Computational and Statistical Learning Theory

TTIC 31120

Prof. Nati Srebro

Lecture 15:

Back to Online Learning

The Perceptron Algorithm

Online Regret

Online Learning Process

- At each time $t = 1,2, \ldots$
  - We receive an instance $x_t \in \mathcal{X}$ (receive an email)
  - We predict a label $\hat{y}_t = h_t(x_t)$ (predict if its spam)
  - We see the correct label y_t of x_t (user tells us if it was really spam)
  - We update the predictor h_{t+1} based on (x_t, y_t)

- Learning rule: mapping $A: (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$
  - $h_t = A\big((x_1, y_1), (x_2, y_2), \ldots, (x_{t-1}, y_{t-1})\big)$

- Goal: make few mistakes $\hat{y}_t \neq y_t$

- Learning Rule A attains mistake bound $M(m)$ on hypothesis class $\mathcal{H}$, if for any sequence $(x_1, y_1) \ldots (x_m, y_m)$ **realizable by $\boldsymbol{\mathcal{H}}$** (i.e. $y_i = h(x_i)$ for some $h \in \mathcal{H}$), the rule A makes at most $M(m)$ mistakes:

$$|\{t \mid h_t(x_t) \neq y_t\}| \leq M(m)$$

$$h_t = A\big((x_1, y_1), \ldots, (x_{t-1}, y_{t-1})\big)$$

Why online?

- Data arrives in a “steam” and needs to be labeled “online”
  - E.g. SPAM, weather, investing, …
- Avoid i.i.d.
  - Allow arbitrary dependence between samples
  - Allow non-stationarity (distribution changes over time)
- More efficient “online”/”realtime” learning rules
  - Small update after each example is received?
- Understand relationship between learning and optimizatoin

Halving

$$HALVING_{\mathcal{H}}(S)(x) = MAJORITY(h(x) \mid h \in \mathcal{H}, L_S(h) = 0)$$

- $HALVING_{\mathcal{H}}$ attains a mistake bound of $M(m) \leq \log|\mathcal{H}|$ on $\mathcal{H}$
- Mistake bound matches $O\left({}^{\log|\mathcal{H}|}/_{\epsilon}\right)$ sample complexity of PAC learning

Online Learning Linear Predictors

$$\mathcal{H} = \{\, x \mapsto sign(\langle w, x \rangle) \mid w \in \mathbb{R}^d \,\}$$

- Can PAC learn with $O\left(\frac{\boldsymbol{d}}{\epsilon^2}\right)$ samples
  $\exists_w L^{01}(w) = 0$ ➔ can get $L^{01}\big(A(S)\big) \leq \epsilon$ using $m = O\left(\frac{\boldsymbol{d}}{\epsilon^2}\right)$ samples

- Can't online learn: even in two dimensions (or one dimension with a bias term), for any learning rule there exists a sequence on which the rule makes a mistake on every point (i.e. $M(m) \geq m$)

- $\exists_w L_S^{\boldsymbol{mrg}}\left(\frac{w}{\boldsymbol{\gamma}}\right) = 0$ (i.e. $\forall_t y_t \langle w, x_t \rangle \geq \boldsymbol{\gamma}$) ➔
  In PAC model, can get $L^{01}\big(A(S)\big) \leq \epsilon$ using $m = O\left(\frac{\|\boldsymbol{w}\|_2^2 (\mathbf{sup}\|\boldsymbol{x}\|_2^2)}{\boldsymbol{\gamma}^2 \epsilon^2}\right)$ samples

Online Perceptron Rule

Init $w_1 = 0$
At iteration t:
- Receive x_t
- Predict $\hat{y}_t = sign(\langle w_t, x_t \rangle)$
- Receive y_t
- If $y_t \neq \hat{y}_t$,
  $$w_{t+1} \leftarrow w_t + y_t x_t$$
  else: $w_{t+1} \leftarrow w_t$

- Theorem: if $\exists_w \forall_t y_t \langle w, x_t \rangle \geq \gamma$ (i.e. $L_S^{mrg}\left(\frac{w}{\gamma}\right) = 0$) then the number of mistakes made by Perceptron is at most

$$M(m) \leq \frac{\|\boldsymbol{w}\|_2^2 (\mathbf{sup} \|\boldsymbol{x}\|_2^2)}{\boldsymbol{\gamma}^2}$$

- Conclusion: Run iterative on a separable training set S (i.e. multiple epochs, use example $i = t \bmod m$ at iteration t), then after $\leq {\|w\|_2^2 \|x\|^2}/{\gamma^2}$ iterations, $L_S^{01}(w) = 0$

Perceptron Analysis

Init $w_1 = 0$
At iteration t:
- Predict $\hat{y}_t = sign(\langle w_t, x_t \rangle)$
- If $y_t \neq \hat{y}_t,\ w_{t+1} \leftarrow w_t + y_t x_t$
  else: $w_{t+1} \leftarrow w_t$

- Denote $M_t = \#mistakes\ in\ rounds\ 1..t$
- Assume $\|x\| \leq 1$ and $y_t \langle \boldsymbol{w}, x_t \rangle \geq \gamma = 1$; prove $M_m \leq \|\boldsymbol{w}\|^2$

Claim 1: $\langle \boldsymbol{w}, w_{t+1} \rangle = \langle \boldsymbol{w}, \sum_{i=1..t, \hat{y}_i \neq y_i} y_i x_i \rangle = \sum_{i=1..t, \hat{y}_i \neq y_i} \underbrace{y_i \langle \boldsymbol{w}, x_i \rangle}_{\geq \gamma = 1} \geq M_t$

Claim 2: $\|w_{t+1}\|_2^2 \leq M_t$

- Induction base: $\|w_1\|_2 = 0$
- If no mistake at round t: $\|w_{t+1}\|_2^2 = \|w_t\|_2^2 = M_{t-1} = M_t$
- If mistake: $\|w_{t+1}\|_2^2 = \|w_t + y_t x_t\|_2^2 = \|w_t\|_2^2 + 2\underbrace{y_t \langle w_t, x_t \rangle}_{\leq 0} + \|x_t\|_2^2$
  $\leq \|w_t\|_2^2 + 1 \leq M_{t-1} + 1 = M_t$

Conclusion: $M_m \underset{\text{claim 1}}{\leq} \langle \boldsymbol{w}, w_{m+1} \rangle \leq \|\boldsymbol{w}\| \cdot \|w_{m+1}\| \underset{\text{claim 2}}{\leq} \|\boldsymbol{w}\| \sqrt{M_m}$ ➔ $M_m \leq \|\boldsymbol{w}\|_2^2$

Online Learning Linear Predictors

- Using Perceptron, can get

$$\exists_w L_S^{mrg}(w) = 0 \Rightarrow M(t) \le \|\boldsymbol{w}\|_2^2 \|\boldsymbol{x}\|_2^2$$

- "Matches" statistical guarantee

$$\exists_w L_{\mathcal{D}}^{mrg}(w) = 0 \Rightarrow L_{\mathcal{D}}^{01}(\widehat{w}^{mrg}) \le \epsilon \text{ using } m = O\left(\frac{\|\boldsymbol{w}\|^2 \|\boldsymbol{x}\|^2}{\epsilon^2}\right)$$

- Where is this coming from???

- Can this be related to ramp/hinge loss?
- Applied to other losses?
- Other hypothesis classes?
- Non realizable?

Non-Realizable Online Learning: Online Regret

- Learning problem specified by: $\ell: \overline{\mathcal{H}} \times \mathcal{Z} \to \mathbb{R}$
  - Supervised learning: $\ell\big(h,(x,y)\big) = loss(h(x);y)$
- Learning rule: $A: \mathcal{Z}^* \to \overline{\mathcal{H}}$
  - $h_t = A(z_1, \ldots, z_{t-1})$
  - Suffer loss $\ell(h_t, z_t) = loss(h_t(x_t); y_t)$
- **Regret** of A on sequence $z_1, z_2, \ldots, z_m$ relative to hypothesis class $\mathcal{H} \subseteq \overline{\mathcal{H}}$:

$$\frac{1}{m}\sum_{t=1}^{m} \ell(A(z_1, \ldots, z_{t-1}), z_t) - \inf_{h \in \mathcal{H}} \frac{1}{m}\sum_{t=1}^{m} \ell(h, z_t)$$

- We say rule A **attains regret $\boldsymbol{Reg(m)}$** on $\mathcal{H}$ if for any sequence:

$$\frac{1}{m}\sum_{t=1}^{m} \ell(A(z_1, \ldots, z_{t-1}), z_t) \leq \inf_{h \in \mathcal{H}} \frac{1}{m}\sum_{t=1}^{m} \ell(h, z_t) + Reg(m)$$

- Mistake bound: $Reg(m) = \frac{M(m)}{m}$ for $loss^{01}$ on realizable sequences

Non-Realizable Online Learning: Online Regret

- Learning problem specified by: $\ell: \overline{\mathcal{H}} \times \mathcal{Z} \to \mathbb{R}$
  - Supervised learning: $\ell\big(h,(x,y)\big) = loss(h(x);y)$
  - Investment: $\ell(w,z) = -\langle w,z \rangle$
- Learning rule: $A: \mathcal{Z}^* \to \overline{\mathcal{H}}$
  - $h_t = A(z_1, \dots, z_{t-1})$
  - Suffer loss $\ell(h_t, z_t) = loss(h_t(x_t); y_t)$ or $-\langle w_t, z_t \rangle$

$w \in \mathbb{R}^d$
investment portfolio
$w[i]$=holding in stock i

$z \in \mathbb{R}^d$
market behavior
$z[i]$=return on stock i

- **Regret** of A on sequence $z_1, z_2, \dots, z_m$ relative to hypothesis class $\mathcal{H} \subseteq \overline{\mathcal{H}}$:

$$\frac{1}{m}\sum_{t=1}^{m} \ell(A(z_1,\dots,z_{t-1}), z_t) - \inf_{h\in\mathcal{H}} \frac{1}{m}\sum_{t=1}^{m} \ell(h, z_t)$$

- We say rule A **attains regret $\boldsymbol{Reg(m)}$** on $\mathcal{H}$ if for any sequence:

$$\frac{1}{m}\sum_{t=1}^{m} \ell(A(z_1,\dots,z_{t-1}), z_t) \le \inf_{h\in\mathcal{H}} \frac{1}{m}\sum_{t=1}^{m} \ell(h, z_t) + Reg(m)$$

- Mistake bound: $Reg(m) = \frac{M(m)}{m}$ for $loss^{01}$ on realizable sequences

Follow The Leader

$$FTL_{\mathcal{H}}(S) = \arg\min_{h\in\mathcal{H}} L_S(h)$$

I.e., at each iteration t:

$$h_t = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{t-1} \ell(h, z_i)$$

Use with h_t and suffer loss $h_t(z_t)$

A rule for prophets—Be The Leader (BTL):

$$h_t = \arg\min_{h\in\mathcal{H}} \sum_{i=1}^{t} \ell(h, z_i)$$

Claim: $Reg_{BTL}(m) \leq 0$

Proof by induction that for any $h \in \mathcal{H}$, $\sum_{i=1}^{t} \ell(h_i, z_i) \leq \sum_{i=1}^{t} \ell(h, z_i)$:

$$\sum_{i=1}^{t-1} \ell(h_i, z_i) + \ell(h_t, z_t) \leq \sum_{i=1}^{t-1} \ell(h_t, z_i) + \ell(h_t, z_t) = \sum_{i=1}^{t} \ell(h_t, z_i) \leq \sum_{i=1}^{t} \ell(h, z_i)$$

Inductive hypothesis, applied to $h = h_t$

Optimality of h_t (definition of BTL)

Stability and Online Regret

- **Definition**: A rule is (leave-last-out) $\beta(m)$**-stable** if, for all $z_1, \dots, z_m$:

$$|\ell(A(z_1, \dots, z_m), z_m) - \ell(A(z_1, \dots, z_{m-1}), z_m)| \le \beta(m)$$

- Follow-The-Leader (FTL): $h_t = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{t-1} \ell(h, z_i)$
- Be-The-Leader (BTL): $h_t = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{t} \ell(h, z_i)$

If FTL is $\beta(m)$-stable:

$$Reg_{FTL}(m) = \frac{1}{m}\sum_{i=1}^{m} \ell(h_i^{FTL}, z_i) \le \frac{1}{m}\sum_{i=1}^{m}\left(\ell\left(h_i^{BTL}, z_i\right) + \beta(i)\right) = \cancel{Reg_{BTL}(m)} + \frac{1}{m}\sum_{i=1}^{m}\beta(i)$$

- **Conclusion**: If FTL is $\beta(m)$-stable, then it has regret

$$Reg(m) \le \frac{1}{m}\sum_{i=1}^{m}\beta(i)$$

When is FTL Stable?

- Example: squared-loss tracking (center of mass)

$$\mathcal{Z} = \left\{z \in \mathbb{R}^d \;\; \|z\|_2 \leq 1\right\} \qquad \mathcal{H} = \mathbb{R}^d \qquad \ell(w,z) = \|w-z\|_2^2$$

- $w_{t+1} = FTL(z_1, \dots, z_t) = \frac{1}{t}\sum_{i=1}^{t} z_i = \left(1-\frac{1}{t}\right) w_t + \frac{1}{t} z_t$

- Hence: $|\ell(FTL(z_1, \dots, z_t), z_t) - \ell(FTL(z_1, \dots, z_{t-1}), z_t)| = |\ell(w_{t+1}, z_t) - \ell(w_t, z_t)|$

$$= \left| \left\| \left(1-\frac{1}{t}\right) w_t + \frac{1}{t} z_t - z_t \right\|^2 + \|w_t - z_t\|^2 \right| = \left| \left\| \left(1-\frac{1}{t}\right)(w_t - z_t) \right\|^2 - \|w_t - z_t\|^2 \right|$$

$$= \left(1 - \left(1-\frac{1}{t}\right)^2\right) \|w_t - z_t\|^2 \leq \frac{2}{t} \|w_t - z_t\|^2 \leq \frac{2}{t} \cdot 4$$

- Conclusion: $FTL_{\mathcal{H}}$ is $\beta(m) = \frac{8}{m}$ stable.

➔ It attains regret $Reg(m) \leq \frac{1}{m}\sum_{i=1}^{m} \frac{8}{i} \leq \frac{8(\ln m + 1)}{m}$

Convex Lipschitz Bounded Problems

- Recall our interest in convex Lipschitz bounded problems:
  - $\mathcal{H} \subseteq \mathbb{R}^d, \forall_{w \in \mathcal{H}} \|w\| \leq B$
  - $\ell(w, z)$ convex and G-Lipschitz wrt $\|w\|$

- Is FTL for a convex Lipschitz bounded problem always stable?
  - Same as asking if ERM is stable—we already saw this is not the case

- Even if perhaps not stable, does it attain diminishing regret?

FTL for a Linear Problem

$$\mathcal{Z} = [-1,1] \qquad \mathcal{H} = [-1,1] \qquad \ell(h,z) = h \cdot z$$

- $FTL(z_1, \dots, z_t) = \begin{cases} -1, & \sum_{i=1}^{t-1} z_i > 0 \\ 1, & \sum_{i=1}^{t-1} z_i < 0 \end{cases}$

- Consider the sequence $0.5, -1, 1, -1, 1, -1, 1, -1, 1, -1, 1, \dots$

- With FTL, $h_t = (-1)^t$ and $Reg(m) = \frac{m-1}{m} - 0 \to 1$

(Can get similar behavior with $\ell(h,z) = loss^{hinge}(hx, y)$)

Follow the Regularized Leader

$$FTRL(S) = \arg\min_{w\in\mathbb{R}^d} L_S(w) + \lambda\|w\|_2^2$$

- Claim: $\ell(w,z)$ is convex and G-Lipschitz wrt $\|w\|_2$,

➔ $FTRL$ is $\frac{G^2}{\lambda m}$-stable (leave last out, as well as replacement)

- Observe: FTRL is FTL for the modified loss $\tilde{\ell}(w,z) = \ell(w,z) + \lambda\|w\|_2^2$

➔ FTL of $\tilde{\ell}$ is stable, and can apply regret guarantee:

For any w,

$$\frac{1}{m}\sum_{i=1}^{m} \ell\,(w_t, z_t) \le \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}(w_t, z_t) \le \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}(w, z_t) + \frac{1}{m}\sum_{i=1}^{m} \frac{G^2}{\lambda i}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \ell\,(w, z_t) + \frac{1}{m}\sum_{i=1}^{m} \lambda\|w\|_2^2 + \frac{1}{m}\sum_{i=1}^{m} \frac{G^2}{\lambda i}$$

Follow the Regularized Leader

$$FTRL(S) = \arg\min_{w \in \mathbb{R}^d} L_S(w) + \lambda \|w\|_2^2$$

- Claim: $\ell(w,z)$ is convex and G-Lipschitz wrt $\|w\|_2$,

  ➔ $FTRL$ is $\frac{G^2}{\lambda m}$-stable (leave last out, as well as replacement)

- Observe: FTRL is FTL for the modified loss $\tilde{\ell}(w,z) = \ell(w,z) + \lambda\|w\|_2^2$

  ➔ FTL of $\tilde{\ell}$ is stable, and can apply regret guarantee:

For any $\|w\|_2^2 \leq B^2$,

$$\frac{1}{m}\sum_{i=1}^m \ell(w_t, z_t) \leq \frac{1}{m}\sum_{i=1}^m \tilde{\ell}(w_t, z_t) \leq \frac{1}{m}\sum_{i=1}^m \tilde{\ell}(w, z_t) + \frac{1}{m}\sum_{i=1}^m \frac{G^2}{\lambda i}$$

$$= \frac{1}{m}\sum_{i=1}^m \ell(w, z_t) + \frac{1}{m}\sum_{i=1}^m \lambda\|w\|_2^2 + \frac{1}{m}\sum_{i=1}^m \frac{G^2}{\lambda i}$$

$$\leq \frac{1}{m}\sum_{i=1}^m \ell(w, z_t) + \lambda B^2 + \frac{\ln m + 1}{m}\frac{G^2}{\lambda} \leq O\left(\sqrt{\frac{G^2 B^2 \log m}{m}}\right)$$

$$\lambda = \sqrt{G^2 \log m / mB^2}$$

Follow the Regularized Leader

$$FTRL(S) = \arg\min_{w \in \mathbb{R}^d} L_S(w) + \lambda \Psi(w)$$

- Claim: $\ell(w, z)$ is convex and G-Lipschitz wrt $\|w\|$, and $\Psi(w) \geq 0$, α-strongly convex wrt $\|w\|$,

  ➔ $FTRL$ is $\frac{2G^2}{\lambda \alpha m}$-stable (even $\frac{2G^2}{\lambda \alpha (2m-1)}$-stable)

- Observe: FTRL is FTL for the modified loss $\tilde{\ell}(w, z) = \ell(w, z) + \lambda \Psi(w)$

  ➔ FTL of $\tilde{\ell}$ is stable, and can apply regret guarantee:

For any $\Psi(w) \leq \tilde{B}^2$,

$$\frac{1}{m}\sum_{i=1}^{m} \ell\left(w_t, z_t\right) \leq \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}(w_t, z_t) \leq \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}(w, z_t) + \frac{1}{m}\sum_{i=1}^{m} \frac{2G^2}{\lambda \alpha i}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \ell\left(w, z_t\right) + \frac{1}{m}\sum_{i=1}^{m} \lambda \Psi(w) + \frac{1}{m}\sum_{i=1}^{m} \frac{2G^2}{\lambda \alpha i}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m} \ell\left(w, z_t\right) + \lambda \tilde{B}^2 + \frac{\ln m + 1}{m} \frac{2G^2}{\lambda \alpha} \leq O\left(\sqrt{\frac{G^2 \tilde{B}^2 \log m}{\alpha\, m}}\right)$$

$$\lambda = \sqrt{G^2 \log m / \alpha m \tilde{B}^2}$$

Refined FTRL

$$FTRL(z_1, \dots, z_{t-1}) = \arg\min_{w \in \mathbb{R}^d} L_{z_1,\dots,z_{t-1}}(w) + \boldsymbol{\lambda_t}\Psi(w)$$

- Claim: $\ell(w,z)$ is convex and G-Lipschitz wrt $\|w\|$, and $\Psi(w) \geq 0$, α-strongly convex wrt $\|w\|$,

$\Rightarrow$ $FTRL$ is $\frac{2G^2}{\boldsymbol{\lambda_m}\alpha m}$-stable

- Observe: FTRL is FTL for the modified loss $\tilde{\ell}(w,z) = \ell(w,z) + \boldsymbol{\lambda_m}\Psi(w)$

$\Rightarrow$ FTL of $\tilde{\ell}$ is stable, and can apply regret guarantee:

For any $\Psi(w) \leq \tilde{B}^2$,

$$\frac{1}{m}\sum_{i=1}^{m} \ell\,(w_t, z_t) \leq \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}(w_t, z_t) \leq \frac{1}{m}\sum_{i=1}^{m} \tilde{\ell}(w, z_t) + \frac{1}{m}\sum_{i=1}^{m} \frac{2G^2}{\boldsymbol{\lambda_i}\alpha i}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m} \ell\,(w, z_t) + \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{\lambda_i}\Psi(w) + \frac{1}{m}\sum_{i=1}^{m} \frac{2G^2}{\boldsymbol{\lambda_i}\alpha i}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m} \ell\,(w, z_t) + \frac{1}{m}\sum_{i=1}^{m} \left(\boldsymbol{\lambda_i}\tilde{B}^2 + \frac{2G^2}{\boldsymbol{\lambda_i}\alpha i}\right) \leq \frac{1}{m}\sum_{i=1}^{m} \ell\,(w, z_t) + \frac{1}{m}\sum_{i=1}^{m} \sqrt{\frac{8G^2\tilde{B}^2}{\alpha i}}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m} \ell\,(w, z_t) + \sqrt{\frac{32G^2\tilde{B}^2}{\alpha m}}$$

$$\boldsymbol{\lambda_i} = \sqrt{2G^2 / \alpha\tilde{B}^2 \boldsymbol{i}}$$

Online Learning
Convex Lipschitz Bounded Problems

FTRL: $h_t = \arg\min_{w \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{t-1} \ell(w, z_i) + \lambda_i \boldsymbol{\Psi}(\boldsymbol{w})$

$\boldsymbol{\Psi}(\boldsymbol{w}) \geq 0$ is α-strongly convex wrt $\|w\|$

- Conclusion: Using $\lambda_i = \sqrt{\frac{\alpha G^2}{\tilde{B}^2 i}}$, FTRL attains regret $O\left(\sqrt{\frac{G^2 \tilde{B}^2}{\alpha m}}\right)$ for convex G-Lipschitz wrt $\|w\|$, and $\sup_{w \in \mathcal{H}} \Psi(w) \leq \tilde{B}^2$

Convex G-Lipschitz, B-Bounded wrt $\|w\|_2$: regret $O\left(\sqrt{\frac{G^2 \|w\|_2^2}{m}}\right)$

- "Matches" statistical excess error
  ("regret" versus best possible expected error)

- But very non-online-ish rule (not a simple update of previous iterate)

	Online	Statistical
Finite Cardinality	$\log \lvert\mathcal{H}\rvert$ Halving **agnostic???**	$\log \lvert\mathcal{H}\rvert$ ERM also agnostic
Finite Dimension	∞	$VCdim$ ERM also agnostic
Convex and Scale Sensitive	$\lVert w\rVert_2^2$ or $\sup \Psi(w)$ FTRL also agnostic	$\lVert w\rVert_2^2$ or $\sup \Psi(w)$ RERM (or ERM if $\ell(w,z) = loss(\langle w, \phi(z)\rangle, z)$ also agnostic

"steaming"/"online-ish" rule?
More like Perceptron?