E0 370 Statistical Learning Theory **Lecture 20 (Oct 24, 2013)**

Online-to-Batch Conversions

Lecturer: *Shivani Agarwal* Scribe: *Y. R. Siddartha*

1 Introduction

In Lecture 17, we saw regret bounds for online algorithms for classification and regression. These bounds hold any sequence of examples, including a worst-case sequence drawn by an adversary who seeks to maximise the regret.

In this lecture, we study online-to-batch conversion schemes that apply online algorithms in a batch setting, where the sequence of examples are assumed to be drawn *i.i.d.* from a joint distribution, and show that the online regret bounds can be used to obtain generalization error bounds and statistical consistency results for the resulting batch algorithms.

Let $\mathcal{X}, \mathcal{Y}, \hat{\mathcal{Y}}$ denote the instance, label and prediction spaces, D a distribution on $\mathcal{X} \times \mathcal{Y}$, $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$ a loss function and $S = ((x_1, y_1), \ldots, (x_m, y_m))$ the training set consisting of m examples drawn *i.i.d.* from D. The goal for a supervised learner is to learn a model $h_S : \mathcal{X} \to \hat{\mathcal{Y}}$ that minimizes the generalization error $\mathrm{er}_D^\ell[h_S] = \mathbf{E}_{(x,y)\sim D}[\ell(y, h_S(x))]$.

If the training examples in S are input sequentially to a given online learning algorithm $\mathcal{A}$, the latter constructs a sequence of models $\{h_t : \mathcal{X} \to \hat{\mathcal{Y}},\ t = 1, \ldots, T+1\}$ and returns predictions based on them.

Online supervised learning

Initialize model $h_1 : \mathcal{X} \to \hat{\mathcal{Y}}$
for $t = 1, 2, \ldots, T$ **do**
  Receive $x_t \in \mathcal{X}$
  Predict $\hat{y}_t = h_t(x_t)$
  Receive $y_t \in \mathcal{Y}$
  Incur loss $\ell(y_t, \hat{y}_t)$
  Update model to h_{t+1}
end for

The performance of the online learner is measured by its cumulative loss $L_S^\ell[\mathcal{A}] = \sum_{t=1}^m \ell(y_t, \hat{y}_t)$, relative to the cumulative loss of the best model in a comparator class $\mathcal{H}$.

An online-to-batch convertor needs to construct a model $h_S : \mathcal{X} \to \hat{\mathcal{Y}}$ based on the set of models $\{h_t\}_{t=1}^{T+1}$ generated by the online learner. There are several possible ways to do this:

- run the online algorithm for a single pass (so that $T = m$) or for multiple passes (so T is a multiple of m) over the training set and return the final model h_{T+1} (this makes sense, for example, when the perceptron algorithm is run to convergence on linearly separable data, but it's not clear if this is a good approach in general.),
- return the longest surviving model among $h_1, \ldots, h_{T+1}$ (the 'pocket' method [Gallant, 1986]),
- return the best model from $h_1, \ldots, h_{T+1}$ based on its empirical performance on a validation or test set [Littlestone,1989],
- randomly choose from $h_1, \ldots, h_{T+1}$ [Helmbold & Warmuth, 1995],
- 'average' over the models in some way [Helmholtz & Warmuth, 1995] (this is also the approach used in the 'voted perceptron' by [Freund and Schapire, 1999]).

Below we focus on the last approach, and see how it can be used to provide generalization/consistency guarantees on the learned model.

2 Online to Batch: Generalization Guarantees

We start with the following lemma, which bounds the *average* generalization errors of the set of models generated by an online algorithm in terms of the observed cumulative loss of the algorithm.

Lemma 2.1 (Cesa-Bianchi *et al*, 2004)**.** Let $h_1, \ldots, h_{m+1} : \mathcal{X} \to \hat{\mathcal{Y}}$ be prediction models generated by an online algorithm $\mathcal{A}$ when run on a training set $S \in (\mathcal{X} \times \mathcal{Y})^m$. Let D be a distribution over $\mathcal{X} \times \mathcal{Y}$, $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$ be a $[0,1]$-bounded loss and $\delta \in (0,1]$. Then *w.p.* $\geq 1-\delta$ over the draws of $S \sim D^m$,

$$\frac{1}{m}\sum_{t=1}^{m} \mathrm{er}_D^{\ell}[h_t] \leq \frac{1}{m} L_S^{\ell}[\mathcal{A}] + \sqrt{\frac{2\ln(1/\delta)}{m}}.$$

Proof. We use the Hoeffding-Azuma concentration inequality for martingales (see Appendix A) to prove this result. To begin with, we construct a martingale sequence as follows: let

$$U_t = \begin{cases} 0, & t = 0; \\ \sum_{i=1}^{t} \left[\mathrm{er}_D^{\ell}[h_i] - \ell(y_i, h_i(x_i))\right], & t \in [m]. \end{cases}$$

Let S_{t-1} denote the partial training sequence $((x_1,y_1),\ldots,(x_{t-1},y_{t-1}))$. We can verify that $\{U_t\}$ is a martingale sequence *w.r.t.* $\{(x_t,y_t)\}$ by showing that $\mathbf{E}_{(x_t,y_t)}[U_t \mid S_{t-1}] = U_{t-1}$ as follows:

$$\begin{aligned}
&\mathbf{E}_{(x_t,y_t)}[U_t \mid S_{t-1}] \\
&= \mathbf{E}_{(x_t,y_t)}\left[\sum_{i=1}^{t}\left[\mathrm{er}_D^{\ell}[h_i] - \ell(y_i,h_i(x_i))\right] \Big| S_{t-1}\right] \\
&= \underbrace{\mathbf{E}_{(x_t,y_t)}\left[\sum_{i=1}^{t-1}\left[\mathrm{er}_D^{\ell}[h_i] - \ell(y_i,h_i(x_i))\right] \Big| S_{t-1}\right]}_{\text{independent of } (x_t,y_t)} + \mathbf{E}_{(x_t,y_t)}\left[\mathrm{er}_D^{\ell}[h_t] - \ell(y_t,h_t(x_t)) \mid S_{t-1}\right] \\
&= \sum_{i=1}^{t-1}\left[\mathrm{er}_D^{\ell}[h_i] - \ell(y_i,h_i(x_i))\right] + \underbrace{\mathbf{E}_{(x_t,y_t)}\left[\mathrm{er}_D^{\ell}[h_t] \mid S_{t-1}\right] - \mathbf{E}_{(x_t,y_t)}[\ell(y_t,h_t(x_t)) \mid S_{t-1}]}_{\text{where } h_t \text{ is a fixed function, determined by } S_{t-1}} \\
&= \sum_{i=1}^{t-1}\left(\mathrm{er}_D^{\ell}[h_i] - \ell(y_i,h_i(x_i))\right) + \mathrm{er}_D^{\ell}[h_t] - \mathrm{er}_D^{\ell}[h_t] \\
&= U_{t-1}.
\end{aligned}$$

We can now define the martingale difference sequence

$$V_t = U_t - U_{t-1} = \mathrm{er}_D^{\ell}[h_t] - \ell(y_t, h_t(x_t))$$

whose terms are bounded by $|V_t| \leq 1$ (because of the assumed $[0,1]$-bound on the loss ℓ). Then the Hoeffding-Azuma inequality gives us the high-probability bound

$$\begin{aligned}
\mathbf{P}\left[\frac{1}{m}\sum_{t=1}^{m}\mathrm{er}_D^{\ell}[h_t] - \frac{1}{m}L_S^{\ell}[\mathcal{A}] \geq \epsilon\right] &= \mathbf{P}\left[\frac{1}{m}\sum_{t=1}^{m}\left(\mathrm{er}_D^{\ell}[h_t] - \ell(y_t,h_t(x_t))\right) \geq \epsilon\right] \\
&= \mathbf{P}\left[\frac{1}{m}U_m \geq \epsilon\right] = \mathbf{P}\left[U_m - U_0 \geq m\epsilon\right] \\
&\leq \exp\left(\frac{-(m\epsilon)^2}{2m}\right) = \exp\left(\frac{-m\epsilon^2}{2}\right).
\end{aligned}$$

For a chosen confidence level δ this gives the high-confidence bound claimed in the lemma with $\epsilon = \sqrt{\frac{2\ln(1/\delta)}{m}}$. □

We will use the above lemma to show that we can construct from $h_1, \ldots, h_{m+1}$ a *single* model whose generalization error can be bounded in terms of the observed cumulative loss.

2.1 Convex (bounded) losses

When the prediction space $\hat{\mathcal{Y}}$ of the online learner is convex, it is possible to define an averaged predictor based on the online models $\{h_t\}_{t=1}^{m+1}$. If the loss function is also convex, then the generalization error of this averaged predictor can be bounded by the average generalization error of all the online models, *i.e.* the quantity bounded in Lemma 2.1.

Theorem 2.2 (Cesa-Bianchi *et al*, 2004)**.** Let $\hat{\mathcal{Y}}$ be a convex set, $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to [0,1]$ be convex in its second argument and $h_1, \ldots, h_{m+1}$ be the models returned by an online algorithm on the sequence of examples $S = ((x_1, y_1), \ldots, (x_m, y_m))$ as before. Define the average predictor

$$\bar{h}_S = \frac{1}{m}\sum_{t=1}^{m} h_t.$$

Let D be any distribution on $\mathcal{X} \times \mathcal{Y}$ and $\delta \in (0,1]$. Then *w.p.* $\geq 1-\delta$ (over the draw of $S \sim D^m$)

$$\mathrm{er}_D^{\ell}[\bar{h}_S] \leq \frac{1}{m} L_S^{\ell}[\mathcal{A}] + \sqrt{\frac{2\ln(1/\delta)}{m}}.$$

Proof. We have the generalization error of the average predictor

$$\begin{aligned}
\mathrm{er}_D^{\ell}[\bar{h}_S] &= \mathbf{E}_{(x,y)\sim D}\left[\ell(y, \bar{h}_S(x))\right] \\
&= \mathbf{E}_{(x,y)\sim D}\left[\ell(y, \frac{1}{m}\sum_{t=1}^{m} h_t(x))\right] \\
&\leq \mathbf{E}_{(x,y)\sim D}\left[\frac{1}{m}\sum_{t=1}^{m} \ell(y, h_t(x))\right] \text{ by the convexity of } \ell \\
&= \frac{1}{m}\sum_{t=1}^{m} \mathrm{er}_D^{\ell}[h_t] \\
&\leq \frac{1}{m} L_S^{\ell}[\mathcal{A}] + \sqrt{\frac{2\ln(1/\delta)}{m}} \text{ by Lemma 2.1.}
\end{aligned}$$

□

The total loss of the online learner plays the role of an empirical loss measure in this generalization error bound.

2.2 General (bounded) losses

We have seen in Lemma 2.1 that atleast one of the models returned by the online learner has low generalization error (bounded in terms of the total online loss). Cesa-Bianchi *et al* show that it is possible to identify such a model with high probability. This is done by choosing the model h_t that minimizing a penalized empirical loss measured over examples $(x_{t+1}, y_{t+1}), \ldots, (x_m, y_m)$ *not seen* by the online algorithm when the model h_t was learnt. The penalty term compensates for the fact that the empirical loss is estimated with "test-sets" of different sizes for the different models (*i.e.* $m - t + 1$ test samples for model h_t).

Theorem 2.3 (Cesa-Bianchi *et al*, 2004)**.** Let $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to [0,1]$ be a loss function and $h_1, \ldots, h_{m+1}$ be the models returned by an online algorithm as before. For any $\delta' \in (0,1]$, let

$$\hat{h}_S^{\delta'} \in \arg\min_{t\in[m]} \left[\frac{1}{m-t+1}\sum_{i=t}^{m} \ell(y_i, h_t(x_i)) + \sqrt{\frac{1}{2(m-t+1)}\ln\left(\frac{m(m+1)}{\delta'}\right)}\right]$$

be the model minimizing the penalized empirical loss. Then for any distribution D and $\delta \in (0,1]$, *w.p.* $\geq 1-\delta$ (over the draw of $S \sim D^m$)

$$\mathrm{er}_D^\ell[h_S^{\delta/2}] \leq \frac{1}{m} L_S^\ell[\mathcal{A}] + 6\sqrt{\frac{1}{m}\ln\left(\frac{2(m+1)}{\delta}\right)}.$$

Details of the proof can be found in [Cesa-Bianchi *et al*, 2004].

3 Online to Batch: Consistency guarantees

Having seen online-to-batch conversions that return models with confidence-bounded generalization errors, we turn to the question of the statistical consistency of the resulting models.

Theorem 3.1. Let $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to [0,1]$, and $\mathcal{A}$ be an online algorithm for which we can show that for a training set $S \in (\mathcal{X} \times \mathcal{Y})^m$, we can construct a model $h_S : \mathcal{X} \to \hat{\mathcal{Y}}$ that achieves, for any distribution D and $\delta \in (0,1]$, *w.p.* $\geq 1-\delta$ (over $S \sim D^m$)

$$\mathrm{er}_D^\ell[h_S] \leq \frac{1}{m} L_S^\ell[\mathcal{A}] + f(m,\delta)$$

where $\forall \delta$, $f(m,\delta) \to 0$ as $m \to \infty$.[1] Further, let $\mathcal{H} \subseteq \hat{\mathcal{Y}}^{\mathcal{X}}$ be the function class from which models are drawn both by the online algorithm and the online-to-batch convertor. Then $\forall D$ and for $\delta \in (0,1]$, *w.p.* $\geq 1-\delta$ (over $S \sim D^m$)

$$\mathrm{er}_D^\ell[h_S] - \inf_{h\in\mathcal{H}} \mathrm{er}_D^\ell[h] \leq \frac{1}{m}\left(L_S^\ell[\mathcal{A}] - \inf_{h\in\mathcal{H}} L_S^\ell[h]\right) + f(m,\delta) + \sqrt{\frac{2\ln(2/\delta)}{m}}.$$

Proof. Let $h^* \in \inf_{h\in\mathcal{H}} \mathrm{er}_D^\ell[h]$ be a function realizing the optimal generalization error in the function class $\mathcal{H}$ (assume the infimum is achieved for simplicity; a similar argument works in the more general case as well). Then *w.p.* $\geq 1-\delta/2$

$$\begin{aligned}
\mathrm{er}_D^\ell[h_S] \leq & \frac{1}{m} L_S^\ell[\mathcal{A}] + f(m,\delta/2) \\
= & \frac{1}{m}\inf_{h\in\mathcal{H}} L_S^\ell[h] + \underbrace{\frac{1}{m}\left(L_S^\ell[\mathcal{A}] - \inf_{h\in\mathcal{H}} L_S^\ell[h]\right) + f(m,\delta/2)}_{\text{denote this as } Q} \qquad \left(\text{by adding and subtracting } \frac{1}{m}\inf_{h\in\mathcal{H}} L_S^\ell[h]\right) \\
= & \inf_{h\in\mathcal{H}} \frac{1}{m}\sum_{t=1}^m \ell(y_t, h(x_t)) + Q \\
\leq & \frac{1}{m}\sum_{t=1}^m \ell(y_t, h^*(x_t)) + Q \\
& \text{where the first term, estimating the loss of } h^* \text{ on } \textit{i.i.d.} \text{ samples } (x_t,y_t) \sim D, \text{ can be upper} \\
& \text{bounded } \textit{w.p.} \geq 1-\delta/2 \text{ by Hoeffding's inequality as} \\
\leq & \mathrm{er}_D^\ell[h^*] + \sqrt{\frac{2\ln(2/\delta)}{m}} + Q.
\end{aligned}$$

Substituting back the definition of Q and rearranging terms gives the result claimed in the theorem. □

The above result bounds the statistical regret (estimation error) of h_S (*w.r.t.* the function class $\mathcal{H}$) in terms of the online per-trial regret of the online algorithm $\mathcal{A}$ (*w.r.t.* the function class $\mathcal{H}$). If the per-trial regret of the online learner goes to zero as $m \to \infty$, then this result implies universal statistical consistency of the online-to-batch conversion within $\mathcal{H}$. If the online-regret tends to zero at a rate independent of the sequence S, then this result also proves learnability of the function class $\mathcal{H}$.

[1] Note that for both the averaged predictor for convex loss and the penalized empirical loss minimizing predictor for general losses, the $f(m,\delta)$ function has this asymptotic behavior.

4 Example: Online Linear Regression by Gradient Descent

We now apply the online-to-batch conversion process to a concrete example, that of online linear regression by gradient descent for which we showed in Lecture 17 a regret bound *w.r.t.* the class of linear regressors with bounded norm.

Fix the instance, label and prediction spaces as $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n \big| \|\mathbf{x}\|_2 \leq 1\}$, $\mathcal{Y} = \hat{\mathcal{Y}} = [-1, 1]$ and the function class of linear regressors with bounded norm as $\mathcal{H} = \{h : \mathcal{X} \to \hat{\mathcal{Y}} \big| h(\mathbf{x}) = \mathbf{u} \cdot \mathbf{x} \text{ for some } \mathbf{u} \in \mathbb{R}^n, \|\mathbf{u}\|_2 \leq 1\}$. We consider the absolute loss $\ell_{\text{abs}} : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+, \ell(y, \hat{y}) = |y - \hat{y}|$, which for the present choice of $\mathcal{Y}$ and $\hat{\mathcal{Y}}$ is bounded in $[0, 2]$.

If we use the online gradient descent algorithm[2] GD(η) with learning rate η to generate the models $\{h_t\}_{t=1}^{m+1}$, then by Theorem 2.2 (adapted to loss bounded in $[0, 2]$) the averaged predictor defined as $\bar{h}_S = {}^1\!/_m \sum_{t=1}^m h_t$ satisfies, *w.p.* $\geq 1 - \frac{\delta}{2}$, the generalization error bound

$$\text{er}_D^{\text{abs}}[\bar{h}_S] \leq \frac{1}{m} L_S^{\text{abs}}[\text{GD}(\eta)] + \sqrt{\frac{8 \ln ({}^2\!/_\delta)}{m}}. \tag{1}$$

We showed in Lecture 17 the following regret bound for GD with its learning-rate optimized to $\eta^* = \frac{U}{R_2 Z \sqrt{m}}$, where R_2 and U are bounds on the 2-norms of the feature and weight vectors respectively and Z is an absolute bound on the sub-gradient of the loss function. In the present case, with $R_2 = U = Z = 1$ and $\eta^* = \frac{1}{\sqrt{m}}$:

$$L_S^{\text{abs}}[\text{GD}(\eta^*)] - \inf_{h \in \mathcal{H}} L_S^{\text{abs}}[h] \leq U R_2 Z \sqrt{m} = \sqrt{m},$$

which implies a per-trial regret of ${}^1\!/_{\sqrt{m}}$. Further, we have ${}^1\!/_m \inf_{h \in \mathcal{H}} L_S^{\text{abs}}[h] \leq {}^1\!/_m L_S^{\text{abs}}[h^*]$ and the latter can be upper bounded *w.p.* $\geq 1 - \frac{\delta}{2}$ in terms of its expectation $\text{er}_D^{\text{abs}}[h^*]$ using Hoeffding's inequality. Putting these together, we have, *w.p.* $\geq 1 - \delta$, a bound on the statistical regret of $\bar{h}_S$,

$$\text{er}_D^{\text{abs}}[\bar{h}_S] - \inf_{h \in \mathcal{H}} \text{er}_D^{\text{abs}}[h] \leq \frac{1}{\sqrt{m}} + 2\sqrt{\frac{8 \ln ({}^2\!/_\delta)}{m}},$$

which tends to zero as $m \to \infty$ for any D at a fixed rate, demonstrating both the universal statistical consistency of the resulting online-to-batch algorithm in $\mathcal{H}$ *w.r.t.* ℓ_{abs}, as well as learnability of the function class $\mathcal{H}$ (*w.r.t.* ℓ_{abs}).

A Hoeffding-Azuma inequality for martingales

Theorem A.1. Let $\{Y_n\}$ be a martingale sequence *w.r.t.* a stochastic process $\{X_n\}$ (both defined over the same probability space); *i.e.*

$$\mathbf{E}_{X_n}[Y_n | X_1, \ldots, X_{n-1}] = Y_{n-1}.$$

Define the martingale difference sequence $V_n = Y_n - Y_{n-1}$ and suppose that there exist $c_i > 0, \forall i$ s.t. $|V_i| \leq c_i, \forall i$; then $\forall \epsilon > 0$,

$$\left.\begin{array}{l} \mathbf{P}[Y_n - Y_0 \geq \epsilon] \\ \mathbf{P}[Y_n - Y_0 \leq -\epsilon] \end{array}\right\} \leq \exp\left(\frac{-\epsilon^2}{2\sum_{i=1}^n c_i^2}\right).$$

References

[1] Nicolo 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.

[2] using normalized weight updates $\mathbf{w}_{t+1} \leftarrow \frac{\mathbf{w}_t - \eta \ell'_{\text{abs}}(y_t, \hat{y}_t)}{\|\mathbf{w}_t - \eta \ell'_{\text{abs}}(y_t, \hat{y}_t)\|_2}$ (where $\ell'_{\text{abs}}(y_t, \hat{y}_t)$ is any subgradient of $\ell_{\text{abs}}(y_t, \cdot)$ at $\hat{y}_t$) to ensure that all models lie within $\mathcal{H}$.