [0, 1]^n$ with margin $\gamma \in (0, \frac{1}{2})$, the number of mistakes made by Winnow is bounded by

$$\mathbb{L}(\mathbf{w}) \leq \frac{1}{\gamma} \mathbb{L}(\mathbf{u}) + \frac{X_\infty^2}{2\gamma^2} (\ln n - H(\mathbf{u}))$$

where $X_\infty = \max_{t=1}^T \max_{i=1}^n \mathbf{x}_{t,i}$, and choosing $\eta = \frac{8\gamma}{X_\infty^2}$

Sketch of Proof

Again, a variant of the proof for multiplicative regression learning that uses absolute loss bounds with a tolerance factor $\frac{\gamma}{2}$ (inspired from Bylander).