Lecture 12: Linear Classes, Perceptron, and Smoothed Analysis (draft)

Han Shao

UML, Sections 9.1.2 and 21.4
Haghtalab, Roughgarden, and Shetty (2020)

Without additional assumptions, linear separators can have infinite Littlestone dimension. Thus, even realizability does not give a finite worst-case mistake bound. In this lecture, we consider two assumptions that make linear classes easier to learn: separation with a positive margin, and smoothness of the distributions generating the instances.
We focus on the homogeneous halfspace class

$$\mathcal{H}_p^{\mathrm{hom}} = \left\{ h_w(x) = \begin{cases} +1 & \text{if } w^\intercal x \geq 0, \\ -1 & \text{if } w^\intercal x < 0, \end{cases} : w \in \mathbb{R}^p \right\}.$$

On the full domain $\mathbb{R}^p$, this class has infinite Littlestone dimension when $p \geq 2$. An affine separator in $\mathbb{R}^p$ can be represented as a homogeneous separator in $\mathbb{R}^{p+1}$ by lifting x to $(x, 1)$, since

$$w^\intercal x + b = (w, b)^\intercal (x, 1).$$

1 Separability with a Large Margin

Consider a sequence

$$S = ((x_1, y_1), \ldots, (x_T, y_T)), \qquad \|x_t\|_2 \leq R \quad \text{for every } t \in [T],$$

where $R > 0$. Suppose there exists $w^\star \in \mathbb{R}^p$ such that

$$y_t (w^\star)^\intercal x_t \geq 1 \qquad \text{for every } t \in [T]. \tag{1}$$

Equivalently, $(w^\star)^\intercal x_t \geq 1$ on positive examples and $(w^\star)^\intercal x_t \leq -1$ on negative examples.
The Euclidean distance from x to the hyperplane $(w^\star)^\intercal x = 0$ is

$$\frac{|(w^\star)^\intercal x|}{\|w^\star\|_2}.$$

Thus (1) guarantees a geometric margin of at least

$$\gamma = \frac{1}{\|w^\star\|_2}.$$

Conversely, if a unit vector $u^\star$ satisfies $y_t (u^\star)^\intercal x_t \geq \gamma > 0$ for every t, then $w^\star = u^\star / \gamma$ satisfies (1).

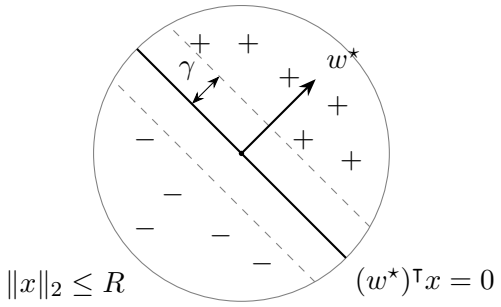


Figure 1: A separating hyperplane and its margin. The dashed lines are $(w^\star)^\intercal x = \pm 1$ under the normalization in (1).

The Perceptron Algorithm Perceptron maintains a weight vector and changes it only when it makes a mistake. A mistaken positive example is added to the vector; a mistaken negative example is subtracted.

Algorithm 1 Perceptron

Initialize $w_1 = 0 \in \mathbb{R}^p$.
for $t = 1, 2, \ldots, T$ **do**
  Receive x_t.
  Predict $\widehat{y}_t = h_{w_t}(x_t)$.
  Observe y_t.
  if $\widehat{y}_t \neq y_t$ **then**
    $w_{t+1} = w_t + y_t x_t$.
  else
    $w_{t+1} = w_t$.
  end if
end for

Theorem 1 (Perceptron mistake bound)**.** *Suppose $\|x_t\|_2 \le R$ and $y_t(w^\star)^\intercal x_t \ge 1$ for every $t \in [T]$. Then Perceptron makes at most*

$$M \le R^2 \|w^\star\|_2^2$$

mistakes. In geometric-margin notation, this is

$$M \le \frac{R^2}{\gamma^2}.$$

Proof. We compare the progress of the weight vector in the direction of $w^\star$ with the growth of its norm.

Progress toward the separator. On every mistake round,

$$\begin{aligned} w_{t+1}^\intercal w^\star &= (w_t + y_t x_t)^\intercal w^\star \\ &= w_t^\intercal w^\star + y_t (w^\star)^\intercal x_t \\ &\ge w_t^\intercal w^\star + 1. \end{aligned}$$

The vector does not change on other rounds. Since $w_1 = 0$, after M mistakes,

$$w_{T+1}^\intercal w^\star \ge M. \tag{2}$$

Controlling the norm. Whenever Perceptron makes a mistake, $y_t w_t^\intercal x_t \le 0$. Consequently,

$$\begin{aligned}\|w_{t+1}\|_2^2 &= \|w_t + y_t x_t\|_2^2 \\ &= \|w_t\|_2^2 + 2y_t w_t^\intercal x_t + \|x_t\|_2^2 \\ &\le \|w_t\|_2^2 + R^2.\end{aligned}$$

After M mistakes, therefore,

$$\|w_{T+1}\|_2^2 \le MR^2, \qquad \|w_{T+1}\|_2 \le R\sqrt{M}.$$

By Cauchy–Schwarz and (2),

$$M \le w_{T+1}^\intercal w^\star \le \|w_{T+1}\|_2 \|w^\star\|_2 \le R\sqrt{M}\,\|w^\star\|_2.$$

If $M = 0$, the claim is immediate. Otherwise, divide by $\sqrt{M}$ and square to obtain $M \le R^2\|w^\star\|_2^2$. □

As proved in the homework, we can convert an online learning algorithm and regret bound to an offline learning algorithm, so to learn a linear separator with γ margin, we need $O(\frac{R^2}{\gamma^2 \varepsilon})$ samples. The dependence on R/γ cannot be improved in general. The following construction uses one previously unseen coordinate per round.

Proposition 1 (Margin lower bound for deterministic learners)**.** *Let $0 < \gamma \le R$ and $n = \lfloor R^2/\gamma^2 \rfloor$. In ambient dimension $p \ge n$, every deterministic learner can be forced to make n mistakes on a sequence with $\|x_t\|_2 \le R$ that is separable with geometric margin at least γ.*

Proof. Let $e_1, \dots, e_n$ be the first n standard basis vectors in $\mathbb{R}^p$. On round $t \in [n]$, present

$$x_t = Re_t$$

and choose $y_t = -\widehat{y}_t$. The learner makes a mistake on every round. Since the learner is deterministic, these labels can be constructed in advance by simulating it.
After constructing the sequence, define

$$w' = \sum_{t=1}^{n} y_t e_t, \qquad w^\star = \frac{w'}{R}.$$

For every $t \in [n]$,

$$y_t (w^\star)^\intercal x_t = y_t \left(\frac{1}{R} \sum_{j=1}^{n} y_j e_j \right)^\intercal Re_t = y_t^2 = 1.$$

Moreover,

$$\|w^\star\|_2 = \frac{\sqrt{n}}{R} \le \frac{1}{\gamma}.$$

Thus the geometric margin is $R/\sqrt{n} \ge \gamma$, as required. When R^2/γ^2 is an integer, this construction forces exactly R^2/γ^2 mistakes. □

Without a perfect separator, we measure how much a comparator violates the unit-margin condition.

Definition 1 (Hinge loss)**.** *The hinge loss of a vector w on a labeled example (x, y) is*

$$\ell_{\text{hinge}}(w, (x, y)) = \max\{0, 1 - yw^\intercal x\}.$$

Its total hinge loss on $S = ((x_t, y_t))_{t=1}^T$ is

$$L_{\text{hinge}}(w, S) = \sum_{t=1}^{T} \ell_{\text{hinge}}(w, (x_t, y_t)).$$

For a positive example, this loss is $\max\{0, 1 - w^\intercal x\}$, which is zero once $w^\intercal x \geq 1$.

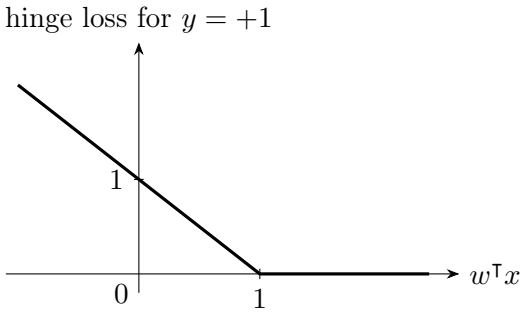


Figure 2: The positive-label hinge loss $\max\{0, 1 - w^\intercal x\}$.

Theorem 2 (Perceptron with hinge-loss comparison; stated without proof)**.** *On any sequence $S = ((x_t, y_t))_{t=1}^T$ with $\|x_t\|_2 \leq R$, Perceptron makes at most*

$$M \leq \min_{w \in \mathbb{R}^p} \left\{ R^2 \|w\|_2^2 + 2L_{\text{hinge}}(w, S) \right\}$$

mistakes.

The comparator need not separate the sequence. The bound trades off its norm against its total hinge loss. When a comparator satisfies the unit-margin condition, its hinge loss is zero and the bound reduces to theorem 1.

2 Smooth Distributions

The second approach restricts how instances are generated, rather than assuming a margin. The motivating i.i.d. benchmark in these notes is a regret guarantee of order $\text{VCDim}(\mathcal{H})\sqrt{T}$. We now relax the i.i.d. assumption by allowing different distributions on different rounds, subject to a smoothness condition.

Smoothness and the Online Protocol Let U be a fixed uniform probability distribution on $\mathcal{X}$. For example, $\mathcal{X}$ may be a finite set, or a bounded subset of $\mathbb{R}^p$ of positive volume. In particular, we do not use a uniform probability distribution on all of $\mathbb{R}^p$.

Definition 2 (σ-smooth distribution)**.** *Let $0 < \sigma \leq 1$. A distribution Q on $\mathcal{X}$ is* σ-smooth *relative to U if its density q satisfies*

$$q(x) \leq \frac{u(x)}{\sigma}$$

almost everywhere, where u is the density of U.
A joint distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$ is called σ-smooth when its instance marginal $\mathcal{D}_{\mathcal{X}}$ is σ-smooth.

The condition prevents the adversary from concentrating too much probability on a small region. It imposes no restriction on the conditional label distribution given the instance.
In smoothed online learning, on each round t the adversary selects a σ-smooth joint distribution $\mathcal{D}_t$, and (x_t, y_t) is drawn from it. The learner receives x_t, predicts $\widehat{y}_t$, and then observes y_t. An adaptive adversary may choose $\mathcal{D}_t$ based on previous examples and previous learner actions, but not on the learner's current random draw.

Non-adaptive adversaries. The proof in this lecture concerns the non-adaptive case: the adversary fixes an unknown sequence

$$\boldsymbol{\mathcal{D}} = (\mathcal{D}_1, \ldots, \mathcal{D}_T)$$

before the interaction, and the pairs $(x_t, y_t) \sim \mathcal{D}_t$ are drawn independently across rounds. The distributions may differ with t. We write

$$S \sim \bigotimes_{t=1}^{T} \mathcal{D}_t.$$

For a realized sequence, define

$$L_T(h) = \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq y_t\}, \qquad L_T(\mathcal{A}) = \sum_{t=1}^{T} \mathbb{1}\{\widehat{y}_t \neq y_t\}.$$

The expected cumulative regret is

$$\mathrm{Reg}_T(\mathcal{A}, \boldsymbol{\mathcal{D}}) = \mathbb{E}_{S,\mathcal{A}} \left[L_T(\mathcal{A}) - \min_{h \in \mathcal{H}} L_T(h) \right]. \tag{3}$$

The expectation includes both data generation and the learner's internal randomness. The best-in-hindsight comparator remains inside the expectation. Average regret would be $\mathrm{Reg}_T(\mathcal{A}, \boldsymbol{\mathcal{D}})/T$.

Theorem 3 (Regret against a non-adaptive smooth adversary)**.** *Let $\mathcal{H}$ be a nonempty binary hypothesis class with finite VC dimension $d \geq 1$, and suppose $T \geq \max\{d, 2\}$. There is a randomized online learning algorithm such that, for every non-adaptive sequence of σ-smooth distributions,*

$$\mathrm{Reg}_T(\mathcal{A}, \boldsymbol{\mathcal{D}}) \leq O\left(\sqrt{dT \ln \frac{eT}{\sigma}} \right).$$

Approximating a Hypothesis Class by a Finite Cover

Definition 3 (ε-cover)**.** *Given a distribution Q on $\mathcal{X}$, a set $\mathcal{H}_\varepsilon \subseteq \mathcal{H}$ is an* ε-cover *of $\mathcal{H}$ under Q if, for every $h \in \mathcal{H}$, there exists $h' \in \mathcal{H}_\varepsilon$ such that*

$$\Pr_{x \sim Q}[h(x) \neq h'(x)] \leq \varepsilon.$$

Lemma 1 (VC covering bound; stated without proof)**.** *If $\mathrm{VCDim}(\mathcal{H}) = d \geq 1$, then, for every distribution Q and $\varepsilon \in (0, 1)$, there is an ε-cover $\mathcal{H}_\varepsilon \subseteq \mathcal{H}$ under Q with*

$$|\mathcal{H}_\varepsilon| \leq \left(\frac{41}{\varepsilon} \right)^d.$$

Example 1 (One-dimensional thresholds)**.** *For a distribution on $[0, 1]$, place thresholds at quantiles so that each interval between successive thresholds has probability mass at most ε. Any threshold can be moved to a nearby chosen threshold, changing its prediction only inside one such interval. Thus $O(1/\varepsilon)$ thresholds give an ε-cover. This is the "one threshold per ε-mass" construction.*

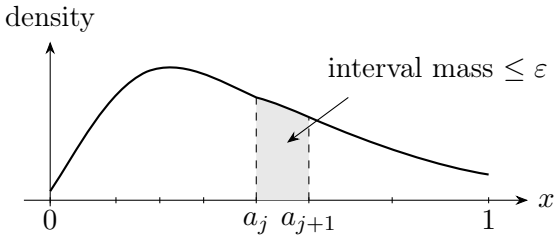


Figure 3: A threshold cover uses probability mass, rather than equal Euclidean spacing, to choose the representatives. The curve is schematic.

Proof of the Smoothed Regret Bound

Proof of theorem 3. Fix $\varepsilon \in (0,1)$, to be chosen at the end. Let $\mathcal{H}_\varepsilon \subseteq \mathcal{H}$ be an ε-cover under U with

$$N := |\mathcal{H}_\varepsilon| \leq \left(\frac{41}{\varepsilon}\right)^d.$$

For each $h \in \mathcal{H}$, fix a representative $\pi(h) \in \mathcal{H}_\varepsilon$ satisfying

$$\Pr_{x\sim U}[h(x) \neq \pi(h)(x)] \leq \varepsilon.$$

The cover and this choice of representatives are fixed before the examples are generated.

Run Hedge on the cover. Treat each $g \in \mathcal{H}_\varepsilon$ as an expert, with round-t loss

$$\ell_t(g) = \mathbb{1}\{g(x_t) \neq y_t\}.$$

Sample an expert h_t according to Hedge's distribution p_t and predict $\widehat{y}_t = h_t(x_t)$. After observing y_t, all expert losses are known. Conditional on any fixed labeled sequence, the Hedge bound gives

$$\mathbb{E}_{\mathcal{A}}[L_T(\mathcal{A})] - \min_{g\in\mathcal{H}_\varepsilon} L_T(g) \leq \sqrt{\frac{T\ln N}{2}}.$$

For $N > 1$, use learning rate $\eta = \sqrt{8\ln N/T}$; for $N = 1$, simply follow the sole expert.

Separate estimation from approximation. Adding and subtracting the best loss within the cover gives

$$\begin{aligned}\mathrm{Reg}_T(\mathcal{A}, \boldsymbol{\mathcal{D}}) = \mathbb{E}_{S,\mathcal{A}}\left[L_T(\mathcal{A}) - \min_{g\in\mathcal{H}_\varepsilon} L_T(g)\right] \\ + \mathbb{E}_S\left[\min_{g\in\mathcal{H}_\varepsilon} L_T(g) - \min_{h\in\mathcal{H}} L_T(h)\right].\end{aligned}$$

For every realized sequence,

$$\begin{aligned}&\min_{g\in\mathcal{H}_\varepsilon} L_T(g) - \min_{h\in\mathcal{H}} L_T(h) \\ &\quad = \max_{h\in\mathcal{H}} \min_{g\in\mathcal{H}_\varepsilon} \big(L_T(g) - L_T(h)\big)\end{aligned}$$

$$\leq \max_{h\in\mathcal{H}} \min_{g\in\mathcal{H}_\varepsilon} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq g(x_t)\}$$
$$\leq \sup_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq \pi(h)(x_t)\}.$$

The inequality uses

$$\mathbb{1}\{g(x_t) \neq y_t\} - \mathbb{1}\{h(x_t) \neq y_t\} \leq \mathbb{1}\{g(x_t) \neq h(x_t)\}.$$

It follows that

$$\mathrm{Reg}_T(\mathcal{A}, \boldsymbol{\mathcal{D}}) \leq \sqrt{\frac{T \ln N}{2}} + A_T, \qquad A_T := \mathbb{E}_S \left[\sup_{h\in\mathcal{H}} \sum_{t=1}^{T} \mathbb{1}\{h(x_t) \neq \pi(h)(x_t)\} \right]. \tag{4}$$

Smoothness controls the population disagreement. For each $h \in \mathcal{H}$, define

$$g_h(x) = \mathbb{1}\{h(x) \neq \pi(h)(x)\}, \qquad \mathcal{G} = \{g_h : h \in \mathcal{H}\}.$$

By smoothness of the instance marginal, for every h and t,

$$\mathbb{E}_{x\sim\mathcal{D}_{t,\mathcal{X}}}[g_h(x)] \leq \frac{1}{\sigma}\mathbb{E}_{x\sim U}[g_h(x)] \leq \frac{\varepsilon}{\sigma}. \tag{5}$$

However, the comparator in A_T can be chosen after seeing the sample. We therefore need a bound uniform over h, not just the expectation for each fixed h.

Uniform convergence for independent, non-identical samples. We will use the following form of the uniform-convergence lemma. As before, $\Pi_{\mathcal{G}}(n)$ denotes the maximum number of distinct labelings of n instances induced by $\mathcal{G} \subseteq \{0,1\}^{\mathcal{X}}$.

Lemma 2 (Independent-sample uniform convergence; stated without proof)**.** *Let $X_1, \ldots, X_T$ be independent, with $X_t \sim Q_t$, where the Q_t need not be identical. For every nonempty binary function class $\mathcal{G}$, there is a universal constant $C > 0$ such that*

$$\mathbb{E}_{X_{1:T}} \left[\sup_{g\in\mathcal{G}} \left(\frac{1}{T}\sum_{t=1}^{T} g(X_t) - \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{X_t\sim Q_t}[g(X_t)] \right) \right] \leq C\sqrt{\frac{\ln\big(2\Pi_{\mathcal{G}}(2T)\big)}{T}}.$$

In particular, if $1 \leq v = \mathrm{VCDim}(\mathcal{G}) \leq T$, the right-hand side is

$$O\left(\sqrt{\frac{v\ln(2eT/v)}{T}}\right).$$

This is the same double-sampling and symmetrization argument as before: for each t, use an independent copy $X_t' \sim Q_t$. The pair (X_t, X_t') can be swapped without changing its distribution. Independence across rounds is enough; identical distributions are not needed. As usual, we assume the relevant functions and suprema are measurable.

Apply uniform convergence to the disagreement class. On any n instances, the labeling induced by g_h is determined by the two labelings induced by h and $\pi(h)$, both of which belong to $\mathcal{H}$. Consequently,

$$\Pi_{\mathcal{G}}(n) \le \Pi_{\mathcal{H}}(n)^2. \tag{6}$$

This checks the complexity of the class to which the lemma is applied; we do not assume that $\mathrm{VCDim}(\mathcal{G}) = d$. Since $T \ge d$, Sauer's lemma yields

$$\Pi_{\mathcal{G}}(2T) \le \left(\frac{2eT}{d}\right)^{2d}.$$

The instances are independent, so lemma 2 implies

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{H}}\left(\sum_{t=1}^{T} g_h(x_t) - \sum_{t=1}^{T}\mathbb{E}_{x\sim\mathcal{D}_{t,\mathcal{X}}}[g_h(x)]\right)\right] \le O\left(\sqrt{dT\ln\frac{2eT}{d}}\right).$$

Together with (5), this gives

$$A_T \le \frac{\varepsilon T}{\sigma} + O\left(\sqrt{dT\ln\frac{2eT}{d}}\right). \tag{7}$$

Here the normalized deviation in lemma 2 is multiplied by T, because A_T is a cumulative quantity.

Choose the cover resolution. Substituting the cover size and (7) into (4) gives

$$\mathrm{Reg}_T(\mathcal{A}, \boldsymbol{\mathcal{D}}) \le \sqrt{\frac{dT}{2}\ln\frac{41}{\varepsilon}} + \frac{\varepsilon T}{\sigma} + O\left(\sqrt{dT\ln\frac{2eT}{d}}\right).$$

Take the resolution from the handwriting,

$$\varepsilon = \frac{\sigma}{\sqrt{T}}.$$

Since $T \ge 2$ and $0 < \sigma \le 1$, we have $0 < \varepsilon < 1$. The bound becomes

$$\sqrt{\frac{dT}{2}\ln\frac{41\sqrt{T}}{\sigma}} + \sqrt{T} + O\left(\sqrt{dT\ln\frac{2eT}{d}}\right) = O\left(\sqrt{dT\ln\frac{eT}{\sigma}}\right).$$

This proves the theorem. □