CMSC 35900 (Spring 2008) Learning Theory **Lecture: 13**

Online to Batch Conversions

Instructors: Sham Kakade and Ambuj Tewari

1 Using Online Algorithms in a Batch Setting

We have recently been studying the case where have a training set T generated from an underlying distribution and our goal is to find some good hypothesis, with respect to the true underlying distribution, using the training set T. We now examine how to use online learning algorithms (which work on individual, arbitrary sequences) in a stochastic setting.

Let us consider the training set T as the *ordered sequence*:

$$T = \{(X_1, Y_1), \dots, (X_m, Y_m)\}$$

and let us run an online learning algorithm on this sequence. In particular, let us say that each round t our algorithm chooses some $\theta \in \Theta$ and we suffer loss $\ell(\theta; (x_i, y_i))$. Here, the decision space/parameter space Θ could be the space corresponding to the parameterization of our hypothesis class. The regret of our algorithm on the sequence is defined as:

$$R_T = \sum_{i=1}^{m} \ell(\theta_i; (x_i, y_i)) - \inf_{\theta \in \Theta} \sum_{i=1}^{m} \ell(\theta; (x_i, y_i))$$

Previously, we studied algorithms which provides bounds for this regret on arbitrary sequences T.

Now if we use an online algorithm on a sequence T, then we would like to use the algorithms behavior to find a hypothesis that is good with respect to the distribution.

2 Martingales

A stochastic process $X_1, X_2, \dots X_m$ is a martingale if $\mathbb{E}[|X_i|] \leq \infty$ and:

$$\mathbb{E}[X_i | X_1, \dots X_{i-1}] = X_{i-1}$$

If we have a filtration $\{H_i\}$ (think of this like a "history") where X_i is measurable with respect to H_i (i.e. X_i is a deterministic function of H_i), then $X_1, X_2, \dots X_m$ is a martingale with respect to this filtration if $\mathbb{E}[|X_i|] \leq \infty$ and:

$$\mathbb{E}[X_i | H_{i-1}] = X_{i-1}$$

The process $Z_1, Z_2, \dots Z_m$ is a martingale difference sequence if $\mathbb{E}[|Z_i|] \leq \infty$ and

$$\mathbb{E}[Z_i | H_{i-1}] = 0$$

Clearly, $Z_i = X_i - X_{i-1}$ is a martingale difference sequence.

A useful property of martingale different sequences is that:

$$\mathbb{E}[Z_i] = 0$$

Here, we have an unconditional expectation.

3 Online to "Batch"

Let us define

$$Z_i = (\ell(\theta_i; (x_i, y_i)) - \mathcal{L}(\theta_i)) - (\ell(\theta^*; (x_i, y_i)) - \mathcal{L}(\theta^*)) \quad .$$

With respect to the history $T_{<i}$, this process is a martingale difference sequence.

The following lemma is useful.

Lemma 3.1. *Assume that each (x_i, y_i) is generated in an i.i.d manner. Assume that θ_i is a deterministic function of $T_{<i}$, where:*

$$T_{<i} = \{(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1})\}$$

Then the process $\{Z_i\}$ is a martingale difference sequence, with respect to the history $T_{<i}$.

Proof. To see that the process is a martingale difference sequence,

$$\begin{aligned}
\mathbb{E}\left[Z_i | T_{<i}\right] &= \mathbb{E}\left[\ell(\theta_i; (x_i, y_i)) - \mathcal{L}(\theta_i) | T_{<i}\right] - \mathbb{E}\left[\ell(\theta^*; (x_i, y_i)) - \mathcal{L}(\theta^*) | T_{<i}\right] \\
&= \mathcal{L}(\theta_i) - \mathcal{L}(\theta_i) - (\mathcal{L}(\theta^*) - \mathcal{L}(\theta^*)) \\
&= 0
\end{aligned}$$

which completes the proof. □

Lemma 3.2. *We have that*

$$\sum_{i=1}^{m} \mathcal{L}(\theta_i) \leq \mathcal{L}(\theta^*) + R_T - \sum_{i=1}^{m} Z_i$$

Proof. To complete the proof:

$$\begin{aligned}
\sum_{i=1}^{m} \mathcal{L}(\theta_i) - \mathcal{L}(\theta^*) &= \sum_{i=1}^{m} \ell(\theta_i; (x_i, y_i)) - \ell(\theta^*; (x_i, y_i)) - Z_i \\
&\leq \sum_{i=1}^{m} \ell(\theta_i; (x_i, y_i)) - \inf_{\theta \in \Theta} \sum_{i=1}^{m} \ell(\theta; (x_i, y_i)) - \sum_{i=1}^{m} Z_i \\
&= R_T - \sum_{i=1}^{m} Z_i
\end{aligned}$$

which completes the proof.

□

The following theorem bounds the expected performance of an online to batch conversion.

Theorem 3.3. *Assume that each (x_i, y_i) is generated in an i.i.d manner. Assume that θ_i is a deterministic function of $T_{<i}$. Let θ^* be defined as:*

$$\theta^* \in \mathrm{argmin}_{\theta \in \Theta} \mathcal{L}(\theta)$$

Let $\theta_1, \ldots \theta_m$ be the random variable corresponding to the output of our online algorithm on the training sequence T (generated in an i.i.d. manner from some distribution). Then:

$$\mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\theta_i)\right] \leq \mathcal{L}(\theta^*) + \frac{1}{m} \mathbb{E}\left[R_T\right]$$

where the expectation is with respect to T. Furthermore, if $\mathcal{L}(\cdot)$ is convex, then:

$$\mathbb{E}\left[\mathcal{L}\left(\frac{1}{m} \sum_{i=1}^{m} \theta_i\right)\right] \leq \mathcal{L}(\theta^*) + \frac{1}{m} \mathbb{E}\left[R_T\right]$$

Proof. Since Z_i is a martingale difference sequence, we have

$$\mathbb{E}\left[\sum_{i=1}^{m} Z_i\right] = 0$$

where the expectation is unconditional. Now just take expectations in the previous lemma. $\square$

3.1 With High Probability

The following concentration result is useful.

Theorem 3.4. *(Hoeffding-Azuma) Let $Z_1, Z_2, \dots Z_m$ be a martingale difference sequence s.t. $|Z_i| \leq B$ (with probability one). For all $\epsilon \geq 0$*

$$\mathbb{P}\left(\sum_{i=1}^{m} Z_i \geq \epsilon\right) \leq e^{-\frac{\epsilon^2}{2B^2 m}}$$

The following high probability statement is now straightforward.

Theorem 3.5. *Assume that each (x_i, y_i) is generated in an i.i.d manner. Assume that θ_i is a deterministic function of $T_{<i}$. Let θ^* be defined as:*

$$\theta^* \in \mathrm{argmin}_{\theta \in \Theta} \mathcal{L}(\theta)$$

Let $\theta_1, \dots \theta_m$ be the random variable corresponding to the output of our online algorithm on the training sequence T (generated in an i.i.d. manner from some distribution). Assuming that the loss is bounded in $[0, 1]$, then with probability greater than $1 - \delta$

$$\frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\theta_i) \leq \mathcal{L}(\theta^*) + \frac{1}{m}\mathbb{E}\left[R_T\right] + 2\sqrt{\frac{2\log\frac{1}{\delta}}{m}}$$

where the expectation is with respect to T.

Proof. Clearly, Z_i is bounded by 2. Hence, with probability greater than $1 - \delta$

$$\frac{1}{m}\sum_{i=1}^{m} Z_i \leq 2\sqrt{\frac{2\log\frac{1}{\delta}}{m}}$$

The proof follows from our earlier lemma. $\square$

4 L1 and L2 constrained problems

In the online learning setting, we restricted model complexity by bounding the decision region. We could consider similar restrictions in the stochastic setting.

For the case with an L_2 bounded decision region, we have:

$$\theta_2^* = \mathrm{argmin}_{\theta : ||\theta||_2 \leq D_2} \mathcal{L}(\theta)$$

where D_2 is some bound on the norm of the decision region. Similarly, we could consider an L_1 constrained decision region, with optimal predictor:

$$\theta_1^* = \mathrm{argmin}_{\theta : ||\theta||_2 \leq D_1} \mathcal{L}(\theta)$$

where, again, D_1 is a bound on the L_1 norm of the decision region.

4.1 Regularization

A natural question is to come up with an estimator $\hat{\theta}$, which is a function of the training sequence, in which the right hand side of

$$\mathcal{L}(\hat{\theta}_2) - \mathcal{L}(\theta_2^*) \leq ??$$
$$\mathcal{L}(\hat{\theta}_1) - \mathcal{L}(\theta_1^*) \leq ??$$

is small.

Two natural estimators are:

$$\hat{\theta}_2 = \text{argmin}_{\theta:||\theta||_2 \leq D_2} \sum_{i=1}^{m} \ell(\theta_i; (x_i, y_i))$$
$$\hat{\theta}_1 = \text{argmin}_{\theta:||\theta||_2 \leq D_1} \sum_{i=1}^{m} \ell(\theta_i; (x_i, y_i))$$

However, it is not year clear if there estimators always perform favorably. The duals of these problem are often referred to as regularization:

$$\hat{\theta}_2 = \text{argmin}_{\theta} \sum_{i=1}^{m} \ell(\theta_i; (x_i, y_i)) + \lambda ||\theta||_2^2$$
$$\hat{\theta}_1 = \text{argmin}_{\theta} \sum_{i=1}^{m} \ell(\theta_i; (x_i, y_i)) + \lambda ||\theta||_1$$

which will consider later.

4.2 Online to Batch Conversions for OLCP

Now we can apply our previous results on Gradient Descent and Exponentiated Gradient descent to this setting.

Corollary 4.1. *Assuming that $\ell(\theta; (x, y))$ is a convex function of θ for all (x, y), then the with probability greater than $1 - \delta$, the output of the gradient descent algorithm satisfies:*

$$\mathcal{L}(\frac{1}{m}\sum_{i=1}^{m}) - \mathcal{L}(\theta_2^*) \leq \frac{G_2 D_2}{\sqrt{m}} + 2\sqrt{\frac{2 \log \frac{1}{\delta}}{m}}$$

where G_2 is an upper bound on $||\nabla \ell(\theta; (x, y))||_2$ (for all (x, y)).

Corollary 4.2. *Assuming that $\ell(\theta; (x, y))$ is a convex function of θ for all (x, y), then the with probability greater than $1 - \delta$, the output of the exponentiated gradient descent algorithm satisfies:*

$$\mathcal{L}(\frac{1}{m}\sum_{i=1}^{m}) - \mathcal{L}(\theta_2^*) \leq 2\frac{G_\infty D_1}{\sqrt{m}} + 2\sqrt{\frac{2 \log \frac{1}{\delta}}{m}}$$

where G_∞ is an upper bound on $||\nabla \ell(\theta; (x, y))||_\infty$ (for all (x, y)).

Proof. The proof directly follow from the previous theorem and the fact that R_T is bounded uniformly (as we saw in an earlier lecture). $\square$