Online-to-batch conversions

Lecturer: Sasha Rakhlin *Scribe: Dapo Omidiran, Adam Pauls, Jacob Abernethy*

1 Intro

Much of today's lecture comes from the paper [2].

Recap: We've shown that low regret algorithms, where regret $\mathcal{R}_T$ is defined as

$$\mathcal{R}_T := \sum_{t=1}^{T} \ell_t(x_t) - \min_{x\in\mathcal{K}} \sum_{t=1}^{T} \ell_t(x)$$

can be obtained by, for each t, finding

$$x_{t+1} = \arg\min_{x\in\mathcal{K}} R(x) + \eta \sum_{s=1}^{t} \ell_s(x)$$

where $R(x)$ is some reguarlizing function (not to be confused with regret $\mathcal{R}$ or risk $R(f)$).

Suppose $\{(x_1, y_1), \ldots, (x_T, y_T)\} = z_1^T$ iid, and we wish to find a function f that predicts y given x (the standard classification setting). In terms of our low regret algorithms, we can think of x as a hypothesis f and $\ell_t(x)$ as $\ell(f(x_t), y_t)$. Then, a plausible training procedure is to run an online algorithm on this sequence, obtaining $f_1, \ldots, f_T$. Let's look at the regret:

$$\frac{1}{T}\mathcal{R}_T = \frac{1}{T}\sum_{t=1}^{T} \ell(f_{t-1}(x_t), y_t) - \min_{f\in\mathcal{K}} \frac{1}{T}\sum_{t=1}^{T} \ell(f(x_t), y_t)$$

We have shown that this regret is "small" ($\mathcal{O}(\frac{1}{\sqrt{T}})$). However, in the classification setting, we would like to find small expected risk $R(f)$ among $f_1, f_2, \ldots, f_T$.

Define:

$$\begin{aligned}
R(f) &:= \mathbb{E}[\ell(f(X), Y)] \\
\hat{R}(f) &:= \frac{1}{T}\sum_{i=1}^{T} \ell(f(X_i), Y_i) \\
\bar{f} &:= \frac{1}{T}\sum_{t=1}^{T} f_t \\
f^* &:= \arg\min_{f\in\mathcal{K}} R(f)
\end{aligned}$$

If $\ell(\hat{y}, y)$ is convex in $\hat{y}$, and $0 \leq \ell(\cdot,\cdot) \leq 1$, then:

$$\begin{aligned}
R(\bar{f}) &\leq \frac{1}{T}\sum_{t=1}^{T} R(f_t) && \text{(by convexity)} \\
&\leq \frac{1}{T}\sum \ell(f_t(x_t), y_t) + \sqrt{\frac{2}{T}\log(\frac{1}{\delta})} && \text{(w.h.p. - see next lemma)} \\
&\leq \min_{f\in K}\frac{1}{T}\sum_{t=1}^{T} \ell(f(x_t), y_t) + \frac{\mathcal{R}_T}{T} + \sqrt{\frac{2}{T}\log(\frac{1}{\delta})} && \text{(due to regret bound)} \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \ell(f^*(x_t), y_t) + \frac{\mathcal{R}_T}{T} + \sqrt{\frac{2}{T}\log(1/\delta)} && \text{(assuming } f^* \text{ is optimal)} \\
&\leq R(f^*) + \frac{\mathcal{R}_T}{T} + 2\sqrt{\frac{2}{T}\log(1/\delta)} && \text{(by def'n of risk)}
\end{aligned}$$

The second inequality is true by the following Lemma:

Lemma 1.1. Define $M_T = \frac{1}{T}\sum_{t=1}^{T} \ell(f_{t-1}(x_t), y_t)$. Then

$$\mathbb{P}\left[\frac{1}{T}\sum_{t=1}^{T} R(f_{t-1}) \leq M_T + \sqrt{\frac{2}{T}\log\left(1/\gamma\right)}\right] \geq 1-\delta \tag{1}$$

Proof. (Using Martingale's) Define

$$V_{t-1} := R(f_{t-1}) - \ell(f_{t-1}(x_t), y_t)$$

Then

$$\frac{1}{T}\sum_{t=1}^{T} V_{t-1} = \frac{1}{T}\sum R(f_{t-1}) - M_T$$

and $-1 \leq V_{t-1} \leq 1$.

If $\mathbb{E}_t[\cdot] = E[\,\cdot\,|(X_1 = x_1, Y_1 = y_1), \ldots, (X_{t-1} = x_{t-1}, Y_{t-1} = y_{t-1})]$. Then:

$$\mathbb{E}_t[V_{t-1}] = R(f_{t-1}) - \mathbb{E}_t[\ell(f_{t-1}(X_{t-1}), Y_t)] = 0$$

by definition of $R(f_{t-1})$. Therefore, V_t forms Martingale sequence. Since $-1 \leq \frac{1}{T}\sum_{t=1}^{T} V_{t-1} \leq 1$, we can apply Azuma-Hoeffding:

$$P\left(\frac{1}{T}\sum_{t=1}^{T} V_{t-1} - \mathbb{E}_t\left[\frac{1}{T}\sum_{t=1}^{T} V_{t-1}\right] > \epsilon\right) = P\left(\frac{1}{T}\sum_{t=1}^{T} V_{t-1} - 0 > \epsilon\right) \leq \exp\left(\frac{-\epsilon^2 T}{2}\right)$$

Note that more details can be found in in the proof McDiarmid Inequality from Lecture 13, where a similar Martingale sequence was constructed. $\square$

The above analysis assumes $\ell(\cdot, y)$ is convex in the first argument. If not, then we can instead use the following "cross-validation" scheme:

$$\begin{aligned}
\text{Define } R(f_t, t+1) &:= \frac{1}{T-t} \sum_{s=t+1}^{T} \ell(f_t(x_s), y_s). \\
\text{Set } c_\delta(t) &:= \sqrt{\frac{1}{2t} \log \frac{2T(T+1)}{\delta}}. \\
\text{Let } \hat{f} &:= \arg \min_{0 \le t \le T} [\hat{R}(f_t, t+1) + c_\delta(T-t)]
\end{aligned}$$

Theorem 1.2. Under some additional assumptions, it can be shown that

$$P\left(R(\hat{f}) \ge M_T + \delta \sqrt{\frac{1}{T} \log \frac{2(T+1)}{\delta}} \right) \le \delta$$

1.1 Example: Kernel Perceptron Classification

Suppose we have a RKHS $\mathcal{H}_K$ with kernel K, and have the 0/1 loss function as our criterion. Then:

$$\not\Vdash_{sign(\hat{y})=y} \le \ell_\gamma(\hat{y}, y) := \max\{0, 1 - \frac{y\hat{y}}{\gamma}\}.$$

Here ℓ_γ is the hinge-loss with x-intercept γ.

Kernel Perceptron gives: $f_t = sign(\sum_{s \in \mu_t} y_s K(x_s, \cdot))$, where μ_t is set of indices of mistakes up to t.

Theorem 1.3. Let $f_0, \dots, f_{T-1}$ be generated by kernel perceptron on Z_1^T and $\hat{f}$ as designed before. Then

$$R(sign(\hat{f})) \le \inf_{f \in \mathcal{H}_k, ||f||_* < 1, \delta > 0} \left\{ \frac{1}{T} \sum_{t=1}^{T} \ell_\gamma(f(x_{t-1}), y_t) + \frac{1}{\gamma T} \sqrt{\sum_{t \in \mu_T} K(x_t, x_t)} + \delta \sqrt{\frac{1}{T} \log \frac{2(T+1)}{\delta}} \right\}$$

with probability exceeding $1 - \delta$.

Now, we will compare the above to a similar result obtained in [1]. First we make the following definitions:

$$\tilde{\ell}_\gamma(\hat{y}, y) := \min\{1, \ell_\gamma(\hat{y}, y)\}$$

$$\tilde{D}_{\gamma,T}(f, Z_1^T) := \frac{1}{T} \sum_{t=1}^{T} \tilde{\ell}_\gamma(f(x_{t-1}), y_t)$$

The following result is proved [1]:

Lemma 1.4. With probability at least $1 - \delta$:

$$R(sign(F)) \le \tilde{D}_{\gamma,T}(F, Z_1^T) + \frac{4B}{\gamma T} \sqrt{\sum K(x_t, x_t)} + \left(\frac{8}{\gamma} + 1 \right) \sqrt{\frac{1}{T} \log(\frac{2(T+1)}{\delta}}$$

Simultaneously for all F of the form $F(\cdot) = \sum_{t=1}^{T} \alpha_t K(x_t, \cdot)$ and coefficients $\alpha_1, \alpha_2, \dots, \in \mathbb{R}$ such that $\sum_{i,j} \alpha_i \alpha_j K(x_i, x_i) \le B^2$.

The two results are arguably similar, yet the former requires much less machinery and arises from analyzing regret, i.e. when we are learning against an adversary.

References

[1] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002. http://www.jmlr.org/papers/volume3/bartlett02a/bartlett02a.pdf.

[2] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions on*, 50(9):2050–2057, Sept. 2004.