Generalization Ability of Online Strongly Convex Learning Algorithms

John Wieting

December 21, 2013

1 Introduction

Online learning, in contrast to batch learning, occurs in a sequence of rounds. At the beginning of a round, an example is presented to the learning algorithm, the learning algorithm uses its current hypothesis to label the example, and then the learning algorithm is presented with the correct label and the hypothesis is updated. It is a different learning paradigm than batch learning where we are given all of our data at once, and we aim to construct a single optimal hypothesis using the entire data set. We hope that the resulting hypothesis will generalize well to unseen data. In online learning, our goal is to minimize the total loss along the entire sequence of training examples and we generate a new hypothesis with nearly every training example.

Online learning can be motivated from situations where it is not feasible or desirable to utilize a batch learning approach. Examples could be situations where there is a huge amount of data where storing and learning from all of it is computationally unfeasible, or perhaps when the distribution generating the data is changing i.e. during the sequence more than one hypothesis is generating the data.

Recently, statistical learning machinery has been used to analyze this paradigm. In [1] the generalization ability of convex functions was analyzed and [2] extends this work by investigating online algorithms with strongly convex loss functions. This analysis can be motivated due to the fact that there exists a large number of optimization problems in machine learning that are strongly convex. For instance all problems that use a log-loss or square-loss loss function or those who use a convex loss function, that is not necessarily strongly convex, and use L_2 regularization or another strongly convex regularizer. The latter case describes the SVM problem which uses a convex loss function (hinge loss) with L_2 regularization.

This paper will discuss [2] and examine the paper through the lens of our CS 598 course. It will also discuss the main application of this paper, which is bounding the convergence rate of the Pegasos algorithm [4], with high probability.

2 Background

A major theme of our course has been bounding the generalization error of the output functions of binary classification algorithms. Ideally, we want the output of our learning algorithm $\widehat{f}_n$ from a function class $\mathscr{F}$ to have similar generalization error to that of the optimal function, $f^* \in \mathscr{F}$. More simply, we want:

$$P(\widehat{f}_n(X) \neq Y) \approx P(f^*(X) \neq Y)$$

We started exploring bounds to quantify this relationship by trying to construct an algorithm that is PAC - which informally means that the probability of generating a bad enough sample that the deviation between $L(f)$ and $L(f^*)$ exceeds ϵ $\forall \epsilon > 0$ goes to 0 as the number of examples approaches ∞. L refers to the expected $0,1$ loss. A good starting place was the ERM algorithm, an algorithm that finds the function $\widehat{f}_n$ that has minimal error on our training set of examples. While intuitively appealing, we needed to specify the conditions under which ERM would be PAC. We were able to show that if some property, known as the Uniform Convergence of Empirical Means, holds than ERM is PAC. The UCEM states that $\forall \epsilon > 0$:

$$\lim_{n\to\infty} \sup_{P\in\mathscr{P}} P(Z^n \in \mathscr{Z}^n : \sup_{f\in\mathscr{F}} \|L(\widehat{f}) - L(f)\| \geq \epsilon) = 0$$

The desire to determine when the ERM is PAC without having to directly check the UCEM property led us to explore the Rademacher average. We were able to show that much of the time we can bound this deviation (known as the standard deviation) by the expectation of the Rademacher. This in turn can often be bounded by

$$C\frac{\sqrt{V(\mathscr{F})}}{\sqrt{n}}$$

Where $V(\mathscr{F})$ is the VC dimension and C is a universal constant. Thus if $\mathscr{F}$ is a VC class (VC dimension less than ∞) then the UCEM property will hold. This analysis also led to a powerful result that holds for functions $f : Z \to [0,1]$ and any distribution from $\mathscr{P}$:

$$L(\widehat{f}_n) \leq L(f^*) + 4ER_n(\mathscr{F}(Z^n)) + \sqrt{\frac{2\log(\frac{1}{\delta})}{n}}$$

Here ER_n is the expected Rademacher average of the function class.

One issue with the above analysis is that we do not have much hope of finding the ERM solution for many interesting function classes (unless $P = NP$). Thus we learned we could replace the $0,1$ loss with another loss function - one that could make finding the ERM more tractable. Generally, we lose then the ability to compare the minimum risk with the generalization error, but it gives us the ability to be clever with our surrogate loss functions to make finding the ERM tractable. We also can still bound the generalization error if the loss function is Lipschitz and bounded.

Further analysis into the properties of surrogate loss functions led us to wonder what can happen if we choose a convex loss function. It turns out that with such a loss function we can show that the unique minimizer of the surrogate loss over all real valued functions on the domain of X, has the property that its sign is equal to the the Bayes classifier - the optimal choice for binary classification. We also were able to bound $L(\widehat{f}_n) - L^*$, where $\widehat{f}_n$ is the empirical surrogate loss minimizer, if we can relate the minimum surrogate loss to the Bayes rate. In practice this can often be done for commonly used loss functions [5].

The common thread through all of this analysis on binary classification, is that we are aiming to achieve generalization bounds by finding the hypothesis that minimizes the loss of some random training set. In online learning, as mentioned previously, our goals are slightly different. We want to minimize the loss along the entire sequence of examples that we are given. More specifically we want the total accumulated loss at the end of the sequence to be close to that of the optimal function in our function class $\mathscr{F}$. This is known as the regret:

$$\sum_{t=1}^{T} l_t(f_t) - \min_{f\in\mathscr{F}} \sum_{t=1}^{T} l_t(f)$$

Interestingly, it turns out that we can use the regret to achieve bounds on the excess risk $R(\widehat{f}) - R(f^*) = E[l(\widehat{f}, Z)] - E(l(f^*, Z)]$ in what is known as an online-to-batch conversion. Thus achieving low regret can be seen as obtaining better generalization - an intuitively appealing notion that illustrates the relationship between online and batch learning paradigms. In these conversions, we must choose a single hypothesis from all of the hypotheses that were generated along the sequence. A common approach in the literature it to average the hypotheses and that is what is done in this paper. An analysis on alternative approaches - like picking the last one or using validation can be found in [3]. One last thing that is important to note about this discussion is that L has been referring to the expectation of the $0,1$ loss which also is equivalent to $P(f(X) \neq Y)$. In this paper the loss function in the risk is not the $0,1$ loss and so $R(f)$ will be as large or larger than $L(f)$ - assuming the expectations are taken with respect to the same distributions. Thus we are not bounding the uniform deviation as we did for most of the bounds for binary classification. We did study though one bound that related the $0,1$ loss to that of the risk with respect to a certain class of loss functions - which fall under those that are studied in this paper. For the analysis done in this paper, bounding the excess risk is appropriate as it allows one to determine the convergence rate of the algorithm. By that I mean, how many examples must we see until we can, with high probability, output a hypothesis whose risk is within ϵ of the optimal f^*.

3 Results

The main result of this paper is a single theorem that relates the excess risk to the regret. It relies on several corollaries and a few assumptions. I will start by introducing some notation and then present the main results with some commentary.

Let $f : S \times Z \rightarrow [0.B]$ where Z is a random variable and S is a convex set with a norm $\|\cdot\|$. Furthermore, let $F(\mathbf{w}) = E[f(\mathbf{w}, Z)]$ and $\{\mathbf{w}^* = \text{argmin}_{\mathbf{w} \in S} F(\mathbf{w})$. Also let $Z_1, ..., Z_T$ be a sequence of independent random variables, then $F(\mathbf{w} = E[f(\mathbf{w}, Z)] = E_{t-1}[f(\mathbf{w}, Z_t)]$ where E_{t-1} refers to the conditioal expectation of the first $t-1$ random variables in the sequence.

The assumptions, known as LIST, have two components. The first is that $f(\mathbf{w}, z)$ is convex in $\mathbf{w}$ and has a Lipschitz constant L with respect to $\|\cdot\|$. Thus by definition, $\forall \mathbf{w}, \mathbf{w}' \in S, \|f(\mathbf{w}, Z) - f(\mathbf{w}', Z)\| \leq L\|\mathbf{w} - \mathbf{w}'\|$. The second assumption is that $f(\mathbf{w}, z)$ is strongly convex with respect to $\|\cdot\|$.

Lastly to make the following discussion less verbose we need to define two equations and a martingale difference sequence:

$$\text{Reg}_T = \sum_{i=1}^{T} f(\mathbf{w_t}, Z_t) - \text{min}_{\mathbf{w} \in S} \sum_{t=1}^{T} f(\mathbf{w}, Z_t)$$

$$\text{Diff}_T = \sum_{i=1}^{T} (F(\mathbf{w_t}) - F(\mathbf{w}^*)$$

$$\zeta_t = F(\mathbf{w}) - F(\mathbf{w}^*) - (f(\mathbf{w_t}, Z_t) - f(\mathbf{w}^*, Z_t))$$

Note that this is clearly a martingale difference sequence because due to the relationship between f and F, ζ_t vanishes when E_{t-1} is applied.

We are now ready to derive the main result. The first step to achieving the main result of this paper is that we need to bound the ζ.

Lemma 3.1 *Suppose the LIST assumptions hold and let ζ_t be the martingale sequence previously defined. Then we have*

$$Var_{t-1}\zeta_t = E_{t-1}[\zeta_T^2] \leq \frac{4L^2}{v}(F(\mathbf{w_t}) - F(\mathbf{w}^*))$$

where v refers to the strong convexity parameter and L refers to the Lipschitz constant.

Proof: The proof is fairly straightforward. The first step is key and then everything else follows from definitions. This step is to notice that

$$E_{t-1}[\zeta_T^2] \leq E_{t-1}[(f(\mathbf{w_t}, Z_t) - f(\mathbf{w}, Z_t))^2]$$

This can be seen by expanding $E_{t-1}[\zeta_T^2$, taking expectations and canceling terms leaving a nonpositive term and the right side of the inequality. Then we have:

$$E_{t-1}[(f(\mathbf{w_t}, Z_t) - f(\mathbf{w}, Z_t))^2] \leq E_{t-1}[L^2\|\mathbf{w_t} - \mathbf{w}^*\|^2] = L^2\|\mathbf{w_t} - \mathbf{w}^*\|^2$$

Where the first inequality stems from the LIST assumption (and that $E[X] \geq E[Y]$ if $Y \geq X$), and the last equality comes from the fact that we have eliminated the random variable. Now due to the strong convexity assumption we have for any $\mathbf{w}$ and $\mathbf{w}' \in S$:

$$\frac{f(\mathbf{w}, Z) + f(\mathbf{w}', Z)}{2} \geq f(\frac{\mathbf{w} + \mathbf{w}'}{2}, Z) + \frac{v}{8}\|\mathbf{w} - \mathbf{w}'\|$$

This can be seen by taking $\theta = \frac{1}{2}$ since the strong convexity inequality holds for any $\theta \in [0, 1]$. Then just take the expectations and substitute in $\mathbf{w_t}$ and $\mathbf{w}^*$:

$$\frac{F(\mathbf{w_t}) + F(\mathbf{w}^*)}{2} \geq F(\frac{\mathbf{w_t} + \mathbf{w}^*}{2}) + \frac{v}{8}\|\mathbf{w_t} - \mathbf{w}^*\| \geq F(\mathbf{w}^*) + \frac{v}{8}\|\mathbf{w_t} - \mathbf{w}^*\|$$

Where the last inequality stems from the w^* being the minimizer of F. Finally rearranging gives:

$$\|\mathbf{w_t} - \mathbf{w}^*\| \leq \frac{4(F(\mathbf{w_t}) + F(\mathbf{w}^*))}{v}$$

The final inequality then comes from combining the results. ■

The next lemma is the longest proof in the paper. I will not detail the proof. The proof relies on Freedman's inequality - which interestingly is a Martingale extension to Bernstein's inequality. Essentially, the argument follows a similar structure to some of the arguments used in the course where we can prove some bound by messaging terms and substituting variables so that a known concentration bound can be used. This approach was used, for instance, in the last assignment where we were able to use Bernstein's inequality after a little thought. The massaging in this case relies on a carefully chosen discretization.

Lemma 3.2 *Suppose X_i is a martingale difference sequence with $\|X_i\| \leq b$. Let*

$$Var_t X_t = Var(X_t \| X_1, ..., X_{t-1})$$

Also let $V = \sum_{i=1}^T Var_t X_t$. Also let $\sigma = \sqrt{V}$. Then for any $\delta < 1/e$ and $T \geq 3$:

$$P(\sum_{t=1}^T X_t > max\{2\sigma, 3b\sqrt{\log(1/\delta)}\sqrt{\log(1/\delta}\}) \leq 4\log(T)\delta$$

The last lemma is also the simplest

Lemma 3.3 *Suppose* $s, r, d, b, \Delta \geq 0$ *and*

$$s - r \leq max\{4\sqrt{ds}, 6b\Delta\}\Delta$$

Then

$$s \leq r + 4\sqrt{dr}\Delta + max\{16d, 6b\}\Delta^2$$

Proof: This proof is very simple. We just show that if

$$s - r - 4\sqrt{ds}\Delta \leq 0$$

Then:

$$(\sqrt{s})^2 - r - 4\sqrt{ds}\Delta \leq 0$$

Therefore $\sqrt{s}$ is less than the largest root of this equation (and should be bigger than the smallest root since it is a parabola). Finding the root, simplifying, and putting the resulting inequalities together gives the desired result. ■

Armed with these 3 lemmas we can now prove the main result and only theorem in the paper.

Theorem 3.4 *Assume the LIST assumptions we have with probability at least* $1 - 4\log(T)\delta$*:*

$$\frac{1}{T}\sum_{t=1}^{T} F(\mathbf{w_t}) - F(\mathbf{w}^*) \leq \frac{Reg_T}{T} + 4\sqrt{\frac{L^2 \log 1/\delta}{v}}\frac{\sqrt{Reg_T}}{T} + max\{\frac{16L^2}{v}, 6B\}\frac{\log(1/\delta)}{T}$$

Proof: We know that from the first lemma:

$$\mathrm{Var}_{t-1}\zeta_t = E_{t-1}[\zeta_T^2] \leq \frac{4L^2}{v}(F(\mathbf{w_t} - F(\mathbf{w}^*))$$

Thus:

$$\sqrt{\sum_{t=1}^{T} \mathrm{Var}_t \zeta_t} \leq \sqrt{\frac{4L^2}{v}\mathrm{Diff}_T}$$

We can then apply the third lemma as $\|\zeta_t\| \leq 2B$ as f is bounded by B (an assumption we made in the beginning that f maps to $[0, B]$) and the ζ_t forms a martingale difference sequence. Thus we have with probability $1 - \delta$:

$$\sum_{t=1}^{T} \zeta_t \leq \max\{2\sigma, 6B\sqrt{\log(1/\delta)}\}\sqrt{\log(1/\delta}$$

Since by definition we have:

$$\mathrm{Diff}_T - \mathrm{Reg}_T \leq \sum_{t=1}^{T} \zeta_t$$

and also by Lemma 3.1 we have:

$$\sigma \leq \sqrt{\frac{4L^2}{v}\text{Diff}_T}$$

gives us:

$$\text{Diff}_T - \text{Reg}_T \leq \max\{2\sqrt{\frac{4L^2}{v}\text{Diff}_T}, 6B\sqrt{\log(1/\delta)}\}\sqrt{\log(1/\delta}$$

Lastly, we apply Lemma 3.3 to achieve the result. ■

I will close this section with three trivial corollaries that use this theorem. These stem from instantiating the very general situation above. If we define our Z_i to be pairs (X_i, Y_i), our loss function $l : D \times Y \rightarrow [0, 1]$ for some space D and our hypothesis $h : X \times S \rightarrow D$ or $h(X, \mathbf{w})$. Now instantiate f to be l and assume l satisfies the LIST assumptions. Lastly, set $R(\mathbf{w}) = E[l(h(X, \mathbf{w}), Y)]$ to be the risk. The last corollary deserves special mention as here we have bounded the excess risk by the regret. This inequality will be used to obtain the convergence rate for Pegasos.

Corollary 3.5 *Assume the LIST assumptions we have with probability at least* $1 - 4\log(T)\delta$

$$F(\frac{1}{T}\sum_{t=1}^{T}\mathbf{w_t}) - F(\mathbf{w}^*) \leq \frac{Reg_T}{T} + 4\sqrt{\frac{L^2\log 1/\delta}{v}}\frac{\sqrt{Reg_T}}{T} + max\{\frac{16L^2}{v}, 6B\}\frac{\log(1/\delta}{T}$$

Proof: Let $\widehat{\mathbf{w}} = \sum_{t=1}^{T}\mathbf{w_t}$. Note that since f is convex $f(\widehat{\mathbf{w}}) \leq \sum_{t=1}^{T} f(\mathbf{w_t})$ therefore $F(\widehat{\mathbf{w}}) \leq \sum_{t=1}^{T} F(\mathbf{w_t})$. ■

Corollary 3.6 *Suppose the LIST assumptions hold for* $l(h(x, \mathbf{w}, y)$ *then with probability at least* $1 - 4\log(T)\delta$*:*

$$R(\widehat{\mathbf{w}}) - R(\mathbf{w}^*) \leq \frac{Reg_T}{T} + 4\sqrt{\frac{L^2\log 1/\delta}{v}}\frac{\sqrt{Reg_T}}{T} + max\{\frac{16L^2}{v}, 6B\}\frac{\log(1/\delta}{T}$$

Proof: Follows from the last corollary by substituting R for F. ■

4 Applications

Pegasos [4] is an iterative algorithms for solving the SVM optimization problem. The run-time of the algorithm does not depend on the size of the data-sets and can be used in an online fashion. This has appeal in the machine learning community because this algorithm has the generalization properties of the SVM, but doesn't require expensive batch training. The original paper did not characterize the convergence in terms of the excess risk. However with the tools that were built up in the previous section, this is now possible. All that needs to be done is to ensure that the objective function of Pegasos fits our assumptions and then substitute the correct variables. We assume that $\|x_i\| < R \; \forall i$ and that in total we have m examples. The objective function is:

$$f(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{k}\sum_{(x,y)\in Z_i} l(\mathbf{w}, (x, y))$$

Taking expectations gives us:

$$F(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{m}\sum_{i=1}^{m} l(\mathbf{w}, (x_i, y_i))$$

To arrive at this, note that:

$$E[l(\mathbf{w}, Z_i)] = \frac{1}{m}\sum_{i=1}^{m} l(\mathbf{w}, (x_i, y_i))$$

Therefore:

$$E[f(\mathbf{w})] = \frac{\lambda}{2}\|\mathbf{w}\|^2 + \frac{1}{k}\sum_{(x,y)\in Z_i} E[l(\mathbf{w}, Z_i)] = \frac{\lambda}{2}\|\mathbf{w}\|^2 + E[l(\mathbf{w}, Z_i)] = F(\mathbf{w})$$

We can also easily show that f is Lipschitz with $L = \sqrt{\lambda} + R$ and strongly convex where $v = \lambda$. Also f maps into $[0, B]$ where $B = \frac{3}{2} + \frac{R}{\sqrt{\lambda}}$. Thus all of the LIST assumptions are satisfied and since [4] showed that:

$$\text{Reg}_T \leq \frac{L^2(1 + \log T)}{2v}$$

We can plug this into Corollary 3.6 and obtain with probability at least $1 - \delta$:

$$R(\widehat{\mathbf{w}}) - R(\mathbf{w}^*) = O(\frac{\log \frac{T}{\delta}}{\lambda T})$$

Assuming $R = 1$ and for small enough λ.

5 Conclusion

Online learning algorithms are an important tool in the analysis of extremely large data sets - a task that is becoming increasingly common. The analysis in this paper applies to a special case of online algorithms - those that satisfy the LIST assumptions. However, in practice this is not overly restrictive as many commonly used algorithms satisfy these conditions. The work showed how to bound the excess risk of an online learning algorithm satisfying these assumptions and this bound was used to characterize the convergence of these algorithms.

This paper is closely related to the material we learned in CS 598 as many of the techniques and ideas that we learned during our study on binary classification carried over - even though online learning and batch learning require a different analysis. This particular problem can be seen as an extension of what we learned and it is also a clear of example of the applicability of statistical learning

References

[1] Nicolo Cesa-Bianchi, Alex Conconi, and Claudio Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50:2050–2057, 2001.

[2] Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms.

[3] Shai Shalev Schwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, Hebrew University, 2007.

[4] Yoram Singer and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *In ICML*, pages 807–814, 2007.

[5] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32:56–134, 2003.