Lecture 17 SVM (Part II) and Online Learning

Lei Li, **Yu-Xiang Wang**

(some slides from my convex optimization class, originally taught by Ryan Tibshirani in CMU)

Recap: Support Vector Machines

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$ having rows $x_1, \ldots x_n$, recall the support vector machine or SVM problem:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \ i = 1, \ldots n \\ & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots n \end{aligned}$$

This is a quadratic program

Recap: Lagrange dual problem

Given a minimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0,\ i = 1, \dots m \\ & \ell_j(x) = 0,\ j = 1, \dots r \end{aligned}$$

we defined the Lagrangian:

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$$

and Lagrange dual function:

$$g(u, v) = \min_x \ L(x, u, v)$$

Recap: Lagrange dual problem

The subsequent dual problem is:

$$\max_{u,v} \quad g(u,v)$$
$$\text{subject to} \quad u \geq 0$$

Important properties:

- Dual problem is always convex, i.e., g is always concave (even if primal problem is not convex)
- The primal and dual optimal values, $f^\star$ and $g^\star$, always satisfy weak duality: $f^\star \geq g^\star$
- Slater's condition: for convex primal, if there is an x such that

$$h_1(x) < 0, \ldots h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots \ell_r(x) = 0$$

then strong duality holds: $f^\star = g^\star$. Can be further refined to strict inequalities over the nonaffine h_i, $i = 1, \ldots m$

Recap: Deriving the dual of SVM

Introducing dual variables $\underbrace{v, w \geq 0}$, we form the Lagrangian:

$$L(\underbrace{\beta, \overset{x}{\beta_0}, \xi}, v, w) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n v_i \xi_i + \sum_{i=1}^n w_i\big(1 - \xi_i - y_i(x_i^T\beta + \beta_0)\big)$$

$$0 = \nabla_\beta L = \beta + \sum_{i=1}^n \underbrace{w_i(-y_i)}_{\text{scalar}} \vec{x}_i$$

$$\nabla_{\beta_0} L = \sum_{i=1}^n -w_i y_i \qquad \underbrace{\beta_0\Big(\sum_{i=1}^n -w_i y_i\Big)}$$

$$\nabla_{\xi_i} L = C - v_i - w_i \qquad \xi_i\underbrace{(C - v_i - w_i)}$$

$v, w \geq 0$

$$\left\{\begin{array}{l} -\sum_{i=1}^n w_i y_i = 0 \\ C - v_i - w_i = 0 \end{array}\right.$$

constraints on w, v

$$\underline{g(w,v)} = \begin{cases} -\infty & \text{if not satisfied} \\ - & \text{next slide} \end{cases}$$

Recap: Dual SVM

Minimizing over β, β_0, ξ gives Lagrange dual function:

$$g(v, w) = \begin{cases} -\frac{1}{2} w^T \tilde{X} \tilde{X}^T w + 1^T w & \text{if } w = C1 - v, \ w^T y = 0 \\ -\infty & \text{otherwise} \end{cases}$$

where $\tilde{X} = \mathrm{diag}(y) X$. Thus SVM dual problem, eliminating slack variable v, becomes

$$\begin{aligned} \max_{w} \quad & -\frac{1}{2} w^T \tilde{X} \tilde{X}^T w + 1^T w \\ \text{subject to} \quad & 0 \leq w \leq C1, \ w^T y = 0 \end{aligned}$$

Check: Slater's condition is satisfied, and we have strong duality. Further, from study of SVMs, might recall that at optimality

$$\beta = \tilde{X}^T w$$

$\text{Score}(x) = \beta^T x + \beta_0$

This is not a coincidence, as we'll later via the KKT conditions

“Kernel trick” in SVM

$(y_1, y_2, \ldots, y_n)\; n \,|\, \boxed{X}\; d$

$\tilde{X} = \mathrm{diag}(y)\, X$

- The dual SVM depends only on inner products

$\forall x \quad k(x,x) \geq 0$

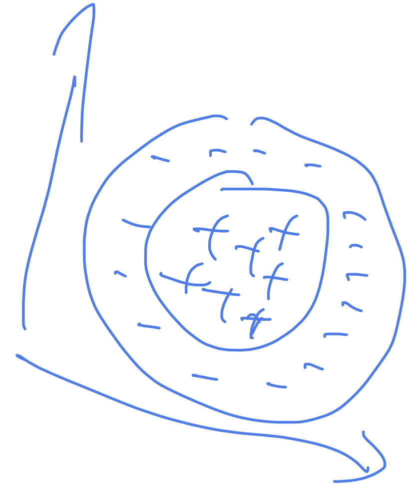
$A \succeq 0$

$x^T A x \geq 0 \;\; \forall x$

$$\max_{w} \quad -\frac{1}{2} w^T \boxed{\tilde{X}\tilde{X}^T} w + 1^T w$$

$$\text{subject to} \quad 0 \leq w \leq C1, \;\; w^T y = 0$$

$$w^T \tilde{X} \tilde{X}^T w = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j y_i y_j x_i^T x_j$$



$\phi(x_i) \in \mathcal{H}$

example: $\phi(x_i) = \exp(-\|x_i - \cdot\|^2)$

$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle = \exp(-\|x_i - x_j\|^2)$

- How to make predictions?

$$\mathrm{sign}\big(x^T \cdot \underbrace{\tilde{X}^T w}_{\beta} + \beta_0\big) = \mathrm{sign}\big(\phi(x)^T \cdot [\phi(x_1) \cdot y_1, \phi(x_2) y_2, \ldots, \phi(x_n) y_n] \cdot w + \beta_0\big)$$

$$= \mathrm{sign}\big([k(x,x_1) \cdot y_1, k(x,x_2) \cdot y_2 \ldots k(x,x_n) \cdot y_n] \cdot w + \beta\big)$$

$$= \Big(\sum_{i=1}^{n} w_i \cdot k(x, x_i)\, y_i\Big) + \beta_0$$

This lecture

- KKT conditions
  - SVM as an example

- Online Learning

Optimality conditions: the conditions that characterizes the optimal solutions

- What you learned in high school

$$\min_{x \in \mathbb{R}} x^2 - 4x + 9 = f(x)$$

$\cancel{x} \in [0,1]$ $\qquad \underline{\nabla f(x^*) = 0}$ $\qquad$ if $\underline{f''(x) \geq 0}$

- Slight generalization: For convex and differentiable objective function

$$\min_{\underline{x \in \mathbb{R}^d}} f(x) \qquad \nabla f(x) = 0 \qquad \nabla^2 f(x) \succeq 0$$

Does not handle non-differentiable functions, does not handle constraints.

Handling constraints with first-order optimality conditions

For a convex problem

$\forall y, x \in C \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \geq f(x)$

$$\min_x \; f(x) \;\; \text{subject to} \;\; x \in C$$

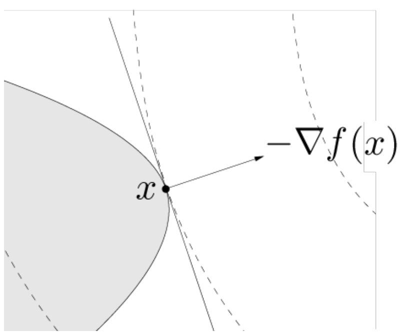
$x^* = \arg\min_{x \in C} f(x)$

and differentiable f, a feasible point x is optimal if and only if

$$\underline{\nabla f(x)^T (y - x) \geq 0} \quad \text{for all } y \in C$$

$f(x) \leq f(x^*)$

$\Rightarrow x$ is optimal



This is called the first-order condition for optimality

In words: all feasible directions from x are aligned with gradient $\nabla f(x)$

Important special case: if $\underline{C = \mathbb{R}^n}$ (unconstrained optimization), then optimality condition reduces to familiar $\nabla f(x) = 0$

Handling non-differentiable functions with **“subgradient”**

Recall that for convex and differentiable f,

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \text{for all } x, y$$

I.e., linear approximation always underestimates f

A subgradient of a convex function f at x is any $g \in \mathbb{R}^n$ such that

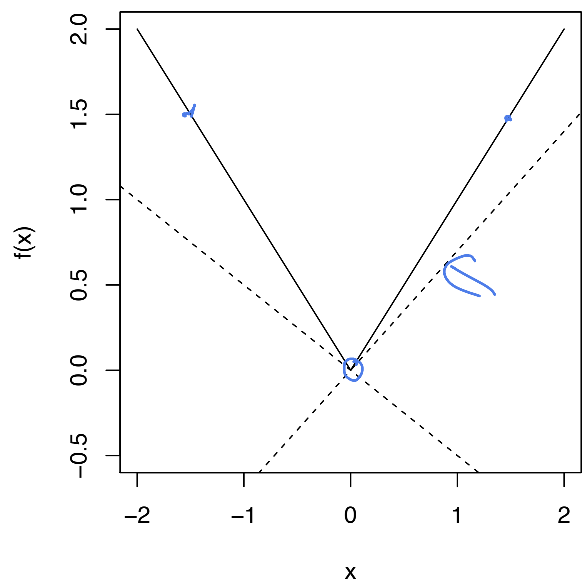
$$f(y) \geq f(x) + g^T (y - x) \quad \text{for all } y$$

- Always exists[1]
- If f differentiable at x, then $g = \nabla f(x)$ uniquely
- Same definition works for nonconvex f (however, subgradients need not exist)

[1] On the relative interior of $\mathrm{dom}(f)$

Examples of subgradients

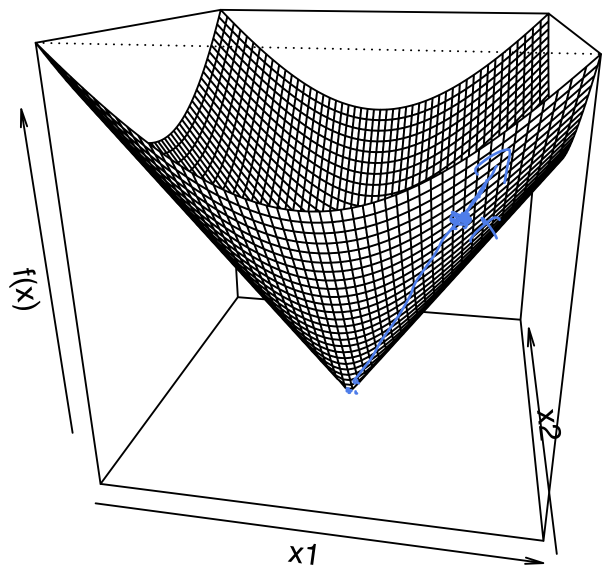
Consider $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$



$f(0) + \theta^T(y-0)$

- For $x \neq 0$, unique subgradient $g = \mathrm{sign}(x)$
- For $x = 0$, subgradient g is any element of $[-1, 1]$

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$



- For $x \neq 0$, unique subgradient $g = \underline{x/\|x\|_2}$
- For $x = 0$, subgradient g is any element of $\{z : \|z\|_2 \leq 1\}$

Subdifferential

Set of all subgradients of convex f is called the subdifferential:

$$\partial f(x) = \{g \in \mathbb{R}^n : g \text{ is a subgradient of } f \text{ at } x\}$$

- Nonempty (only for convex f)
- $\partial f(x)$ is closed and convex (even for nonconvex f)
- If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- If $\partial f(x) = \{g\}$, then f is differentiable at x and $\nabla f(x) = g$

First order optimality condition with subgradient

$$\boxed{\begin{array}{l} \min f_0(x) \\ \text{s.t. } x \in C \end{array}}$$

$$f(x) = f_0(x) + I_C(x)$$

For any f (convex or not),

$$f(x^\star) = \min_x f(x) \quad \Longleftrightarrow \quad 0 \in \partial f(x^\star)$$

I.e., $x^\star$ is a minimizer if and only if 0 is a subgradient of f at $x^\star$. This is called the subgradient optimality condition

Why? Easy: $g = 0$ being a subgradient means that for all y

$$f(y) \geq f(x^\star) + 0^T(y - x^\star) = f(x^\star)$$

Note the implication for a convex and differentiable function f, with $\partial f(x) = \{\nabla f(x)\}$

Karush-Kuhn-Tucker conditions

Given general problem

$$
\begin{aligned}
\min_{x} \quad & f(x) \\
\text{subject to} \quad & h_i(x) \le 0, \ i = 1, \dots m \\
& \ell_j(x) = 0, \ j = 1, \dots r
\end{aligned}
$$

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- $0 \in \partial\Big(f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)\Big)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \le 0, \ \ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \ge 0$ for all i (dual feasibility)

$x \in \operatorname{argmin}_x L(x, u, v)$

Necessity

Let $x^\star$ and $u^\star, v^\star$ be primal and dual solutions with zero duality gap (strong duality holds, e.g., under Slater's condition). Then

$$
\begin{aligned}
f(x^\star) &= g(u^\star, v^\star) \\
&= \min_x \; f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_j^\star \ell_j(x) \\
&\le f(x^\star) + \sum_{i=1}^m u_i^\star h_i(x^\star) + \sum_{j=1}^r v_j^\star \ell_j(x^\star) \\
&\le f(x^\star)
\end{aligned}
$$

stationarity

optimality

$f(x) \geq L(x,u,v)$

$= 0$

0 $\quad$ 0

complementary slackness

In other words, all these inequalities are actually equalities

Two things to learn from this:

- The point $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$—this is exactly the stationarity condition
- We must have $\sum_{i=1}^m \underline{u_i^\star h_i(x^\star)} = 0$, and since each term here is ≤ 0, this implies $u_i^\star h_i(x^\star) = 0$ for every i—this is exactly complementary slackness

Primal and dual feasibility hold by virtue of optimality. Therefore:

> If $x^\star$ and $u^\star, v^\star$ are primal and dual solutions, with zero duality gap, then $x^\star, u^\star, v^\star$ satisfy the KKT conditions

(Note that this statement assumes nothing a priori about convexity of our problem, i.e., of f, h_i, ℓ_j)

Sufficiency

If there exists $x^\star, u^\star, v^\star$ that satisfy the KKT conditions, then

$\min_x L(x, u^\star, v^\star)$

$$\begin{aligned} g(u^\star, v^\star) &= f(x^\star) + \underbrace{\sum_{i=1}^m u_i^\star h_i(x^\star)}_{=0} + \underbrace{\sum_{j=1}^r v_j^\star \ell_j(x^\star)}_{=0} \\ &= f(x^\star) \end{aligned}$$

where the first equality holds from stationarity, and the second holds from complementary slackness

Therefore the duality gap is zero (and $x^\star$ and $u^\star, v^\star$ are primal and dual feasible) so $x^\star$ and $u^\star, v^\star$ are primal and dual optimal. Hence, we've shown:

> If $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions, then $x^\star$ and $u^\star, v^\star$ are primal and dual solutions

Putting it together

In summary, KKT conditions:

- always sufficient
- necessary under strong duality

Putting it together:

For a problem with strong duality (e.g., assume Slater's condition: convex problem and there exists x strictly satisfying non-affine inequality contraints),

$$x^\star \text{ and } u^\star, v^\star \text{ are primal and dual solutions}$$
$$\iff x^\star \text{ and } u^\star, v^\star \text{ satisfy the KKT conditions}$$

(Warning, concerning the stationarity condition: for a differentiable function f, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless f is convex! There are other versions of KKT conditions that deal with local optima.)

Example: support vector machines

Given $y \in \{-1, 1\}^n$, and $X \in \mathbb{R}^{n \times p}$, the support vector machine problem is:

$$
\begin{aligned}
\min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^n \xi_i \\
\text{subject to} \quad & \xi_i \geq 0, \; i = 1, \ldots n \\
& y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \; i = 1, \ldots n
\end{aligned}
$$

$\beta^* = X^{uT} w^*$

Introduce dual variables $v, w \geq 0$. KKT stationarity condition:

$$
0 = \sum_{i=1}^n w_i y_i, \quad \beta = \sum_{i=1}^n w_i y_i x_i, \quad w = C1 - v
$$

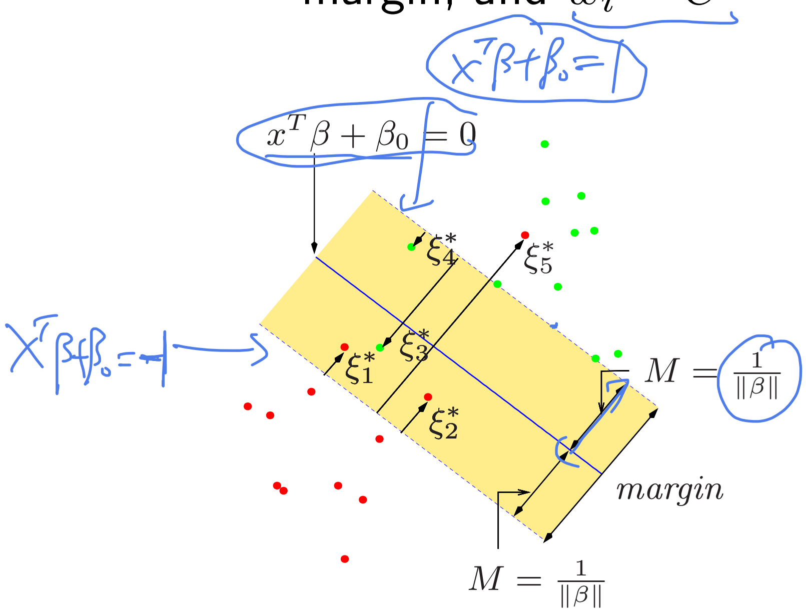
$\nabla_\beta L(\beta, w, v) = 0$

Complementary slackness:

$$
v_i \xi_i = 0, \; w_i\big(1 - \xi_i - y_i(x_i^T \beta + \beta_0)\big) = 0, \quad i = 1, \ldots n
$$

Hence at optimality we have $\beta = \sum_{i=1}^{n} w_i y_i x_i$, and w_i is nonzero only if $y_i(x_i^T \beta + \beta_0) = 1 - \xi_i$. Such points i are called the support points

- For support point i, if $\xi_i = 0$, then x_i lies on edge of margin, and $w_i \in (0, C]$;
- For support point i, if $\xi_i \neq 0$, then x_i lies on wrong side of margin, and $w_i = C$



$\frac{1-0}{\|\beta\|_2}$ distance between two hyperplanes

KKT conditions do not really give us a way to find solution, but gives a better understanding

In fact, we can use this to screen away non-support points before performing optimization

Checkpoint: KKT conditions and SVM

- A generalized set of conditions that characterizes the optimal solutions
  - Stationarity, complementary slackness, primal / dual feasibility
  - Always sufficient for optimality
  - Necessary when we have strong duality

- Complementary slackness implies
  - SVM dual solutions are sparse!
  - The number of “support vector”s is small

This lecture

- KKT conditions
  - SVM as an example

- Online Learning

Batch

Recap: Statistical Learning Setting

$(x_1, y_1) \dots (x_n, y_n) \overset{iid}{\sim} D$ $\quad x \in \mathcal{X} = \mathbb{R}^d$ or $\{0,1\}^d$

$y \in \mathcal{Y} = \{0,1\}$

$\mathcal{H}$ hypothesis class $\quad h \in \mathcal{H} \quad h: \mathcal{X} \to \mathcal{Y}$

Realizable case $\quad \exists\, h^* \in \mathcal{H}$ s.t. w.p. 1 $\quad h^*(x) = y$

$(x, y) \sim D$

Goal of learning: find $h \in \mathcal{H}$ s.t.

$$\mathrm{err}_D(h) = \mathop{\mathbb{E}}_{(x,y)\sim D}\left[\mathbb{1}\left(\underbrace{h^*(x)}_{=\, y} \neq h(x)\right)\right] \xrightarrow[n \to \infty]{} 0$$

function of data

(Adversarial) Online Learning Setting

- Data points show up sequentially (non-iid), learner makes online predictions

x_1 chosen by nature

$h_1 \leftarrow x_1$, predict $\hat{y}_1 = h_1(x_1)$

y_1 is revealed by nature,

⋮

$h_t \leftarrow (x_1, y_1) \ldots (x_{t-1}, y_{t-1}), x_t$, predict $h_t(x_t)$, receive y_t

$\text{loss}_t = \mathbb{1}(\hat{y}_t \neq y_t)$

- Performance metric: Mistake bounds M

if Alg A satisfies # of mistakes A makes $\leq M$

for all seq. of $(x_1, h^*(x_1)), (x_2, h^*(x_2)), \ldots$

Then Alg A has a ~~mistake~~ mistake bound of M.

Algorithm A "Consistency"

1. $V_1 = \mathcal{H}$
2. for $t = 1, 2, 3, \ldots$

   Receive x_t, pick any $h \in V_t$

   prediction $\hat{y}_t = h(x_t)$

   Receive $y_t = h^*(x_t)$

   update $V_{t+1} = \{ h \in V_t \mid h(x_t) = y_t \}$

$\boxed{\text{Check: } \forall h \in V_{t+1},\ h^*(x_i) = y_i \text{ for } \forall i = 1, 2, \ldots, t}$

Each mistake, we can eliminate at least 1 hypothesis

trivial upper bound

$M[\text{consistency}] \le |\mathcal{H}| - 1$

Example: $\mathcal{X} = \{1, 2, \ldots, |\mathcal{H}|\}$

$\mathcal{H} = \{h_1, \ldots, h_{|\mathcal{H}|}\}$

$h_i(x) = \begin{cases} 0 & \text{if } x < i \\ 1 & \text{otherwise} \end{cases}$

$x_1 = 1,\ x_2 = 2,\ \ldots,\ x_{|\mathcal{H}|} = |\mathcal{H}|,\ \ldots$

$y_1 = 0,\ y_2 = 0,\ \ldots,\ y_{|\mathcal{H}|} = 1,$

$h^* = h_{|\mathcal{H}|}$

predict $h_1, h_2, \ldots, h_{|\mathcal{H}|-1}, h_{|\mathcal{H}|}$

h_t classifies (x_i, y_i) $i = 1, \ldots, t-1$

but $h_t(x_t) \ne y_t$

Algorithm B "Halfing"

1. $V_1 = \mathcal{H}$
2. for $t = 1, 2, 3, \ldots$
   Receive x_t
   predict $\hat{y}_t = \underset{h \in V_t}{\mathrm{Vote}}(h(x_t)) = \underset{r \in \{0,1\}}{\arg\max} \left|\{h \in V_t \mid h(x_t) = r\}\right|$
   receive $y_t = h^*(x_t)$
   update $V_{t+1} = \{h \in V_t \mid h(x_t) = y_t\}$

Claim: $M(\text{halfing}) \le \log_2(|\mathcal{H}|)$

Proof: for each mistake, at least $\frac{|V_t|}{2}$ hypotheses are wrong.

$$|V_{t+1}| \le |V_t| \cdot \frac{1}{2}$$

$$1 \le |V_{t+1}| \le |\mathcal{H}| \cdot 2^{-M}$$

$$2^M \le |\mathcal{H}| \Rightarrow M \le \log_2 |\mathcal{H}| \quad \square$$

Now let's get rid of "Realizability". The setting is called "Agnostic learning"

Compete v.s. the best $h \in \mathcal{H}$ in hindsight

$$\text{Regret} = \sum_{t=1}^{T} \mathbb{1}(h_t(x_t) \neq y_t) - \min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathbb{1}(h(x_t) \neq y_t)$$

(x_t, y_t) chosen by adversary

as $T \to \infty$

if $\text{Regret}_T(T) = o(T)$

$\frac{1}{T} \text{Regret}(T) = o(1)$

Example: Stock forecasting

n experts

	Exp1 (Siqi)	Exp2 (Esha)	Exp3 (Lei)	Exp4 (Raffles the cat)	Outcome
Day1	Down	Up	Up	Down	Down
Day2	Up	Up	Down	Down	Down
Day3	Up	Down	Up	Up	Up
Weighted Majority	1	1	1	1	~~Day~~ Day1
	1	$\frac{1}{2}$	$\frac{1}{2}$	1	Day
	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	1	

Alg C Weighted Majority

predict $\hat{y}_t = \mathbb{1}\left(\underbrace{\sum_{h\in H} w_h^t \cdot h(x_t)}_{\text{weighted "Up"}} \geq \frac{1}{2} \underbrace{\sum_{h\in H} w_h^t}_{\text{total weight}}\right)$

receive y_t. discount $w_h^{t+1} = w_h^t \cdot \frac{1}{2}$ for $\forall h$ that made a mistake

Thm: $\underset{\text{\# of error WM makes}}{M} \leq 2.41(\underset{\text{min \# of mistakes}}{m} + \log \underset{\text{\# of experts}}{n})$

$\underline{M-m}$: regret

W is total weights $\boxed{\sum_{h\in H} w_h^t}$

Proof: claim Each Mistake $\Rightarrow$ W drop at least by 25%

$\underset{\substack{\text{best expert} \\ \text{is part of the pool}}}{\left(\frac{1}{2}\right)^m} \leq W \leq n \cdot \left(\frac{3}{4}\right)^M$

$M \cdot \log \frac{4}{3} \leq \log 2 \cdot m + \log n$

$M \leq \frac{\log 2 \cdot m + \log n}{\log \frac{4}{3}} \leq 2.41 \cdot (m + \log n)$

not 1

How do we fix "weighted majority"? Instead of discounting by 1/2, let's try discounting by 1-ϵ

Fact: For all $0 \le x \le 0.5$

$$-x - x^2 \le \log(1-x) \le -x$$

- Following the same analysis

$$(1-\varepsilon)^m \le W \le n \cdot \left(\frac{1}{2} + \frac{\varepsilon}{2}\right)^m$$

$$m \log(1-\varepsilon) \le \log n + m \log\left(\frac{1}{2} + \frac{\varepsilon}{2}\right)$$

$$m \le \frac{\log n - m\log(1-\varepsilon)}{-\log\left(\frac{1}{2} + \frac{\varepsilon}{2}\right)} = \frac{-\log(1-\varepsilon)}{-\log\left(\frac{1}{2}(1+\varepsilon)\right)} m - \frac{\log n}{\log\left(\frac{1}{2} + \frac{\varepsilon}{2}\right)}$$

$$\le \frac{\varepsilon + \varepsilon^2}{\frac{1}{2}(1-\varepsilon)} m + O(\log n)$$

$$= 2(1 + O(\varepsilon)) m + O(\log n)$$

Algorithm D: Randomized Weighted Majority

$n = |\mathcal{H}|$

At t, F_t: fraction of the weights of the expert making mistake.

$$(1-\varepsilon)^m \le \underbrace{W = n \cdot (1-\varepsilon F_1)(1-\varepsilon F_2)(1-\varepsilon F_3)\cdots(1-\varepsilon F_T)}$$

$$\to \underline{W_{t+1}} = (1-F_t) W_t + F_t W_t (1-\varepsilon)$$

$$= W_t - F_t W_t \varepsilon = \underline{(1-F_t\varepsilon) W_t}$$

$\boxed{\text{lem: } \log(1-x) \le -x}$

$$\log W = \log n + \sum_{t=1}^{T} \log(1-\varepsilon F_t)$$

$$\le \log n + \left(-\varepsilon \sum_t F_t\right)$$

$$= \log n - \varepsilon \cdot \underline{E[M]}$$

$(1-\varepsilon)^m \le W$

$(1-\varepsilon)^m \le (1-\varepsilon F_1)\cdots(1-\varepsilon F_m)$

$= W_{h^*} \le W$

$$m \log(1-\varepsilon) \le \log n - \varepsilon \underline{E[M]}$$

$$(1+\varepsilon) \;\; \| $$

$\varepsilon = \frac{1}{\sqrt{m}}$

$$E[M] \le \frac{m \cdot (-\log(1-\varepsilon))}{\varepsilon} + \frac{\log n}{\varepsilon} \le m\left(\frac{\varepsilon+\varepsilon^2}{\varepsilon}\right) + \frac{\log n}{\varepsilon} \le m + \sqrt{m} + \sqrt{m}\log n$$

$$\le m + \tilde{O}(\sqrt{T})$$

Analysis of RWM

From mistake bounds to loss minimization

- Loss function $\quad 0 \le \ell_t(h) \le 1 \quad \forall h$
- Regret $\quad \sum_t \ell_t(h_t) - \min_{h \in \mathcal{H}} \sum_t \ell_t(h)$
- The "Hedge" Algorithm:

① Pick expert h_t w.p. $\frac{W_t(h)}{\sum_{\tilde{h}} W_t(\tilde{h})}$

② incur loss $\ell_t(h_t)$

③ $W_h^{t+1} = W_h^t \cdot \underbrace{e^{-\varepsilon \ell_t(h)}}$ for $h \in \mathcal{H}$

OCO

optization view

$\min_{\theta \in \mathcal{H}} \sum_{t=1}^{T} f_t(\theta)$

online

f_t is revealed one at a time.

$f_t(\theta) = \left\langle \begin{pmatrix} \ell_t(h=1) \\ \ell_t(h=2) \\ \vdots \\ \ell_t(h=|\mathcal{H}|) \end{pmatrix}, \theta \right\rangle$

$\begin{cases} \|\theta\|_1 \le 1 \\ \theta \ge 0 \end{cases}$

$\|\ell_T\|_\infty \le 1$

Checkpoint: Online Learning

- Learning with expert advice
  - A summary of regret bound: # mistakes - Oracle # of mistakes

	Consistency	Halfing	Weighted Majority	Randomized WM
Realizable setting	$\min(T, \lvert\mathcal{H}\rvert)$	$\min(T, \log\lvert\mathcal{H}\rvert)$	$\min(T, \log\lvert\mathcal{H}\rvert)$	$\min(T, \log\lvert\mathcal{H}\rvert)$
Agnostic setting	n.a.	n.a.	$(2 + \epsilon)m + \log\lvert\mathcal{H}\rvert/\epsilon$	$\sqrt{m \log\lvert\mathcal{H}\rvert} = O(\sqrt{T \log\lvert\mathcal{H}\rvert})$

Next lecture

- Online Learning (Part II)
  - Online Gradient Descent
- Reinforcement Learning
  - Markov Decision Processes