Online Learning, Stability, and Stochastic Gradient Descent

May 26, 2011

Tomaso Poggio, Stephen Voinea, Lorenzo Rosasco

CBCL, McGovern Institute, CSAIL, Brain Sciences Department, Massachusetts Institute of Technology

Abstract

In batch learning, stability together with existence and uniqueness of the solution corresponds to well-posedness of Empirical Risk Minimization (ERM) methods; recently, it was proved that CV_{loo} *stability* is necessary and sufficient for generalization and consistency of ERM ([2]). In this note, we introduce CV_{on} *stability*, which plays a similar role in online learning. We show that stochastic gradient descent (SDG) with the usual hypotheses is CV_{on} stable and we then discuss the implications of CV_{on} stability for convergence of SGD.

This report describes research done within the Center for Biological & Computational Learning in the Department of Brain & Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research was sponsored by grants from: AFSOR, DARPA, NSF. Additional support was provided by: Honda R&D Co., Ltd., Siemens Corporate Research, Inc., IIT, McDermott Chair.

Contents

1 Learning, Generalization and Stability

In this section we collect some basic definition and facts.

1.1 Basic Setting

Let Z be a probability space with a measure ρ. A training set S_n is an i.i.d. sample z_i, $i, = 0, \dots, n-1$ from ρ. Assume that a hypotheses space $\mathcal{H}$ is given. We typically assume $\mathcal{H}$ to be a Hilbert space and sometimes a p-dimensional Hilbert Space, in which case, without loss of generality, we identify elements in $\mathcal{H}$ with p-dimensional vectors and $\mathcal{H}$ with $\mathbb{R}^p$. A loss function is a map $V : \mathcal{H} \times Z \to \mathbb{R}_+$. Moreover we assume that

$$I(f) = \mathbb{E}_z V(f, z),$$

exists and is finite for $f \in \mathcal{H}$. We consider the problem of finding a minimum of $I(f)$ in $\mathcal{H}$. In particular, we restrict ourselves to finding a minimizer of $I(f)$ in a closed subset K of $\mathcal{H}$ (note that we can of course have $K = \mathcal{H}$). We denote this minimizer by f_K so that

$$I(f_K) = \min_{f \in K} I(f).$$

Note that in general, existence (and uniqueness) of a minimizer is not guaranteed unless some further assumptions are specified.

Example 1. *An example of the above set is supervised learning. In this case X is usually a subset of $\mathbb{R}^d$ and $Y = [0, 1]$. There is a Borel probability measure ρ on $Z = X \times Y$. and S_n is an i.i.d. sample $z_i = (x_i, y_i)$, $i, = 0, \ldots, n-1$ from ρ. The hypotheses space $\mathcal{H}$ is a space of functions from X to Y and a typical example of loss functions is the square loss $(y - f(x))^2$.*

1.2 Batch and Online Learning Algorithms

A batch learning algorithm A maps a training set to a function in the hypotheses space, that is

$$f_n = A(S_n) \in \mathcal{H},$$

and is typically assumed to be *symmetric*, that is, invariant to permutations in the training set. An online learning algorithm is defined recursively as $f_0 = 0$ and

$$f_{n+1} = A(f_n, z_n).$$

A weaker notion of an online algorithm is $f_0 = 0$ and $f_{n+1} = A(f_n, S_{n+1})$. The former definition gives a memory-less algorithm, while the latter keeps memory of the past. Clearly, the algorithm obtained from either of these two procedures will not in general be symmetric.

Example 2 (ERM)**.** *The prototype example of batch learning algorithm is empirical risk minimization, defined by the variational problem*

$$\min_{f \in \mathcal{H}} I_n(f),$$

where $I_n(f) = \mathbb{E}_n V(f, z)$, $\mathbb{E}_n$ being the empirical average on the sample, and $\mathcal{H}$ is typically assumed to be a proper, closed subspace of $\mathbb{R}^p$, for example a ball or the convex hull of some given finite set of vectors.

Example 3 (SGD)**.** *The prototype example of online learning algorithm is stochastic gradient descent, defined by the recursion*

$$f_{n+1} = \Pi_K(f_n - \gamma_n \nabla V(f_n, z_n)), \tag{1}$$

where z_n is fixed, $\nabla V(f_n, z_n)$ is the gradient of the loss with respect to f at f_n, and γ_n is a suitable decreasing sequence. Here K is assumed to be a closed subset of $\mathcal{H}$ and $\Pi_K : \mathcal{H} \to K$ the corresponding projection. Note that if K is convex then Π_K is a contraction, i.e. $\|\Pi_K\| \leq 1$ and moreover if $K = \mathcal{H}$ then $\Pi_K = I$.

1.3 Generalization and Consistency

In this section we discuss several ways of formalizing the concept of generalization of a learning algorithm. We say that an algorithm is weakly consistent if we have convergence of the risks in probability, that is for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(I(f_n) - I(f_K) > \epsilon) = 0, \tag{2}$$

and that it is strongly consistent if convergence holds almost surely, that is

$$\mathbb{P}\left(\lim_{n\to\infty} I(f_n) - I(f_K) = 0\right) = 1.$$

A different notion of consistency, typically considered in statistics, is given by convergence in expectation

$$\lim_{n\to\infty} \mathbb{E}[I(f_n) - I(f_K)] = 0.$$

Note that, in the above equations, probability and expectations are with respect to the sample S_n. We add three remarks.

Remark 1. *A more general requirement than those described above is obtained by replacing $I(f_K)$ by $\inf_{f\in\mathcal{H}} I(f)$. Note that in this latter case no extra assumptions are needed.*

Remark 2. *Yet a more general requirement would be obtained by replacing $I(f_K)$ by $\inf_{f\in\mathcal{F}} I(f)$, $\mathcal{F}$ being the largest space such that $I(f)$ is defined. An algorithm having such a consistency property is called universal.*

Remark 3. *We note that, following Alon at al., the convergence* (2) *corresponds to the definition of learnability of the class $\mathcal{H}$.*

1.3.1 Other Measures of Generalization.

Note that alternatively one could measure the error with respect to the norm in $\mathcal{H}$, that is $\|f_n - f_K\|$, for example

$$\lim_{n\to\infty} \mathbb{P}(\|f_n - f_K\| > \epsilon) = 0. \tag{3}$$

A different requirement is to have convergence in the form

$$\lim_{n\to\infty} \mathbb{P}(|I_n(f_n) - I(f_n)| > \epsilon) = 0. \tag{4}$$

Note that for both the above error measures one can consider different notions of convergence (almost surely, in expectation) as well convergence rates, hence finite sample bounds.

For certain algorithms, most notably ERM, under mild assumptions on the loss functions, the convergence (4) implies weak consistency[1]. For general algorithms there is no straightforward connection between (4) and consistency (2).

Convergence (3) is typically stronger than (2), in particular this can be seen if the loss satisfies the Lipschitz condition

$$|V(f,z) - V(f',z)| \leq L \|f - f'\|, \qquad L > 0, \tag{5}$$

for all $f, f' \in \mathcal{H}$ and $z \in Z$, but also for other loss function which do not satisfy (5) such as the square loss.

1.4 Stability and Generalization

Different notions of stability are sufficient to imply consistency results as well as finite sample bounds.

A strong form of stability is uniform stability

$$\sup_{z \in Z} \sup_{z_1,\ldots,z_n} \sup_{z' \in Z} |V(f_n, z) - V(f_{n,z'}, z)| \leq \beta_n$$

where $f_{n,z'}$ is the function returned by an algorithm if we replace the i-th point in S_n by z' and β_n is a decreasing function of n.

Bousquet and Eliseef prove that the above condition, for algorithms which are *symmetric*, gives exponential tail inequalities on $I(f_n) - I_n(f_n)$ meaning that we have $\delta(\epsilon, n) = e^{-C\epsilon^2 n}$ for some constant C. Wibisono et al. show that ERM with a strongly convex loss function is always uniformly stable. Weaker requirements can be defined by replacing one or more supremums with expectation or statements in probability. Exponential inequalities will in general be replaced by weaker concentration, Poggio et al. and Rakhlin et al. give a thorough discussion and list of references.

2 Stability and SGD

Here we focus on online learning and in particular on SGD and discuss the role played by the following definition of stability, that we call CV$_{on}$ stability

Definition 2.1. *We say that an online algorithm is CV_{on} stable with rate β_n if for $n > N$ we have*

$$-\beta_n \leq \mathbb{E}_{z_n}[V(f_n, z_n) - V(f_{n+1}, z_n)|S_n] < 0, \tag{6}$$

where $S_n = z_0, \ldots, z_{n-1}$ and $\beta_n \geq 0$ goes to zero with n.

[1] In fact for ERM

$$\mathbb{P}(I(f_n) - I(f_K) > \epsilon) \leq \mathbb{P}(I(f_n) - I_n(f_n) + I_n(f_K) - I_n(f_n) + I_n(f_*) - I(f_*) > \epsilon) \leq$$
$$\mathbb{P}(I(f_n) - I_n(f_n) > \epsilon) + \mathbb{P}(I_n(f_*) - I_n(f_n) > \epsilon) + \mathbb{P}(I_n(f_*) - I(f_*) > \epsilon)$$

The first term goes to zero because of (4), the second term has probability zero since f_n minimizes I_n, the third term goes to zero if $V(f_*, z)$ is a well behaved random variable (for example if the loss is bounded but also under weaker moment/tails conditions).

The definition above is of course equivalent to

$$0 \leq \mathbb{E}_{z_n}[V(f_n, z_n) - V(f_{n+1}, z_n)|S_n] \leq \ \beta_n. \tag{7}$$

In particular, we assume $\mathcal{H}$ to be a p-dimensional Hilbert Space and $V(\cdot, z)$ to be convex and twice differentiable in the first argument for all values of z. We discuss the stability property of (1) when K is a closed, convex subset[2] of $\mathcal{H}$.

2.1 Setting and Preliminary Facts

We recall the following standard result, see Lelong and references therein for a proof.

Theorem 1. *Consider* (8) *and assume that,*

- *There exists $f_K \in K$, such that $\nabla I(f_K) = 0$, and for all $f \in \mathcal{H}$, $\langle f - f_K, \nabla I(f) \rangle > 0$.*
- $\sum_n \gamma_n = \infty$, $\sum_n \gamma_n^2 = 0$.
- *There exists $D > 0$, such that for all $f \in \mathcal{H}$,*

$$\mathbb{E}_{z_n}[\|\nabla V(f_n, z)\|^2 |S_n] \leq D(1 + \|f_n - f_K\|^2). \tag{9}$$

Then,

$$\mathbb{P}(\lim_{n\to\infty} \|f_n - f_K\| = 0) = 1.$$

The following result will be also useful.

Lemma 1. *Under the same assumptions of Theorem 1, if f_K belongs to the interior of K, then there exists $N > 0$ such that for $n > N$, $f_n \in K$ so that the projection in* (8) *is not necessary.*

2.1.1 Stability of SGD

Throughout this section we assume that

$$\langle f, H(V(f, z))f \rangle \geq 0 \qquad \|H(V(f, z))\| \leq M < \infty, \tag{10}$$

for any $f \in \mathcal{H}$ and $z \in Z$.

Theorem 2. *Under the same assumption of Theorem 1, there exists N such that for $n > N$, SGD satisfies CV_{on} with $\beta_n = C\gamma_n$, where C is a universal constant.*

[2] Note that a particular case is $K = \mathcal{H}$, that is

$$f_{n+1} = f_n - \gamma_n \nabla V(f_n, z_n). \tag{8}$$

Proof. Note that from Taylor's formula,

$$[V(f_{n+1},z_n)-V(f_n,z_n)] = \langle f_{n+1}-f_n, \nabla V(f_n,z_n)\rangle + 1/2\,\langle f_{n+1}-f_n, H(V(f,z_n))(f_{n+1}-f_n)\rangle\,, \quad (11)$$

with $f=\alpha f_n+(1-\alpha)f_{n+1}$ for $0\leq\alpha\leq 1$. We can use the definition of SGD and Lemma 1 to show there exists N such that for $n>N$, $f_{n+1}-f_n=\gamma_n V(f_n,z_n)$. Hence changing signs in (11) and taking the expectation w.r.t. z_n conditioned over $S_n=z_0,\ldots,z_{n-1}$, we get

$$\mathbb{E}_{z_n}[V(f_n,z_n)-V(f_{n+1},z_n)|S_n] =$$

$$\gamma_n\,\mathbb{E}_{z_n}[\|\nabla V(f_n,z_n)\|^2\,|S_n] + 1/2\gamma_n^2\,\mathbb{E}_{z_n}[\langle\nabla V(f_n,z_n), H(V(f,z_n))\nabla_f V(f_n,z_n)\rangle\,|S_n]. \quad (12)$$

The above quantity is clearly non negative, in particular the last term is non negative because of (10). Using (9) and (10) we get

$$\mathbb{E}_{z_n}[V(f_n,z_n)-V(f_{n+1},z_n)|S_n] = (\gamma_n+1/2\gamma_n^2 M)D(1+\mathbb{E}_{z_n}[\|f_n-f_K\|\,|S_n])\leq C\gamma_n,$$

if n is large enough. □

A partial converse result is given by the following theorem.

Theorem 3. *Assume that,*

- *There exists $f_K\in K$, such that $\nabla I(f_K)=0$, and for all $f\in\mathcal{H}$, $\langle f-f_K,\nabla I(f)\rangle>0$.*
- $\sum_n\gamma_n=\infty$, $\sum_n\gamma_n^2=0$.
- *There exists $C,N>0$, such that for all $n>N$, (7) holds with $\beta_n\leq C\gamma_n$.*

Then,

$$\mathbb{P}\left(\lim_{n\to\infty}\|f_n-f_K\|=0\right)=1. \quad (13)$$

Proof. Note that from (11) we also have

$$\mathbb{E}_{z_n}[V(f_{n+1},z_n)-V(f_n,z_n)|S_n] =$$

$$-\gamma_n\,\mathbb{E}_{z_n}[\|\nabla V(f_n,z_n)\|^2\,|S_n] + 1/2\gamma_n^2\,\mathbb{E}_{z_n}[\langle\nabla V(f_n,z_n), H(V(f,z_n))\nabla_f V(f_n,z_n)\rangle\,|S_n].$$

so that using the stability assumption and (10) we obtain,

$$-\beta_n\leq(1/2\gamma_n^2-\gamma_n)\,\mathbb{E}_{z_n}[\|\nabla V(f_n,z_n)\|^2\,|S_n]$$

that is,

$$\mathbb{E}_{z_n}[\|\nabla V(f_n,z_n)\|^2\,|S_n]\leq\frac{\beta_n}{(\gamma_n-M/2\gamma_n^2)}\leq\frac{C\gamma_n}{(\gamma_n-M/2\gamma_n^2)}.$$

From Lemma 1 for n large enough we obtain

$$\begin{aligned}\left\|f_{n+1}-f_K\right\|^2 &\leq \left\|f_n-\gamma_n \nabla V(f_n,z_n)-f_K\right\|^2\\ &\leq \left\|f_n-f_K\right\|^2+\gamma_n^2\left\|\nabla V(f_n,z_n)\right\|^2-2\gamma_n\left\langle f_n-f_K,\nabla V(f_n,z_n)\right\rangle\end{aligned}$$

so that taking the expectation w.r.t. z_n conditioned to S_n and using the assumptions, we write

$$\begin{aligned}\mathbb{E}_{z_n}[\left\|f_{n+1}-f_K\right\|^2|S_n] &\leq \left\|f_n-f_K\right\|^2+\gamma_n^2\,\mathbb{E}_{z_n}[\left\|\nabla V(f_n,z_n)\right\|^2|S_n]-2\gamma_n\left\langle f_n-f_K,\mathbb{E}_{z_n}[\nabla V(f_n,z_n)|S_n]\right\rangle\\ &\leq \left\|f_n-f_K\right\|^2+\gamma_n^2\frac{D\gamma_n}{(\gamma_n-M/2\gamma_n^2)}-2\gamma_n\left\langle f_n-f_K,\nabla I(f_n)\right\rangle,\end{aligned}$$

since $\mathbb{E}_{z_n}[\nabla V(f_n,z_n)|S_n]=\nabla I(f_n)$. The series $\sum_n \gamma_n^2\frac{D\gamma_n}{(\gamma_n-M/2\gamma_n^2)}$ converges and the last inner product is positive by assumption, so that the Robbins-Siegmund's theorem implies (13) and the theorem is proved. □

A Remarks: assumptions

- The assumptions will be satisfied if the loss is convex (and twice differentiable) and $\mathcal{H}$ is compact. In fact, a convex function is always locally Lipschitz so that if we restrict $\mathcal{H}$ to be a compact set, V satisfies (5) for

  $$L=\sup_{f\in\mathcal{H},z\in Z}\left\|\nabla V(f,z))\right\|<\infty.$$

  Similarly since V is twice differentiable and convex, we have that the Hessian $H(V(f,z))$ of V at any $f\in\mathcal{H}$ and $z\in Z$ is identified with a bounded, positive semi-definite matrix, that is

  $$\langle f,H(V(f,z))f\rangle\geq 0 \qquad \left\|H(V(f,z))\right\|\leq 1<\infty,$$

  for any $f\in\mathcal{H}$ and $z\in Z$, where for the sake of simplicity we took the bound on the Hessian to be 1.

- The gradient in the SGD update rule can be replaced by a stochastic subgradient with little changes in the theorems.

B Learning Rates, Finite Sample Bounds and Complexity

B.1 Connections Between Different Notions of Convergence.

It is known that both convergence in expectation and strong convergence imply weak convergence. On the other hand if we have weak consistency and

$$\sum_{n=1}^{\infty}\mathbb{P}(I(f_n)-I(f_K)>\epsilon)<\infty$$

for all $\epsilon>0$, then weak consistency implies strong consistency by the Borel-Cantelli lemma.

B.2 Rates and Finite Sample Bounds.

A stronger result is weak convergence with a *rate*, that is

$$\mathbb{P}(I(f_n) - I(f_K) > \epsilon) \geq \delta(n, \epsilon),$$

where $\delta(n, \epsilon)$ decreases in n for all $\epsilon > 0$. We make two observations. First, one can see that the Borel-Cantelli lemma imposes a rate on the decay of $\delta(n, \epsilon)$. Second, typically $\delta = \delta(n, \epsilon)$ is invertible in ϵ so that we can write the above result as a finite sample bound

$$\mathbb{P}(I(f_n) - I(f_K) \leq \epsilon(n, \delta)) \geq 1 - \delta.$$

B.3 Complexity and Generalization

We say that a class of real valued functions $\mathcal{F}$ on Z is uniform Glivenko-Cantelli if the following limit exists

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{F\in\mathcal{F}} |\,\mathbb{E}_n(F) - \mathbb{E}(F)| > \epsilon\right) = 0.$$

for all $\epsilon > 0$. If we consider the class of functions induced by V and $\mathcal{H}$, that is $F(\cdot) = V(f, \cdot)$, $f \in \mathcal{H}$, the above properties can be written as

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{f\in\mathcal{H}} |I_n(f) - I(f)| > \epsilon\right) = 0. \tag{14}$$

Clearly the above property implies (4), hence consistency of ERM if $f_{\mathcal{H}}$ exists and under mild assumption on the loss – see previous footnote.

It is well known that UGC classes can be completely characterized by suitable capacity/complexity measures of $\mathcal{H}$. In particular a class of binary valued functions is UGC if and only if the VC-dimension is finite. Similarly a class of bounded functions is UGC if and only if the fat- shattering dimension is finite. See Alon et al. and reference therein.

Finite complexity of $\mathcal{H}$ is hence a sufficient condition for the consistency of ERM.

B.4 Necessary Conditions

One natural question is weather the above conditions are also necessary for consistency of ERM in the sense of (2), or in other words if consistency of ERM on $\mathcal{H}$ implies that $\mathcal{H}$ is UGC class.

An argument in this direction is given by Vapnik which call the result the key theorem in learning (together with the converse direction). Vapnik argues that (2) must be replaced by a much stronger notion of convergence essentially holding if we replace $\mathcal{H}$ with $\mathcal{H}_\gamma = \{f \in \mathcal{H} \mid I(f) \geq \gamma\}$, for all γ.

Another result in this direction is given *without proof* in Alon et al.

B.5 Robbins Siegmund's Lemma

We use the stochastic approximation framework described by Duflo ([1], pp 6-15).

We assume a sequence of *data* z_i defined by a probability space Ω, A, P and a filtration $\mathrm{F} = (\mathcal{F})_{n \in \mathbb{N}}$ where $\mathcal{F}_n$ is a σ-field and $\mathcal{F}_n \in \mathcal{F}_{n+1}$. In addition a sequence $\mathcal{X}_n$ of measurable functions from Ω, A to another measurable space is defined to be *adapted* to F if for all n, $\mathcal{X}_n$ is $\mathcal{F}_n$-measurable.

Definition *Suppose that $X = (X_n)$ is a sequence of random variables adapted to the filtration* F. *X is a supermartingale if it is integrable (see [1]) and if*

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n$$

The following is a key theorem ([1]).

Theorem B.1. *(Robbins-Siegmund) Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let (V_n), (β_n), (χ_n), (η_n) be finite non-negative $\mathcal{F}_n$-mesurable random variables, where $\mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n \subset \cdots$ is a sequence of sub-σ-algebras of $\mathcal{F}$. Suppose that (V_n), (β_n), (χ_n), (η_n) are four positive sequences adapted to* F *and that*

$$\mathbb{E}[V_{n+1}|\mathcal{F}_n] \leq V_n(1 + \beta_n) + \chi_n - \eta_n.$$

Then if $\sum \beta_n < \infty$ and $\sum \chi_n < \infty$, almost surely (V_n) converges to a finite random variable and the series $\sum \eta_n$ converges.

We provide a short proof of a special case of the theorem.

Theorem B.2. *Suppose that (V_n) and (η_n) are positive sequences adapted to $\mathbb{F}$ and that*

$$\mathbb{E}[V_{n+1} \mid \mathcal{F}_n] \leq V_n - \eta_n.$$

Then almost surely (V_n) converges to a finite random variable and the series $\sum \eta_n$ converges.

Proof

Let $Y_n = V_n + \sum_{k=1}^{n-1} \eta_k$. Then we have

$$\mathbb{E}[Y_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[V_{n+1} \mid \mathcal{F}_n] + \sum_{k=1}^{n} \mathbb{E}[\eta_k \mid \mathcal{F}_n] \leq V_n - \eta_n + \sum_{k=1}^{n} \eta_k = Y_n.$$

So (Y_n) is a supermartingale, and because (V_n) and (η_n) are positive sequences, (Y_n) is also bounded from below by 0, which implies it converges almost surely. It follows that both (V_n) and $\sum \eta_n$ converge.

References

[1] M. Duflo. *Random Iterative Models*. Pringer, New York, 1991.

[2] S. Mukherjee T. Poggio, R. Rifkin and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428:419–422, 2004.