Advanced Course in Machine Learning Spring 2010

Online Learning with Shifting and Drifting Experts

Handouts are jointly prepared by Shie Mannor and Shai Shalev-Shwartz

In the previous lecture we learned how to track the best expert in an online manner. In particular, we analyzed the regret of the algorithm by comparing its performance to the performance of the best *fixed* experts. In this lecture we allow the competing expert to change over time. We present two types of changes: shift and drift.

1 Shifting experts

In the previous lecture we presented the Weighted Majority algorithm and showed that this algorithm guarantees:

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] \leq \min_i \sum_{t=1}^{T} |f_i^t - y_t| + \sqrt{2\log(d)T}\,,$$

where f_i^t is the prediction of expert i at round t, d is the number of experts, T is the number of rounds, $\hat{y}_t$ is the prediction of the learner at round t, and the expectation is w.r.t. the algorithm's own randomization.

The above bound is meaningful if one of the experts makes a small number of mistakes. In many situations, one expert performs very well on part of the sequence while other experts perform better on other parts of the sequence. In particular, consider the k-shifting regret:

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] - \min_{0=t_1<\ldots<t_{k+1}=T} \sum_{j=0}^{k} \min_i \sum_{t=t_j+1}^{t_{j+1}} |f_i^t - y_t|\,. \tag{1}$$

That is, we allow the competing expert to change k times. Can we have a vanishing regret for this setting?

1.1 A non-efficient solution

A simple way to achieve a bound on the k-shifting regret is by a straightforward reduction to the usual experts setting. Formally, we will construct a meta expert for each sequence $0 = t_1 < \ldots < t_{k+1}$ and a sequence of best expert for each sequence , $i_1, \ldots, i_k$. The meta expert will predict $f_i^t = f_{i_j}^t$ if $t \in (t_j + 1, \ldots, t_{j+1})$.

The number of constructed experts is at most $(dT)^k$. Therefore, using the bound for weighted majority we have constructed in the previous lecture we obtain a regret bound of the from:

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] - \min_{0=t_1<\ldots<t_{k+1}=T} \sum_{j=0}^{k} \min_i \sum_{t=t_j+1}^{t_{j+1}} |f_i^t - y_t| \;\leq\; \sqrt{2\log(dT)kT}\,. \tag{2}$$

The above regret bound tells us that we can have a vanishing regret as long as $k = o(T/\log(T))$. However, the computational complexity of the resulting algorithm is exponential in k.

To derive a better approach we first present a variant of Weighted Majority in which we allow a prior distribution over experts.

1.2 Experts with prior

The algorithm is identical to the Weighted Majority algorithm described in the previous lecture except that the initial value of $\boldsymbol{\theta}^0$ is now initialized to be a vector such that $\mathbf{w}^0$ is some prior distribution over the experts. It is relatively easy to adapt the proof for Weighted Majority to this setting and to get the following.

Theorem 1 *Let $\mathbf{w}^0$ be a prior distribution over experts and let $\boldsymbol{\theta}^0$ be a vector s.t. $w_i^0 = \exp(\theta_i^0)/Z_0$. Then, running the weighted majority algorithm with this initial value of $\boldsymbol{\theta}^0$ gives the regret bound:*

$$\forall i, \quad \mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] - \sum_{t=1}^{T} |f_i^t - y_t| \leq \frac{-\ln(w_i^0)}{\eta} + \frac{\eta T}{8} \ .$$

In particular, if we set $\eta = \sqrt{-8\ln(\lambda)/T}$ for some λ, then for all i s.t. $w_i^0 \geq \lambda$ we have

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] - \sum_{t=1}^{T} |f_i^t - y_t| \leq 4\sqrt{2\ln(1/\lambda)\, T} \ .$$

The proof is left as an exercise.[1]

1.3 An efficient algorithm

Using the Experts-with-prior algorithm, we are now ready to construct an efficient algorithm for shifting experts. The main idea is to define a prior over meta-experts that will allow us to efficiently update $\mathbf{w}^t$ and calculate $\hat{y}_t$, using dynamic programming.

For any sequence of experts, $\mathbf{z} = (z_1, \ldots, z_T) \in [d]^T$, we will construct a meta expert, $f_{\mathbf{z}}$, such that $f_{\mathbf{z}}^t = f_{z_t}^t$. Obviously, the number of meta-experts is d^T which makes the regret bound for Weighted Majority meaningless. However, we will define a prior over meta-experts such that the regret bound of the Experts-with-prior algorithm will be meaningful for any meta-expert that shifts the active expert a small number of times.

The prior we define assumes the sequence of experts is a Markov chain. That is, the initial value of $w_{\mathbf{z}}^0$ is set according to:

$$\mathbb{P}[z_1 = i] = 1/d \quad ; \quad \mathbb{P}[z_{t+1} = i | z_t = j] = \begin{cases} 1-\alpha & \text{if } i = j \\ \alpha/d - 1 & \text{else} \end{cases}$$

where $\alpha \in (0,1)$ will be specified momentarily. This means that if there are k shifts in the sequence $\mathbf{z}$ (namely, k times in which $z_{t+1} \neq z_t$) then

$$-\log(w_{\mathbf{z}}^0) = -\log\left[\frac{1}{d}\left(\frac{\alpha}{d-1}\right)^k (1-\alpha)^{T-k-1}\right] \ .$$

Choosing $\alpha = k/(T-1)$ (which is the minimum of the above over α) yields

$$\begin{aligned} -\log(w_{\mathbf{z}}^0) &= \log(d) + k\log\frac{(d-1)(T-1)}{k} - (T-k-1)\log\left(1 - \tfrac{k}{T-1}\right) \\ &\leq \log(d) + k\log\frac{(d-1)(T-1)}{k} + k \ , \end{aligned}$$

where the last inequality is because for any $0 < a < b$ we have

$$-(a-b)\log(1-\tfrac{b}{a}) = -(a-b)\log\tfrac{a-b}{a} = (a-b)\log\tfrac{a}{a-b} \leq (a-b)\left(\tfrac{a}{a-b} - 1\right) = b \ .$$

Overall, by applying the regret bound of Experts-with-prior we obtain the regret bound

$$\mathbb{E}\left[\sum_{t=1}^{T} |\hat{y}_t - y_t|\right] - \min_{0=t_1<\ldots<t_{k+1}=T} \sum_{j=0}^{k} \min_i \sum_{t=t_j+1}^{t_{j+1}} |f_i^t - y_t| \ \leq \ O\left(\sqrt{k\log(dT/k)T}\right) \ , \tag{3}$$

[1] Hint: Follow the proof of the regret bound for the Weighted Majority, but instead of the strongly convex function: $f(\mathbf{w}) = \sum_i w_i \log(w_i) + \log(n)$ use the strongly convex function: $g(\mathbf{w}) = f(\mathbf{w}) + \langle \mathbf{w}, \mathbf{w}^0 \rangle$. Show that $g^\star(\theta) = g^\star(\theta - \theta^0)$.

which is similar to the bound we have derived for the non-efficient solution.

It is left to show that it is possible to run the Experts-with-prior algorithm efficiently. To make the prediction on round t it suffices to know the total weight of meta-experts for which $z_t = i$. Denote this quantity by $Q_{t-1}(i)$. Formally,

$$Q_{t-1}(i) = \sum_{\mathbf{z}:z_t=i} \exp(\theta^t_{\mathbf{z}}) \ .$$

Using $Q_{t-1}(i)$ we can efficiently calculate the prediction $\hat{y}_t$ by simply choosing i_t according to the distribution $Q_{t-1}(i)/\sum_j Q_{t-1}(j)$ and returning $f^t_{i_t}$. We will show how to update Q efficiently.

Initially, $Q_0(i) = 1/d$ because $\boldsymbol{\theta} = \log(\mathbf{w}^0)$ is set according to the Markov model and the initial state is chosen uniformly at random. Also note that for any t we have that

$$\sum_{\mathbf{z}:z_{t+1}=i \wedge z_t=i} \exp(\theta^0_{\mathbf{z}}) = \mathbb{P}[z_{t+1} = z_t = i] = \mathbb{P}[z_t = i]\,\mathbb{P}[z_{t+1} = i | z_t = i] = Q_{t-1}(i)\,(1-\alpha) \ .$$

Similarly, for $j \neq i$ we have

$$\sum_{\mathbf{z}:z_{t+1}=i \wedge z_t=j} \exp(\theta^0_{\mathbf{z}}) = \mathbb{P}[z_t = i]\,\mathbb{P}[z_{t+1} = J | z_t = i] = Q_{t-1}(i)\,\frac{\alpha}{d-1} \ .$$

Recall that the update of $\boldsymbol{\theta}$ is $\theta^t_{\mathbf{z}} = \theta^{t-1}_{\mathbf{z}} - \eta|f^t_{z_t} - y_t|$. Therefore, the above two equalities will hold for $t > 0$ as well. Now, suppose that at round t we have the correct value of $Q_{t-1}(i)$ and let us calculate $Q_t(i)$. We have:

$$\begin{aligned}
Q_t(i) &= \sum_{\mathbf{z}:z_{t+1}=i} \exp(\theta^t_{\mathbf{z}}) \\
&= \sum_{\mathbf{z}:z_{t+1}=i \wedge z_t=i} \exp(\theta^t_{\mathbf{z}}) + \sum_{j \neq i} \sum_{\mathbf{z}:z_{t+1}=i \wedge z_t=j} \exp(\theta^t_{\mathbf{z}}) \\
&= \sum_{\mathbf{z}:z_{t+1}=i \wedge z_t=i} \exp(\theta^{t-1}_{\mathbf{z}} - \eta|f^t_i - y_t|) + \sum_{j \neq i} \sum_{\mathbf{z}:z_{t+1}=i \wedge z_t=j} \exp(\theta^{t-1}_{\mathbf{z}} - \eta|f^t_j - y_t|) \\
&= \exp(-\eta|f^t_i - y_t|) \sum_{\mathbf{z}:z_{t+1}=i \wedge z_t=i} \exp(\theta^{t-1}_{\mathbf{z}}) + \sum_{j \neq i} \exp(-\eta|f^t_j - y_t|) \sum_{\mathbf{z}:z_{t+1}=i \wedge z_t=j} \exp(\theta^{t-1}_{\mathbf{z}}) \\
&= \exp(-\eta|f^t_i - y_t|)(1-\alpha)Q_{t-1}(i) + \sum_{j \neq i} \exp(-\eta|f^t_j - y_t|)\,\tfrac{\alpha}{d-1} Q_{t-1}(j) \ .
\end{aligned}$$

That is, Q_t can be calculated based on Q_{t-1} in time $O(d)$ by calculating once $\sum_j \exp(-\eta|f^t_j - y_t|)Q_{t-1}(j)$ and then for each i subtracting from the above the i'th summand, and multiplying the two terms appropriately.

2 Drifting hypothesis

The regret again k-shifting experts makes sense in the experts setting, where changes are discrete. In the general online convex optimization problem we studied in the previous section changes in $\mathbf{w}_t$ are continuous. In this case, it also makes sense to talk about a hypothesis drift. One way to define such a drift is by defining the total change of a sequence of vectors:

$$\sum_{t=1}^{T} \|\mathbf{u}_t - \mathbf{u}_{t-1}\| \ ,$$

where for convenience define $\mathbf{u}_0$ to be the zero vector. Now, we can define a regret bound with respect to a sequence of vectors:

$$\sum_{t=1}^{T} g_t(\mathbf{w}_t) - \min_{\mathbf{u}_1,\ldots,\mathbf{u}_T : \max_t \|\mathbf{u}_t\| \leq U,\ \sum_{t=1}^{T} \|\mathbf{u}_t - \mathbf{u}_{t-1}\| \leq S} g_t(\mathbf{u}_t) \ .$$

Zinkevich derived a regret bound for the above that takes the form:

$$O((U + \sqrt{US})\sqrt{T}) \ .$$

The interesting reader is referred to Zinkevich's paper: "Online convex programming and generalized infinitesimal gradient ascent" from ICML 2003.