Generalization Bounds for Averaged Classifiers/Data Dependent Concentration Bounds for Sequential Prediction Algorithms

Yoav Freund, Yishay Mansour and Robert Schapire, Annal of Statistics, 2004/Tong Zhang, Conference of Learning Theory, 2005

Feb. 6th, 2014

presented by Chicheng Zhang chichengzhang@ucsd.edu

Outline

Section 1

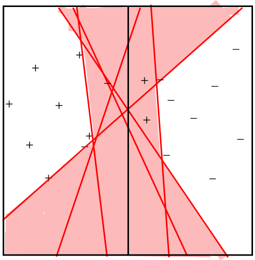
Generalization Bounds for Averaged Classifiers

Summary

- Proposed a pseudo-Bayesian algorithm
- Introduced a reject option for classification
- The error rate independent of complexity of $\mathcal{H}$
- The rejection rate upper bound decreases asymptotically comparable to error of ERM classifier

Introduction

- Overfitting
  - A classification problem w/ insufficient training data
  - ERM: w.p. $1-\delta$, $\epsilon(\hat{h}) \le \epsilon(h^*) + \sqrt{\epsilon(h^*)\frac{\ln|\mathcal{H}|+\ln\frac{1}{\delta}}{m}} + \frac{\ln|\mathcal{H}|+\ln\frac{1}{\delta}}{m}$
- Avoid: Saying Don't Know on part of test examples
  - Agnostic selective classification [EYW11]: Version Space, Disagreement-Based
  - This work: weighted average over $\mathcal{H}$, taking voting "margin" into account



- Generally, Pr(Don't Know) $\uparrow$, Pr(Mistake) $\downarrow$.
- Goal: find an alg. such that both Pr(Don't Know) and Pr(Mistake) can be controlled.

Preliminaries

- Batch Learning(Rather than online learning!), distribution $\mathcal{D}$ defined over $\mathcal{X} \times \{-1, +1\}$, $(x_1, y_1), \ldots, (x_m, y_m) \sim \mathcal{D}$ iid.
- a hypothesis class $\mathcal{H}$, Each classifier $h \in \mathcal{H}$, $h : \mathcal{X} \to \{-1, +1\}$
- True error $\epsilon(h) = \Pr(h(X) \neq Y)$, error of the optimal classifier $\epsilon = \epsilon(h^*) = \min_{h \in \mathcal{H}} \epsilon(h)$, Empirical error $\hat{\epsilon}(h) = \frac{1}{m} \sum_{i=1}^{m} I(h(x_i) \neq y_i)$, empirical risk minimizer $\hat{\epsilon}(\hat{h}) = \min_{h \in \mathcal{H}} \hat{\epsilon}(h)$
- $\mathcal{H}_x^+ = \{h \in \mathcal{H} : h(x) = +1\}$, $\mathcal{H}_x^- = \{h \in \mathcal{H} : h(x) = -1\}$

Algorithm: Intuition

- Inspired by Exponential Weight algorithm. Recall: "Bayesian" $w_{i,t+1} \propto e^{-\eta L_{i,t}}$.
- Translated into binary classification in batch case: $w(h) \propto e^{-\eta \hat{\epsilon}(h)}$
- "Softly" Put higher weight over classifiers performing well.
- Algorithm:
  - $\hat{\ell}(x) = \frac{1}{\eta} \ln(\frac{\sum_{h(x)=+} e^{-\eta \hat{\epsilon}(h)}}{\sum_{h(x)=-} e^{-\eta \hat{\epsilon}(h)}}) = \frac{1}{\eta} \ln(\frac{\sum_{h \in \mathcal{H}_x^+} e^{-\eta \hat{\epsilon}(h)}}{\sum_{h \in \mathcal{H}_x^-} e^{-\eta \hat{\epsilon}(h)}})$
  - if $|\hat{\ell}(x)| \leq \Delta$, then predict 0 (Saying Don't Know).
  - otherwise, predict w/ sign($\hat{\ell}(x)$).
- How to choose Δ? Will analyze a related quantity $\ell(x) = \frac{1}{\eta} \ln(\frac{\sum_{h \in \mathcal{H}_x^+} e^{-\eta \epsilon(h)}}{\sum_{h \in \mathcal{H}_x^-} e^{-\eta \epsilon(h)}})$ first.

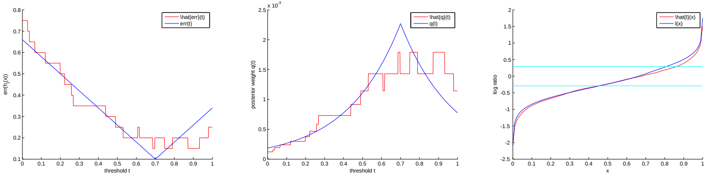
A Toy Example

- Threshold classifier $h_t = 2I(x > t) - 1$, uniform distribution on [0,1],

$$\Pr(y = +1|x) = \begin{cases} 0.9 & x \geq 0.7 \\ 0.1 & x < 0.7 \end{cases}$$

  $h^* = h_{0.7}$

- Discretize $\mathcal{H} = \{h_t, t = 0, 0.001, \ldots, 1\}$
- $m = 20$



Original Theorem and Proof

- Intuitively, $\eta \uparrow$, the performance of $\ell(x)$ gets closer to performance of h^*.
- What if there are two h^*'s disagreeing on a non-negligible region?

Theorem

Let $\eta > 0$*,* $\Delta \geq 0$*,* $\Delta\eta \leq 1/2$*. Then* $\forall \gamma \geq \frac{\ln 8|\mathcal{H}|}{\eta}$*,*

$$\Pr(y\ell(x) \leq 0) \leq 2(1 + 2|\mathcal{H}|e^{-\eta\gamma})(\epsilon + \gamma)$$

$$\begin{aligned}\Pr(y\ell(x) \leq 2\Delta) &\leq (1 + e^{2\Delta\eta})(1 + 2|\mathcal{H}|e^{\eta(2\Delta-\gamma)})(\epsilon + \gamma) \\ &\leq 4(1 + 2|\mathcal{H}|e^{\eta(2\Delta-\gamma)})(\epsilon + \gamma)\end{aligned}$$

- The factor 2 is unavoidable for such voting methods.

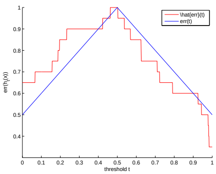
A Bad Example for Voting

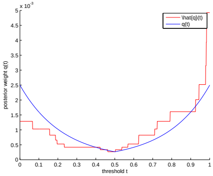
- Threshold classifier $h_t = 2I(x > t) - 1$, uniform distribution on [0,1],

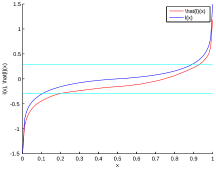
$$\Pr(y = +1|x) = \begin{cases} 0 & x \geq 0.5 \\ 1 & x < 0.5 \end{cases}$$

$$h^* = h_0, h_1$$

- For any finite η,$\Pr(y\ell(x) \leq 0) = 1$!







Proof I

- Define "weak" classifiers: weak $= \{h \in \mathcal{H} : \epsilon(h) \geq \epsilon + \gamma\}$, otherwise, call them "strong".
- Intuition: "strong" classifiers dominate in the final weight
- Fix (x, y), weight of each group of classifiers:

$$W_s^{\checkmark}(x,y) = \frac{\sum_{h(x)=y,\epsilon(h)<\epsilon+\gamma} e^{-\eta\epsilon(h)}}{\sum_{\mathcal{H}} e^{-\eta\epsilon(h)}}, W_s^{\text{X}}(x,y) = \frac{\sum_{h(x)\neq y,\epsilon(h)<\epsilon+\gamma} e^{-\eta\epsilon(h)}}{\sum_{\mathcal{H}} e^{-\eta\epsilon(h)}}$$

$$W_w = \frac{\sum_{\epsilon(h)\geq\epsilon+\gamma} e^{-\eta\epsilon(h)}}{\sum_{\mathcal{H}} e^{-\eta\epsilon(h)}} \leq \frac{|\mathcal{H}|e^{-\eta(\gamma+\epsilon)}}{e^{-\eta\gamma}} = |\mathcal{H}|e^{-\eta\gamma} \leq \frac{1}{8}$$

Proof II

- When $y\ell(x) \leq 2\Delta$, restricting our scope in "strong" classifiers:

$$2\Delta \geq \frac{1}{\eta} \ln \frac{W^{\checkmark}(x,y)}{W^{\text{X}}(x,y)} \geq \frac{1}{\eta} \ln \frac{W_s^{\checkmark}(x,y)}{W_s^{\text{X}}(x,y) + W_w}$$

Hence

$$W_s^{\text{X}}(x,y) + W_w \geq \frac{1}{1 + e^{2\Delta\eta}} =: c \Rightarrow \frac{W_s^{\text{X}}(x,y)}{W_s^{\text{X}}(x,y) + W_s^{\checkmark}(x,y)} \geq \frac{c - W_w}{1 - W_w}$$

On this event, there are a constant fraction of "strong" classifiers making mistake, but since they have small error, this cannot happen often.

Proof III

- Following the reasoning, we have

$$
\begin{aligned}
\Pr(y\ell(x) \leq 2\Delta) &\leq \Pr\left(\frac{W_s^{\text{X}}(x,y)}{W_s^{\text{X}}(x,y) + W_s^{\checkmark}(x,y)} \geq \frac{c - W_w}{1 - W_w}\right) \\
&\leq \Pr_{(x,y)\sim\mathcal{D}}\left(\Pr_{h\sim w|s}(h(x) \neq y) \geq \frac{c - W_w}{1 - W_w}\right) \\
&\leq \mathbb{E}_{(x,y)\sim\mathcal{D}} \Pr_{h\sim w|s}(h(x) \neq y)\frac{1 - W_w}{c - W_w} \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_{h\sim w|s} I(h(x) \neq y)\frac{1 - W_w}{c - W_w} \\
&= \mathbb{E}_{h\sim w|s}\mathbb{E}_{(x,y)\sim\mathcal{D}} I(h(x) \neq y)\frac{1 - W_w}{c - W_w} \\
&\leq (\epsilon + \gamma)\frac{1 - W_w}{c - W_w} \\
&\leq (\epsilon + \gamma)(1 + 2W_w e^{2\Delta\eta})(1 + e^{2\Delta\eta}) \\
&\leq (1 + e^{2\Delta\eta})(1 + 2|\mathcal{H}|e^{\eta(2\Delta-\gamma)})(\epsilon + \gamma)
\end{aligned}
$$

A (somewhat) Simplified Treatment

Theorem

For any $\Delta \geq 0$, we have:

$$\Pr(y\ell(x) \leq 2\Delta) \leq (1 + e^{2\eta\Delta})(\epsilon + \frac{\ln|\mathcal{H}|}{\eta})$$

In particular, let $\Delta = 0$, we have:

$$\Pr(y\ell(x) \leq 0) \leq 2(\epsilon + \frac{\ln|\mathcal{H}|}{\eta})$$

Eliminates the technical conditons in the last theorem statement

Proof I

- Regardless of "weak" or "strong"(non-intuitive)
- Fix an example (x, y), "correct" and "incorrect" weight

$$W^{\checkmark}(x,y) = \frac{\sum_{h(x)=y} e^{-\eta\epsilon(h)}}{\sum_{\mathcal{H}} e^{-\eta\epsilon(h)}}, W^{\mathsf{X}}(x,y) = \frac{\sum_{h(x)\neq y} e^{-\eta\epsilon(h)}}{\sum_{\mathcal{H}} e^{-\eta\epsilon(h)}}$$

Markov's Inequality:

$$\begin{aligned}
& I(y\ell(x) \leq 2\Delta) \\
= \;& I(W^{\checkmark}(x,y) \leq e^{2\eta\Delta} W^{\mathsf{X}}(x,y)) \\
= \;& I(W^{\mathsf{X}}(x,y) \geq \frac{1}{1+e^{2\eta\Delta}}) \\
\leq \;& (1+e^{2\eta\Delta}) W^{\mathsf{X}}(x,y) \\
= \;& (1+e^{2\eta\Delta}) \frac{\sum_{h(x)\neq y} e^{-\eta\epsilon(h)}}{\sum_{h} e^{-\eta\epsilon(h)}}
\end{aligned}$$

Proof II

Taking expectations on both sides:

$$Pr(y\ell(x) \leq 2\Delta) \leq (1 + e^{2\eta\Delta})\frac{\sum_h \epsilon(h)e^{-\eta\epsilon(h)}}{\sum_h e^{-\eta\epsilon(h)}}$$

i.e. the (margin)error of voting classifier is related to that of its corrseponding stochastic classifier.

- Convexity of $x \ln x, x > 0$:

$$\mathbb{E}X \ln X \geq \mathbb{E}X \ln(\mathbb{E}X)$$

- Take $X(h) = e^{-\eta\epsilon(h)}$, uniform distribution over $\mathcal{H}$:

$$\frac{\sum_h \epsilon(h)e^{-\eta\epsilon(h)}}{\sum_h e^{-\eta\epsilon(h)}} \leq -\frac{1}{\eta}\ln(\frac{1}{|\mathcal{H}|}\sum_h e^{-\eta\epsilon(h)})$$

- Familiar "singleton" bound: $\leq \epsilon + \frac{\ln|\mathcal{H}|}{\eta}$.

Generalization to uncountably infinite hypothesis class I

- Have a "prior" μ (hopefully) puts more weights for "good" classifiers
- define $\ell(x)$ slightly differently: $\ell(x) = \frac{\int_{h(x)=+} e^{-\eta\epsilon(h)}\mathrm{d}\mu}{\int_{h(x)=-} e^{-\eta\epsilon(h)}\mathrm{d}\mu}$. Then same argument goes:

$$Pr(y\ell(x) \leq 2\Delta) \leq (1 + e^{2\eta\Delta})(-\frac{1}{\eta} \ln \int e^{-\eta\epsilon(h)}\mathrm{d}\mu(h))$$

- Applying Compression Lemma:

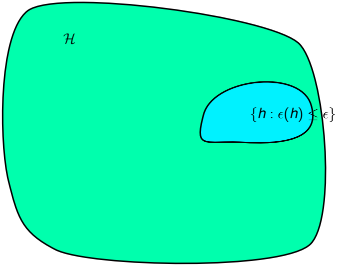
$$-\frac{1}{\eta} \ln \int e^{-\eta\epsilon(h)}\mathrm{d}\mu(h) \leq \int \epsilon(h)\mathrm{d}\nu(h) + \frac{D(\nu||\mu)}{\eta}$$

Where $D(\nu||\mu) = \int \ln \frac{\mathrm{d}\nu}{\mathrm{d}\mu}\mathrm{d}\nu$ is the relative entropy.

Generalization to uncountably infinite hypothesis class II

- Taking $V_\epsilon = \int_{\epsilon(h)\leq\epsilon} \mathrm{d}\mu(h)$, $\mathrm{d}\nu(h) = \frac{I(\epsilon(h)\leq\epsilon)}{V_\epsilon}\mathrm{d}\mu(h)$:

$$-\frac{1}{\eta}\ln\int e^{-\eta\epsilon(h)}\mathrm{d}\mu(h) \leq \epsilon + \frac{\ln 1/|V_\epsilon|}{\eta}$$



$\hat{\ell}(x)$ converges to $\ell(x)$

Theorem

For any $\mathcal{D}$, any $x \in \mathcal{X}$, any $\lambda, \eta > 0$:

$$\Pr_{S \sim \mathcal{D}^m}(|\ell(x) - \hat{\ell}(x)| \geq 2\lambda + \frac{\eta}{8m}) \leq 4e^{-2m\lambda^2}$$

- Intuitively, $\eta \uparrow$, the convergence become worse(in contrast to performance of $\ell(x)$).
- Note the convergence rate does not depend on $|\mathcal{H}|$!
- Define

$$\hat{R}_\eta(\mathcal{K}) = \frac{1}{\eta}\ln(\sum_{h \in \mathcal{K}} e^{-\eta\hat{\epsilon}(h)}), R_\eta(\mathcal{K}) = \frac{1}{\eta}\ln(\sum_{h \in \mathcal{K}} e^{-\eta\epsilon(h)})$$

- Note $\hat{\ell}(x) = \hat{R}_\eta(\mathcal{H}_x^+) - \hat{R}_\eta(\mathcal{H}_x^-)$, $\ell(x) = R_\eta(\mathcal{H}_x^+) - R_\eta(\mathcal{H}_x^-)$
- We will prove it based on covergence of $\hat{R}_\eta(\mathcal{H}_x^+)$ to $R_\eta(\mathcal{H}_x^+)$, and $\hat{R}_\eta(\mathcal{H}_x^-)$ to $R_\eta(\mathcal{H}_x^-)$

Proof (1): $\hat{R}_\eta(\mathcal{K})$ converges to $\mathbb{E}\hat{R}_\eta(\mathcal{K})$

- Note

$$\hat{R}_\eta(\mathcal{K})((x_1, y_1), \ldots, (x_m, y_m)) = \frac{1}{\eta} \ln\left(\sum_{h \in \mathcal{K}} e^{-\eta \hat{\epsilon}(h)}\right)$$

  satisfies bounded difference
- Suppose we have modified training set $S' = (S \setminus \{(x_i, y_1)\}) \cup \{(x_i', y_i')\}$. denote $\hat{\epsilon}'(h)$ the empirical error of h in S'. Then $\sup_{h \in \mathcal{K}} |\hat{\epsilon}'(h) - \hat{\epsilon}(h)| \leq \frac{1}{m}$.

$$-\frac{1}{m} \leq \frac{1}{\eta} \inf_{h \in \mathcal{K}} \ln\left(\frac{e^{-\eta \hat{\epsilon}(h)}}{e^{-\eta \hat{\epsilon}'(h)}}\right) \leq \frac{1}{\eta} \ln\left(\frac{\sum_{h \in \mathcal{K}} e^{-\eta \hat{\epsilon}(h)}}{\sum_{h \in \mathcal{K}} e^{-\eta \hat{\epsilon}'(h)}}\right) \leq \frac{1}{\eta} \sup_{h \in \mathcal{K}} \ln\left(\frac{e^{-\eta \hat{\epsilon}(h)}}{e^{-\eta \hat{\epsilon}'(h)}}\right) \leq \frac{1}{m}$$

- By McDiarmid's Lemma, w.p. $1 - 2e^{-2m\lambda^2}$

$$|\hat{R}_\eta(\mathcal{K}) - \mathbb{E}\hat{R}_\eta(\mathcal{K})| \leq \lambda$$

- How does $\mathbb{E}\hat{R}_\eta(\mathcal{K})$ relate to $R_\eta(\mathcal{K})$?

Proof (2): $\mathbb{E}\hat{R}_\eta(\mathcal{K})$ converges to $R_\eta(\mathcal{K})$ I

Lemma

$$R_\eta(\mathcal{K}) \le \mathbb{E}\hat{R}_\eta(\mathcal{K}) \le R_\eta(\mathcal{K}) + \frac{\eta}{8m}$$

- The first inequality directly follows from convexity of $f(x) = \ln \sum_i e^{x_i}$

Proof (2): $\mathbb{E}\hat{R}_\eta(\mathcal{K})$ converges to $R_\eta(\mathcal{K})$ II

- The second uses Hoeffding's Inequality:
  $X \in [a, b] \Rightarrow \mathbb{E}e^X \le e^{\mathbb{E}X} e^{(b-a)^2/8}$.

$$
\begin{aligned}
\mathbb{E}\hat{R}_\eta(\mathcal{K}) &= \mathbb{E}\frac{1}{\eta}\ln\Big(\sum_{h\in\mathcal{K}} e^{-\eta\hat{\epsilon}(h)}\Big) \\
&\le \frac{1}{\eta}\ln\Big(\sum_{h\in\mathcal{K}} \mathbb{E}e^{-\eta\hat{\epsilon}(h)}\Big) \\
&= \frac{1}{\eta}\ln\Big(\sum_{h\in\mathcal{K}} \big(\mathbb{E}e^{-\frac{\eta}{m}I(h(x_i)\neq y_i))}\big)^m\Big) \\
&\le \frac{1}{\eta}\ln\Big(\sum_{h\in\mathcal{K}} \big(e^{-\frac{\eta}{m}\epsilon(h)} e^{\frac{\eta^2}{8m^2}}\big)^m\Big) \\
&\le \frac{1}{\eta}\ln\Big(\sum_{h\in\mathcal{K}} e^{-\eta\epsilon(h)}\Big) + \frac{\eta}{8m}
\end{aligned}
$$

Proof (3): Combine $\mathcal{H}_x^+$ and $\mathcal{H}_x^-$

- w.p. $1 - 4e^{-2m\lambda^2}$ the following hold simultaneously:

$$\hat{R}_\eta(\mathcal{H}_x^+) \leq \mathbb{E}\hat{R}_\eta(\mathcal{H}_x^+) + \lambda \leq R_\eta(\mathcal{H}_x^+) + \lambda + \frac{\eta}{8m}$$

$$\hat{R}_\eta(\mathcal{H}_x^-) \geq \mathbb{E}\hat{R}_\eta(\mathcal{H}_x^+) - \lambda \geq R_\eta(\mathcal{H}_x^+) - \lambda$$

- Hence $\hat{\ell}(x) \leq \ell(x) + 2\lambda + \frac{\eta}{8m}$
- Analogously, $-\hat{\ell}(x) \leq -\ell(x) + 2\lambda + \frac{\eta}{8m}$
- Proof generalized into uncountably infinite $\mathcal{H}$, with technicalities resolved in paper

Bounding the fraction of "atypical" test examples I

Theorem

For any $\delta > 0$ and $\eta > 0$, if we set $\Delta = 2\sqrt{\frac{\ln(2/\delta)}{m}} + \frac{\eta}{8m}$, then w.p. $1 - \delta$ over choice of S,

$$\Pr_{(x,y)\sim\mathcal{D}}(|\ell(x) - \hat{\ell}(x)| \geq \Delta) \leq \delta$$

Note that setting e.g. $\delta = O(m^{-10})$ won't affect much of the bound

Proof.

Bounding the fraction of "atypical" test examples II

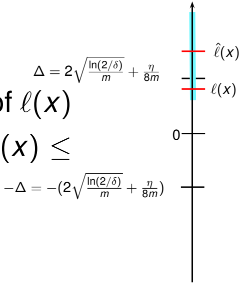
Taking $\lambda = \sqrt{\frac{\ln(2/\delta)}{m}}$ using the previous theorem,

$$
\begin{aligned}
& \delta^2 \\
\geq \quad & \mathbb{E}_{(x,y)\sim\mathcal{D}} \Pr_{S\sim\mathcal{D}^m}(|\ell(x) - \hat{\ell}(x)| \geq \Delta) \\
= \quad & \mathbb{E}_{(x,y)\sim\mathcal{D}} \mathbb{E}_{S\sim\mathcal{D}^m} I(|\ell(x) - \hat{\ell}(x)| \geq \Delta) \\
\geq \quad & \mathbb{E}_{S\sim\mathcal{D}^m} \mathbb{E}_{(x,y)\sim\mathcal{D}} I(|\ell(x) - \hat{\ell}(x)| \geq \Delta) \\
\geq \quad & \mathbb{E}_{S\sim\mathcal{D}^m} \Pr_{(x,y)\sim\mathcal{D}}(|\ell(x) - \hat{\ell}(x)| \geq \Delta)
\end{aligned}
$$

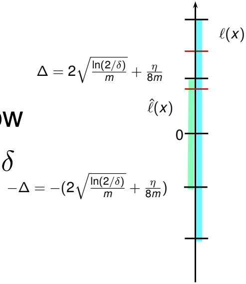
The theorem follows by Markov's Inequality. $\square$

Implication of Deviation

- Relate the error of $\hat{\ell}(x)$ to the error of $\ell(x)$
- $\Pr(|\hat{\ell}(x)| > \Delta \wedge y\hat{\ell}(x) \leq 0) \leq \Pr(y\ell(x) \leq 0) + \delta$



- The probability of s aying Don't Know
- $\Pr(|\hat{\ell}(x)| \leq \Delta) \leq \Pr(|\ell(x)| \leq 2\Delta) + \delta$



Putting them together

- Mistake Bound

$$\begin{aligned}\Pr(\text{Mistake}) &= \Pr(|\hat{\ell}(x)| > \Delta \wedge y\hat{\ell}(x) \le 0) \\ &\le \Pr(y\ell(x) \le 0) + \delta \\ &\le 2(\epsilon + \frac{\ln|\mathcal{H}|}{\eta}) + \delta\end{aligned}$$

- Don't know Bound

$$\begin{aligned}\Pr(\text{Don't Know}) &= \Pr(|\hat{\ell}(x)| \le \Delta) \\ &\le \Pr(|\ell(x)| \le 2\Delta) + \delta \\ &\le \Pr(y\ell(x) \le 2\Delta) + \delta \\ &\le (1 + e^{2\Delta\eta})(\epsilon + \frac{\ln|\mathcal{H}|}{\eta}) + \delta\end{aligned}$$

Putting them together

- Simple tuning($\eta = \ln|\mathcal{H}|m^{1/2}$):

$$\text{Pr(Mistake)} \leq 2(\epsilon + m^{-1/2}) + \delta$$

$$\text{Pr(Don't Know)} \leq e^{\sqrt{\ln\frac{1}{\delta}\ln|\mathcal{H}| + (\ln|\mathcal{H}|)^2}}(\epsilon + m^{-1/2}) + \delta$$

- In region of prediction, the probability of making a mistake is independent of complexity of $\mathcal{H}$ any more!
- The upper bound of probability of saying Don't Know might be loose; should be estimated by unlabelled data(or test data in tranductive setting) in practice.

Putting them together(2)

- A subtler tuning($\eta = \ln|\mathcal{H}|m^{1/2-\theta}$):

$$\Pr(\text{Mistake}) \le 2(\epsilon + m^{-1/2+\theta}) + \delta$$

$$\Pr(\text{Don't Know}) \le (1 + e^{2\frac{\sqrt{\ln 1/\delta}\ln|\mathcal{H}|}{m^\theta} + \frac{(\ln|\mathcal{H}|)^2}{m^{2\theta}}})(\epsilon + m^{-1/2+\theta}) + \delta$$

- When $m \le O((\ln|\mathcal{H}| + \ln(1/\delta))^{1/\theta})$, the mistake bound improves over ERM.
- OTOH, $m \ge \Omega((\ln|\mathcal{H}|\ln(1/\delta))^{1/\theta})$,

$$2\frac{\sqrt{\ln 1/\delta}\ln|\mathcal{H}|}{m^\theta} + \frac{(\ln|\mathcal{H}|)^2}{m^{2\theta}} \le 1$$

$\Pr(\text{Don't Know}) \le 5(\epsilon + m^{-1/2+\theta})$, almost as small as the error guarantee for ERM.

Aside: Proof of Compression Lemma

Lemma

If μ, ν are two probability measures, $\nu \ll \mu$, then

$$\int f(h)\mathrm{d}\nu(h) \leq D(\nu||\mu) + \ln(\int e^{f(h)}\mathrm{d}\mu(h))$$

Proof.

Define a new probability measure $\mathrm{d}\hat{\mu}(h) = e^{f(h)}\mathrm{d}\mu(h)/Z$, $Z = \int e^{f(h)}\mathrm{d}\mu(h)$. Since $D(\nu||\hat{\mu}) \geq 0$, expanding,

$$\int \ln(\frac{\mathrm{d}\nu Z}{\mathrm{d}\mu e^f(h)})\mathrm{d}\nu \geq 0$$

i.e.

$$\ln Z + D(\nu||\mu) \geq \int f(h)\mathrm{d}\nu(h)$$

$\square$

Section 2

Concentration Bounds for Sequential Prediction

Summary

- Online Learning with Stochastic data
- A general martingale technique for analysis
- Small mistake bounds $\Rightarrow$ good generalization
- Analysis of Exponential Weight Algorithm's generalization error

Online Learning with Stochastic Data I

- Perceptron Algorithm:
- $w_1 = 0$;
  For $t = 1, 2, \ldots, m$:
  Observing x_t, predicts $\hat{y}_t = \mathrm{sign}(w_t \cdot x_t) =: h_t(x_t)$.
  Receive y_t, incur loss $I(y_t \neq \mathrm{sign}(w_t \cdot x_t))$.
  Update w_{t+1} based on w_t, (x_t, y_t).
- Mistake Bound: Suppose $||X||_2 \leq X$, Hinge Loss of u:

$$L_{m,\gamma}(u) = \sum_{t=1}^{m} (\gamma - y_t u^T x_t)_+$$

$$\begin{aligned} & \sum_{t=1}^{m} I(y_t \neq \mathrm{sign}(w_t \cdot x_t)) \\ \leq \quad & \inf_{u, \gamma > 0} \left(\frac{L_{m,\gamma}(u)}{\gamma} + \frac{X^2 ||u||^2}{\gamma^2} + \sqrt{\frac{L_{m,\gamma}(u)}{\gamma} \frac{X^2 ||u||^2}{\gamma^2}} \right) \end{aligned}$$

Online Learning with Stochastic Data II

- it holds for arbitrary sequence, hence for stochastic sequence as well
- can we relate $\{\epsilon(h_t)\}_{t=1}^m$ to $\{I(h_t(x_t) \neq y_t)\}_{t=1}^m$? (Online to Batch Conversion) Also answered in [CBG05, CBCG04].

A Basic Inequality I

- Consider a sequence of iid random variables $Z_1, \ldots, Z_m$, and functions $\xi_1(z_1), \xi_2(z_1, z_2), \ldots, \xi_m(z_1, z_2, \ldots, z_m)$.
- Specifically in our context: $z_t = (x_t, y_t)$, $\xi_t(z_1, z_2, \ldots, z_t) = I(h_t(x_t) \neq y_t)$, h_t depends only on $(x_1, y_1), \ldots, (x_{t-1}, y_{t-1})$
- Notice $\mathbb{E}_{Z_t}\xi_t(Z_1, Z_2, \ldots, Z_t) = \epsilon(h_t)$. Find the relationship between $\{\mathbb{E}_{Z_t}\xi_t\}_{t=1}^m$ and $\{\xi_t\}_{t=1}^m$. Specifically, $\mu_m = \frac{1}{m}\sum_{t=1}^m \mathbb{E}_{Z_t}\xi_t$, $s_m = \frac{1}{m}\sum_{t=1}^m \xi_t$

A Basic Inequality II

Lemma

For any functions $\zeta_1(x_1), \ldots, \zeta_m(x_1, \ldots, x_m)$

$$\left\{\frac{e^{\zeta_1(Z_1)}}{\mathbb{E}_{Z_1} e^{\zeta_1(Z_1)}} \frac{e^{\zeta_2(Z_1,Z_2)}}{\mathbb{E}_{Z_2} e^{\zeta_2(Z_1,Z_2)}} \cdots \frac{e^{\zeta_t(Z_1,\ldots,Z_t)}}{\mathbb{E}_{Z_t} e^{\zeta_t(Z_1,\ldots,Z_t)}}\right\}_{t=1}^{m}$$

is a martingale. Hence

$$\mathbb{E}\frac{e^{\zeta_1(Z_1)}}{\mathbb{E}_{Z_1} e^{\zeta_1(Z_1)}} \frac{e^{\zeta_2(Z_1,Z_2)}}{\mathbb{E}_{Z_2} e^{\zeta_2(Z_1,Z_2)}} \cdots \frac{e^{\zeta_m(Z_1,\ldots,Z_m)}}{\mathbb{E}_{Z_m} e^{\zeta_m(Z_1,\ldots,Z_m)}} = 1$$

By Markov's Inequality, w.p. $1-\delta$,

$$\begin{aligned} & \zeta_1(Z_1) + \zeta_2(Z_1, Z_2) + \ldots + \zeta_m(Z_1, \ldots, Z_m) \\ \leq \quad & \ln \mathbb{E}_{Z_1} e^{\zeta_1(Z_1)} + \ln \mathbb{E}_{Z_2} e^{\zeta_2(Z_1,Z_2)} + \ldots + \ln \mathbb{E}_{Z_m} e^{\zeta_m(Z_1,\ldots,Z_m)} + \ln \frac{1}{\delta} \end{aligned}$$

Relative Entropy Inequalities I

- Recall: how do we prove Chernoff bound?

$$\Pr\left(\sum_t X_t \le m(p-\epsilon)\right) \le e^{-mD(p-\epsilon||p)}$$

- Assume $\xi_t \in [0,1]$
- Taking $\zeta_t = -\rho\xi_t$, rearranging,

$$\begin{aligned}
& \rho\xi_1 + \rho\xi_2 + \ldots + \rho\xi_m + \ln\frac{1}{\delta} \\
\ge\ & -\ln \mathbb{E}_{Z_1} e^{-\rho\xi_1} - \ln \mathbb{E}_{Z_2} e^{-\rho\xi_2} - \ldots - \ln \mathbb{E}_{Z_n} e^{-\rho\xi_m} \\
\ge\ & \sum_{t=1}^{m} -\ln(1-(1-e^{-\rho})\mathbb{E}_{Z_t}\xi_t) \\
\ge\ & -m\ln(1-(1-e^{-\rho})\mu_m))
\end{aligned}$$

Equivalent to $-\rho s_m - \ln(1-(1-e^{-\rho})\mu_m) \le \frac{\ln 1/\delta}{m}$

Relative Entropy Inequalities II

- $D(q||p) = \sup_{\rho \geq 0}(-\rho q - \ln(1-(1-e^{-\rho})q)), p \geq q,$
- But: ρ cannot be tuned apriori!
- after some manipulations: $\forall \alpha \in [0,1], t \geq 0,$

$$\Pr(\mu_m \geq D^{-1}(\alpha, \frac{\ln(1/\delta)}{m}), s_m \leq \alpha) \leq \delta$$

  where $D^{-1}(\alpha, \frac{\ln(1/\delta)}{m}) = \inf\{\beta \geq \alpha : D(\alpha||\beta) \geq \frac{\ln(1/\delta)}{m}\}$
  (Reason: Take $\rho_0 = -\ln \frac{\alpha(1-D^{-1}(\alpha, \ln(1/\delta)/m))}{D^{-1}(\alpha, \ln(1/\delta)/m)(1-\alpha)}$, then
  $-\rho_0\alpha - \ln(1-(1-e^{-\rho_0})D^{-1}(\alpha, \ln(1/\delta)/m)) = \ln(1/\delta)/m$,
  so $-\rho_0 s_m - \ln(1-(1-e^{-\rho_0})\mu_m) \geq \ln(1/\delta)/m$, this happens w.p. $\leq \delta$)
- Taking union bound over all $(\alpha, \delta) \in \{(0, \frac{\delta_0}{2^2}), (\frac{1}{m}, \frac{\delta_0}{3^2}), \ldots, (1, \frac{\delta_0}{(m+2)^2})\}$:

$$\Pr(\mu_m \geq D^{-1}(\frac{\lceil ms_m \rceil}{m}, \frac{2\ln(\lceil ms_m \rceil + 2)}{m} + \frac{\ln(1/\delta_0)}{m})) \leq \delta_0$$

Relative Entropy Inequalities III

- Since $D^{-1}(\alpha, \ln(1/\delta)/m) \leq \alpha + 2\ln(1/\delta)/m + \sqrt{2\alpha\ln(1/\delta)/m}$, we get w.p. $1-\delta$

$$\mu_m \leq \frac{\lceil ms_m \rceil}{m} + 2\frac{2\ln(\lceil ms_m \rceil + 2) + \ln\frac{1}{\delta}}{m} + \sqrt{2(\frac{\lceil ms_m \rceil}{m}\frac{2\ln(\lceil ms_m \rceil + 2) + \ln\frac{1}{\delta}}{m}}$$

- Simplifying:

$$\mu_m \leq s_m + O(\frac{\ln m + \ln(1/\delta)}{m} + \sqrt{s_m\frac{\ln m + \ln(1/\delta)}{m}})$$

- Specifically for perceptron:

$$\mu_m \leq \frac{L_{\gamma,m}(u)}{m\gamma} + O(\frac{X^2||u||^2}{m\gamma^2} + \frac{\ln m + \ln(1/\delta)}{m} + \sqrt{\frac{L_{\gamma,m}(u)}{m\gamma}\frac{X^2||u||^2}{m\gamma^2} + \frac{\ln m + \ln(1/\delta)}{m}})$$

Bennett Inequalities I

- Assume $\xi_t - \mathbb{E}_{Z_t}\xi_t \geq -1$
- Taking $\zeta_t = -\rho(\xi_t - \mathbb{E}_{Z_t}\xi_t)$, rearranging, since

$$\mathbb{E}_{Z_t} e^{-\rho(\xi_t - \mathbb{E}_{Z_t}\xi_t)} \leq 1 + \mathbb{E}_{Z_t}(\xi_t - \mathbb{E}_{Z_t}\xi_t)^2(e^\rho - \rho - 1)$$

$$\begin{aligned}
& \rho(\xi_1 - \mathbb{E}_{Z_1}\xi_1) + \rho(\xi_2 - \mathbb{E}_{Z_2}\xi_2) + \ldots + \rho(\xi_m - \mathbb{E}_{Z_m}\xi_m) + \ln\frac{1}{\delta} \\
\geq\ & -\ln \mathbb{E}_{Z_1} e^{-\rho(\xi_1 - \mathbb{E}_{Z_1}\xi_1)} - \ldots - \ln \mathbb{E}_{Z_m} e^{-\rho(\xi_m - \mathbb{E}_{Z_m}\xi_m)} \\
\geq\ & -(e^\rho - \rho - 1)(\mathbb{E}_{Z_1}(\xi_1 - \mathbb{E}_{Z_1}\xi_1)^2 + \ldots + \mathbb{E}_{Z_m}(\xi_m - \mathbb{E}_{Z_m}\xi_m)^2)
\end{aligned}$$

Bennett Inequalities II

- Imposing the assumption that $\mathbb{E}_{Z_t}(\xi_t - \mathbb{E}_{Z_t}\xi_t)^2 \leq b\mathbb{E}_{Z_t}\xi_t$(e.g. $\mathbb{E}_{(x_t,y_t)}(I(h(x_t) \neq y_t) - \epsilon(h_t))^2 \leq \epsilon(h_t) - \epsilon(h_t)^2 \Rightarrow b = 1$)

$$
\begin{aligned}
& (\mathbb{E}_{Z_1}\xi_1 + \mathbb{E}_{Z_2}\xi_2 + \ldots + \mathbb{E}_{Z_m}\xi_m) - (\xi_1 + \xi_2 + \ldots + \xi_m) \\
\leq \quad & \frac{\ln(1/\delta)}{\rho} + \frac{\rho}{2(1-\rho/3)}(\mathbb{E}_{Z_1}(\xi_1 - \mathbb{E}_{Z_1}\xi_1)^2 + \ldots + \mathbb{E}_{Z_m}(\xi_m - \mathbb{E}_{Z_m}\xi_m)^2 \\
\leq \quad & \frac{\ln(1/\delta)}{\rho} + \frac{b\rho}{2(1-\rho/3)}(\mathbb{E}_{Z_1}\xi_1 + \mathbb{E}_{Z_2}\xi_2 + \ldots + \mathbb{E}_{Z_m}\xi_m)
\end{aligned}
$$

- after some manipulations, $\exists c_b$, $\forall \alpha > 0$,

$$
\Pr(\mu_n \geq \alpha + \sqrt{2b\alpha\frac{\ln(1/\delta)}{m}} + c_b\frac{\ln(1/\delta)}{m}, s_m \leq \alpha) \leq \delta
$$

Bennett Inequalities III

- Same as before, taking union bound over all $(\alpha, \delta) \in \{(0, \frac{\delta_0}{2^2}), (\frac{1}{m}, \frac{\delta_0}{3^2}), \ldots, (1, \frac{\delta_0}{(m+2)^2})\}$, w.p. $1 - \delta_0$:

$$\mu_m \leq s_m + c_b \frac{2\ln(\lceil ms_m \rceil + 2) + \ln\frac{1}{\delta_0}}{m} + \sqrt{2b(\frac{\lceil ms_m \rceil}{m}\frac{2\ln(\lceil ms_m \rceil + 2) + \ln\frac{1}{\delta_0}}{m}}$$

- Similar bounds obtained for perceptron

Applications to Exponential Weight Algorithm I

- Reminder:

$$
\begin{aligned}
& \rho\xi_1 + \rho\xi_2 + \ldots + \rho\xi_m + \ln\frac{1}{\delta} \\
\geq \quad & -\ln \mathbb{E}_{Z_1} e^{-\rho\xi_1} - \ln \mathbb{E}_{Z_2} e^{-\rho\xi_2} - \ldots - \ln \mathbb{E}_{Z_m} e^{-\rho\xi_m}
\end{aligned}
$$

- Consider applying Hedge(η) to iid sequence $\{(x_t, y_t)\}_{t=1}^m$, Denote $\mu_m(\eta) = \frac{1}{m}\sum_{t=1}^m \mathbb{E}_{h\sim w_t}\epsilon(h_t)$,$s_m(\eta) = \frac{1}{m}\sum_{t=1}^m \mathbb{E}_{h\sim w_t} I(h_t(x_t) \neq y_t)$
- Let $\xi_t = -\ln \mathbb{E}_{h\sim w_t} e^{-\eta I(h(x_t)\neq y_t)}$, $\rho = 1$:

$$
\begin{aligned}
& -\ln(\mathbb{E}_{h\sim w_1} e^{-\eta n\hat{\epsilon}(h)}) + \ln(1/\delta) \\
\geq \quad & -\ln(1-(1-e^{-\eta})\mathbb{E}_{h\sim w_1}\epsilon(h)) - \ldots - \ln(1-(1-e^{-\eta})\mathbb{E}_{h\sim w_m}\epsilon(h)) \\
\geq \quad & -m\ln(1-(1-e^{-\eta})\mu_m(\eta)))
\end{aligned}
$$

Applications to Exponential Weight Algorithm II

Hence

$$\mu_m(\eta) \leq \frac{\eta\hat{\epsilon}(\hat{h}) + \ln(|\mathcal{H}|/\delta)/m}{1 - e^{-\eta}}$$

Tuning η with hindsight, again applying union bound(details omitted)

$$\mu_m(\eta) \leq \hat{\epsilon}(\hat{h}) + 2\frac{\ln(\frac{|\mathcal{H}|m}{\delta})}{m} + \sqrt{2\hat{\epsilon}(\hat{h})\frac{\ln(\frac{|\mathcal{H}|m}{\delta})}{m}}$$

- Recall Hedge proof:

$$\begin{aligned} & -\ln(\mathbb{E}_{h\sim w_1} e^{-\eta m\hat{\epsilon}(h)}) \\ = \quad & -\ln(1 - (1 - e^{-\eta})\mathbb{E}_{h\sim w_1} I(h(x_1) \neq y_1)) - \dots \\ & -\ln(1 - (1 - e^{-\eta})\mathbb{E}_{h\sim w_m} I(h(x_m) \neq y_m)) \\ \geq \quad & -m\ln(1 - (1 - e^{-\eta})s_m(\eta))) \end{aligned}$$

Applications to Exponential Weight Algorithm III

We have

$$s_m(\eta) \leq \frac{\eta\hat{\epsilon}(\hat{h}) + \ln(|\mathcal{H}|)/m}{1 - e^{-\eta}}$$

References I

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile.
On the generalization ability of on-line learning algorithms.
IEEE Transactions on Information Theory, 50(9):2050–2057, 2004.

Nicolò Cesa-Bianchi and Claudio Gentile.
Improved risk tail bounds for on-line algorithms.
In *NIPS*, 2005.

R. El-Yaniv and Y. Wiener.
Agnostic selective classification.
In *NIPS*, 2011.

Yoav Freund, Yishay Mansour, and Robert E. Schapire.
Generalization bounds for averaged classifiers.
The Annals of Statistics, 32(4):1698–1722, 08 2004.

References II

Tong Zhang.
Data dependent concentration bounds for sequential prediction algorithms.
In *COLT*, pages 173–187, 2005.