CSE522, Winter 2011, Learning Theory Lecture 18 - 02/15/2011

Online to Batch Conversion

Lecturer: Ofer Dekel *Scribe: Krishnamurthy Dvijotham*

1 Recap of Online Learning Algorithms

Traditional supervised learning is formulated as learning from a given data set while being able to generalize to unseen data. It is usually assumed that both the given and unseen data are drawn iid from the same (unknown) distribution. In online learning, we make no assumption about the source of data. One simply observes a stream of data coming from some arbitrary source one by one. At every step, the online learning algorithm tries to make a correct prediction for the next data point. After making the prediction, it observes the true label and updates its hypothesis in light of the new evidence. The goal is to minimize **regret**:

$$\sum_{i=1}^{m} l(h_{i-1}; (x_i, y_i)) - \min_{h \in \mathcal{H}} \sum_{i=1}^{m} l(h; (x_i, y_i))$$

This basically says that the online learning algorithm wants performance close to the single best hypothesis chosen in hindsight given all the data. This is a very general notion of performance that makes sense even when the data is generated by an adversary who is trying to trick the algorithm. In fact, it can be seen as a zero-sum repeated game where the algorithm is trying to minimize the regret while the adversary generating the data is trying to maximize it.

We can take any online learning algorithm and use it for supervised learning as follows: Run the online algorithm on the fixed data set (given in any arbitrary order) and output a randomized hypothesis that has a uniform distribution Q on the hypotheses $h_0, h_1, \ldots, h_{m-1}$. We showed in the last lecture, using an argument based on a Doob Martingale+Azuma's inequality that:

$$l\left(Q; \mathcal{D}\right) \leq \frac{1}{m}\left(\sum_{i} l\left(h_{i-1}; (x_i, y_i)\right)\right) + c\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}} \tag{1}$$

Thus, if we have bounds on the regret of an online learning algorithm, we can combine the above bound with the Hoeffding bound to obtain a bound on the excess risk of the randomized hypothesis Q.

2 Constrained Subgradient Descent (GD)

Consider the class of linear predictors $\mathcal{H} = \{w \in \mathbb{R}^n : \|w\|_2 \leq B\}$ on the input space $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_2 \leq X\}$. Let l be a λ-Lipschtiz convex loss function bounded in [0,c]. Let $l_i(w) = l(w; (x_i, y_i))$ and $\nabla l_i(w)$ denote a subgradient of l_i at w. Let $\prod[w]$ denote the projection of w onto the $\mathcal{H}$ (here $\prod[w] = \frac{B}{\|w\|} w$). The constrained subgradient descent algorithm works as follows:

2.1 Bounding the Regret

Theorem 1. *Let $\mathcal{H} = \{w \in \mathbb{R}^n : \|w\|_2 \leq B\}, \mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_2 \leq X\}$. Let l be a convex λ-Lipschitz loss function. Let $w^* \in \mathcal{H}$ be any fixed hypothesis and $S = \{(x_i, y_i)\}_{i=1}^{m} \subset (\mathcal{X} \times \mathcal{Y})^m$ be arbitrary. Let $w_0, w_1, \ldots, w_m$ be the hypotheses generated by running the GD algorithm with stepsize $\eta = \frac{B}{\lambda X \sqrt{m}}$ on S(given*

Algorithm 1 Constrained Gradient Descent (GD)

```
$w_0 \leftarrow \vec{0}$
**for** i=1,...,m **do**
  Receive $x_i$, Predict $\langle w_{i-1}, x \rangle$, Receive $y_i \in \{-1, +1\}$, suffer loss $l\left(w_{i-1}; (x_i, y_i)\right)$
  $w_i \leftarrow \prod \left[w_{i-1} - \eta \nabla l_i(w_{i-1})\right]$
**end for**
```

in any arbitrary order). Then we have:

$$\left(\sum_i l(w_{i-1}; (x_i, y_i))\right) \leq \left(\sum_i l(w^*; (x_i, y_i))\right) + \lambda X B \sqrt{m}$$

In particular this holds for $w^* = \operatorname{argmin}_{w \in \mathcal{H}} l(w; S)$*, giving us the regret bound*

$$\text{Regret} \leq \lambda B X \sqrt{m}$$

Proof. Define

$$\alpha_i = \frac{1}{2}\left\|w_{i-1} - w^*\right\|^2 - \frac{1}{2}\left\|w_i - w^*\right\|^2$$

We know that $\|w_i - w^*\| \leq \|w_{i-1} - \eta \nabla l_i(w_{i-1}) - w^*\|$, since the projection operator brings $w_{i-1} - \eta \nabla l_i(w_{i-1})$ *uniformly* closer to every element of $\mathcal{H}$ (specifically to w^*). Using this, we get

$$\begin{aligned}
\alpha_i &\geq \frac{1}{2}\left\|w_{i-1} - w^*\right\|^2 - \frac{1}{2}\left\|w_{i-1} - \eta \nabla l_i(w_{i-1}) - w^*\right\| \\
&= -\frac{\eta^2}{2}\left\|\nabla l_i(w_{i-1})\right\|^2 - \eta \left\langle w^* - w_{i-1}, \nabla l_i(w_{i-1}) \right\rangle \\
&\geq -\frac{\eta^2}{2}\left\|\nabla l_i(w_{i-1})\right\|^2 + \eta(l_i(w_{i-1}) - l_i(w^*)) \quad \text{(Since } \nabla l_i(w_{i-1}) \text{ is a subgradient)} \\
&\geq -\frac{\eta^2 \lambda^2 X^2}{2} + \eta(l_i(w_{i-1}) - l_i(w^*)) \quad \text{(Since } l \text{ is } \lambda\text{-Lipschitz)} \qquad (2)
\end{aligned}$$

Summing over i $\alpha_i = \frac{1}{2}\left\|w_{i-1} - w^*\right\|^2 - \frac{1}{2}\left\|w_i - w^*\right\|^2$, all terms except the last negative and the first positive term cancel out, and we get

$$\sum_i \alpha_i = \frac{1}{2}\left\|w_0 - w^*\right\|^2 - \frac{1}{2}\left\|w_m - w^*\right\|^2 \leq \frac{1}{2}\left\|w_0 - w^*\right\|^2 = \frac{\left\|w^*\right\|^2}{2} \leq \frac{B^2}{2}$$

since $w_0 = 0, \|w^*\| \leq B$. Using the lower bound (2), we get

$$\sum_i \alpha_i \geq -m\frac{\eta^2 \lambda^2 X^2}{2} + \eta\left(\sum_i l(w_{i-1}; (x_i, y_i)) - l(w^*; (x_i, y_i))\right) = -\frac{\eta^2 \lambda^2 X^2 m}{2} + \eta \text{Regret}$$

Comparing the lower bound to the upper bound and dividing by η, we get

$$\left(\sum_i l(w_{i-1}; (x_i, y_i))\right) - \left(\sum_i l(w^*; (x_i, y_i))\right) \leq \frac{B^2}{2\eta} + \frac{\eta \lambda^2 X^2 m}{2}$$

We choose $\eta = \frac{B}{\lambda X \sqrt{m}}$ to minimize the RHS, and get

$$\left(\sum_i l(w_{i-1}; (x_i, y_i))\right) \leq \left(\sum_i l(w^*; (x_i, y_i))\right) + \lambda X B \sqrt{m}$$

□

Thus, we have a regret that grows sub-linearly with the number of samples observed. Asymptotically, this gives us an average regret per sample ($\frac{\text{Regret}}{m}$) that goes to 0 at rate $\frac{1}{\sqrt{m}}$.

2.2 Application to Statistical Learning

Theorem 2. *Let $\mathcal{H} = \{w \in \mathbb{R}^n : \|w\|_2 \leq B\}, \mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_2 \leq X\}$. Let l be a convex λ-Lipschitz loss function bounded in $[0, c]$. Let $S \sim \mathcal{D}^m$ and let Q be the uniform distribution over hypotheses generated by applying the GD algorithm to S (given in any arbitrary order). Let $w^* \in \operatorname{argmin}_{w \in \mathcal{H}} l(w; \mathcal{D})$. Then $w.p \geq 1 - \delta$ over the sampling of S, the excess risk of Q is bounded as:*

$$l\left(Q; \mathcal{D}\right) - l(w^*; \mathcal{D}) \leq \frac{\lambda X B}{\sqrt{m}} + 2c\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2m}}$$

Proof. Let $S \sim \mathcal{D}^m$. Then $w.p \geq 1 - \delta$, we have (from (1))

$$l\left(Q; \mathcal{D}\right) \leq \frac{1}{m}\left(\sum_i l\left(w_{i-1}; (x_i, y_i)\right)\right) + c\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}$$

From theorem 1, we know that

$$\sum_i l\left(w_{i-1}; (x_i, y_i)\right) \leq m l(w^*; S) + \lambda X B\sqrt{m}$$

Thus, we have $w.p \geq 1 - \delta$

$$l\left(Q; \mathcal{D}\right) \leq l(w^*; S) + \frac{\lambda X B}{\sqrt{m}} + c\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}$$

From the Hoeffding bound applied to $l(w^*; S)$, we have $w.p \geq 1 - \delta$

$$l(w^*; S) \leq l(w^*; \mathcal{D}) + c\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}$$

Combining the above inequalities using the union bound, we get, $w.p \geq 1 - 2\delta$

$$l\left(Q; \mathcal{D}\right) - l(w^*; \mathcal{D}) \leq \frac{\lambda X B}{\sqrt{m}} + 2c\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}$$

Restating this, we get the following bound on the excess risk of the GD algorithm for supervised learning: $w.p \geq 1 - \delta$, we get

$$l\left(Q; \mathcal{D}\right) - l(w^*; \mathcal{D}) \leq \frac{\lambda X B}{\sqrt{m}} + 2c\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2m}} \tag{3}$$

□

2.3 Derandomization

We discuss various schemes to de-randomize Q and get a single hypothesis w with good generalization properties:

1 **Mean:** $\mathcal{H}$ is convex, so $\frac{\sum_{i=1}^m w_i}{m} = \text{Exp}_Q [w] \in \mathcal{H}$. Since l is convex, using Jensen's inequality, we get

$$l\left(\frac{\sum_{i=1}^m w_i}{m}; \mathcal{D}\right) \leq l(Q; \mathcal{D})$$

Thus we can output the average as a derandomized single hypothesis.

2 **Majority vote:** The majority vote hypothesis is defined as:

$$h_{maj}(x) = \text{sign}\left(\frac{\sum_{i=1}^m \text{sign}\left(\langle w_i, x\rangle\right)}{m}\right)$$

We know that the $0-1$ risk of the majority vote hypothesis is at most twice the $0-1$ risk of Q. Thus, if l is a convex upper bound on the $0-1$ loss (like the hinge loss/exponential loss), we have

$$l_{0-1}(h_{maj}(x); \mathcal{D}) \leq 2l_{0-1}(Q; \mathcal{D}) \leq 2l(Q; \mathcal{D})$$

3 **Dynamic Validation:** Note that w_{i-1} is independent of $\{(x_k, y_k)\}_{k=i}^m$. Using a Hoeffding+Union bound, w.p. at least $1-\delta$,

$$\forall i \quad \left| l\left(h_{i-1}; \{x_k, y_k\}_{k=i}^m\right) - l\left(h_{i-1}; \mathcal{D}\right) \right| \leq c\sqrt{\frac{\log\left(m/\delta\right)}{2(m-i+1)}}$$

Let

$$\tilde{h} = \underset{\{h_0, h_1, \ldots, h_{m-1}\}}{\text{argmin}} \; l\left(h_{i-1}; \{x_k, y_k\}_{k=i}^m\right) + c\sqrt{\frac{\log\left(m/\delta\right)}{2(m-i+1)}}$$

Define $\tilde{\tilde{h}}$ to be the output of ERM run on the first half of the dataset $S' = \{(x_i, y_i\}_{i=1}^{m/2-1}$. We know that $\tilde{h}$ is at least as good as $\tilde{\tilde{h}}$. Let h^* be the true risk minimizer among $h_0, \ldots, h_{m/2-1}$. We know that

$$l(\tilde{\tilde{h}}; \mathcal{D}) \leq l(\tilde{\tilde{h}}; S') + c\sqrt{\frac{\log m/\delta}{2(m/2)}}$$

with probability at least δ. Now, by definition $l(\tilde{\tilde{h}}; S') \leq l(h^*; S')$ and by a Hoeffding bound $l(h^*; S') \leq l(h^*; \mathcal{D}) + c\sqrt{\frac{\log(m/\delta)}{2(m/2)}}$ with probability at least $1-\delta$. Thus, with probability at least $1-2\delta$

$$l(\tilde{\tilde{h}}; \mathcal{D}) \leq l(h^*; \mathcal{D}) + 2c\sqrt{\frac{\log\left(m/\delta\right)}{m}}$$

so that with probability at least $1-\delta$,

$$l(\tilde{h}; \mathcal{D}) \leq l(h^*; \mathcal{D}) + 2c\sqrt{\frac{\log\left(2m/\delta\right)}{m}}$$

3 Exponentiated Gradient (EG)

We now describe an alternate online learning algorithm that works in the "experts" setting: Here we assume that we have a set of experts making each of whom predictions (possibly with some confidence) and the algorithm tries to weight them outputs an optimal convex combination of the predictions that is as accurate as possible. More concretely, Consider the class of linear predictors $\mathcal{H} = \{w \in \mathbb{R}^n : w \geq 0 \sum_i w_i = 1\}$ on the input space $\mathcal{X} = \{x \in \mathbb{R}^n : \|x\|_\infty \leq X\}$. Let l be a λ-Lipschtiz convex loss function. Let $l_i(w) = l(w; (x_i, y_i))$ and $\nabla l_i(w)$ denote a subgradient of l_i at w and $\nabla_j l_i(w)$ denote its j-th component.

Algorithm 2 Exponentiated Gradient (EG)

```
$w_0 \leftarrow \{\frac{1}{n}, \dots, \frac{1}{n}\}$
for i=1,...,m do
    Receive $x_i$, Predict $\langle w_{i-1}, x \rangle$, Receive $y_i \in \{-1, +1\}$, suffer loss $l\left(w_{i-1}; (x_i, y_i)\right)$
    Update: $\forall j \quad w_{i,j} \leftarrow \frac{w_{i-1,j}\exp(-\eta \nabla_j l_i(w_{i-1}))}{\sum_k w_{i-1,k}\exp(-\eta \nabla_k l_i(w_{i-1}))}$
end for
```

Theorem 3. *Let $w^* \in \mathcal{H}$ be any fixed hypothesis and $S = \{(x_i, y_i)\}_{i=1}^m \subset (\mathcal{X} \times \mathcal{Y})^m$ be arbitrary. Let $w_0, w_1, \dots, w_m \in \mathcal{H}$ be the hypotheses generated by running the EG algorithm with stepsize $\eta = \frac{1}{\lambda X}\sqrt{\frac{2\log(n)}{m}}$ on S(given in any arbitrary order). Then we have:*

$$\left(\sum_i l(w_{i-1}; (x_i, y_i))\right) \leq \left(\sum_i l(w^*; (x_i, y_i))\right) + \lambda X \sqrt{2n \log(m)}$$

In particular this holds for $w^* = \operatorname{argmin}_{w \in \mathcal{H}} l(w; S)$, giving us the regret bound

$$\text{Regret} \leq \lambda X \sqrt{2m \log(n)}$$

Proof. Each $w \in \mathcal{H}$ can be regarded as a probability distribution on $\{1, 2, \dots, n\}$. Define

$$\alpha_i = \mathrm{KL}\left(w^* \parallel w_{i-1}\right) - \mathrm{KL}\left(w^* \parallel w_i\right)$$

$$\begin{aligned}
\alpha_i &= \sum_{i=1}^n w_j^* \log\left(\frac{w_{i,j}}{w_{i-1,j}}\right) \quad \text{(Def of KL)} \\
&= -\eta \sum_{j=1}^n w_j^* \nabla_j l_i(w_{i-1}) - \log\left(\sum_{j=1}^n w_{i-1,j}\exp\left(-\eta \nabla_j l_i(w_{i-1})\right)\right) \quad \text{(Def of EG Update)} \\
&= -\eta \left\langle w^*, \nabla l_i(w_{i-1})\right\rangle - \log\left(\sum_{j=1}^n w_{i-1,j}\exp\left(-\eta \nabla_j l_i(w_{i-1})\right)\right) \\
&= -\eta \left\langle w^* - w_{i-1}, \nabla l_i(w_{i-1})\right\rangle - \log\left(\sum_{j=1}^n w_{i-1,j}\exp\left(\eta(-\nabla_j l_i(w_{i-1}) + \langle w_{i-1}, \nabla l_i(w_{i-1})\rangle)\right)\right) \\
&\geq \eta\left(l_i(w_{i-1}) - l_i(w^*)\right) - \log\left(\operatorname*{Exp}_{w_{i-1}}\left[\exp\left(Z - \operatorname{Exp}[Z]\right)\right]\right)
\end{aligned}$$

where Z is a discrete-valued random variable defined as

$$Z = -\eta \nabla_j l_i(w_{i-1}) \quad w.p \quad w_{i-1,j}$$

Using the λ-Lipschitz property of l and the l_∞ boundedness of $\mathcal{X}$, one can show that $|Z| \leq \eta\lambda X$ w.p. 1. Then, we have the following bound:

$$\operatorname{Exp}\left[\exp\left(Z - \operatorname{Exp}[Z]\right)\right] \leq \exp\left(\frac{\eta^2\lambda^2 X^2}{2}\right)$$

Therefore,

$$\alpha_i \geq \eta\left(l_i(w_{i-1}) - l_i(w^*)\right) - \frac{\eta^2\lambda^2 X^2}{2}$$

Also,

$$\sum_i \alpha_i = \sum_i \mathrm{KL}\left(w^* \parallel w_{i-1}\right) - \mathrm{KL}\left(w^* \parallel w_i\right) = \mathrm{KL}\left(w^* \parallel w_0\right) - \mathrm{KL}\left(w^* \parallel w_i\right) \leq \mathrm{KL}\left(w^* \parallel w_0\right) \leq \log\left(n\right)$$

Comparing the above bounds, we get

$$\sum_{i=1}^{m}\left(l_i(w_{i-1}) - l_i(w^*)\right) \leq \frac{\log\left(n\right)}{\eta} + \eta\frac{m\lambda^2 X^2}{2}$$

We can choose $\eta = \frac{1}{\lambda X}\sqrt{\frac{2\log(n)}{m}}$ to minimize the RHS and get

$$\sum_{i=1}^{m}\left(l(w_{i-1};(x_i,y_i)) - l(w^*;(x_i,y_i))\right) \leq \lambda X\sqrt{2m\log\left(n\right)}$$

□

References