Randomisation of Positive Linear Learning Algorithms in Banach Function Lattices

Richard F. Bonner

Centre for Mathematical Modelling
Department of Mathematics and Physics
Mälardalen University, Västeras, Sweden.
Richard.Bonner@mdh.se

Abstract

When knowledge is represented in a Banach lattice, relationships between knowledge correspond to Banach lattice homomorphisms $T : E \to F$. In the context of learning, one may, for example, think of the teacher's knowledge as a point x in E, and let Tx be the target knowledge of the student. A learning algorithm is then a uniformly bounded sequence ω of (linear) positive operators $T_n : E \to F$ of finite rank. A subset M of E is T-learnable by ω if the sequence $T_n x$ converges to Tx for every x of M. A classical theory of approximation of Banach lattice homomorphisms, known as the Korovkin theory, is invoked for finite learnability conditions. For Banach *function* lattices in the sense of Luxemburg and Zaanen, a randomised Korovkin-type theorem is proposed. Interpreted for learning, the result asserts the existence of a finite set $M \subset E$ of concepts which are "most difficult to learn" for any learning algorithm: if a random algorithm is likely to learn these concepts in a weak sense, then it almost surely learns all concepts in E.

1 Introduction

Lattice theory has many lives [16]. In mathematical analysis, for example, lattice structure occurs naturally in combination with that of a Banach space, resulting in the classical concept of Banach lattice; see the standard reference [17] or the Appendix in [12] for a primer. The Euclidean spaces R^n, $n \geq 1$, the uniform function algebra $C(X)$, and the familiar spaces $L^p(X, \mu)$, $1 \leq p \leq \infty$, all are Banach (function) lattices, with the lattice structure inherited from the reals. Roughly, a Banach lattice is a Banach space with a projection onto a closed conic subset of its "positive elements".

It makes sense to consider models of learning of the COLT type [20] in Banach lattices, quite apart from all the deeper reasons for representing knowledge by (non-linear) lattice structure [18]. The step from binary to real-valued concepts is indeed a step from Boolean structures into Banach lattices. The lattice $L^p(X, \mu)$, for example, is the smallest linear lattice containing a given σ-algebra and complete in the $\|\cdot\|_p$ norm.

Recall the general idea of computational learning models. Think of some knowledge at hand as a point x in a Banach space E, and think of the learning target - the approximate representation of the knowledge - as a point Tx in another Banach space F, so that $T : E \to F$ is a continuous map. In particular, if the learning target is the point x itself, the operator T is the identity on E. A learning algorithm is then naturally a sequence of "computable" operators $T_n : E \to F$,

and the algorithm "learns" knowledge Tx if the sequence $T_n x$ converges to Tx in a suitable sense. In the present note all operators in the learning algorithms are assumed *linear* and of finite rank. Learning in the strong operator topology is then in general *not* possible in Banach spaces [5]: there are bounded linear maps $T : E \to F$ between Banach spaces for which Tx cannot be learned for some x in E. However, learning *is* possible in a large class of Banach lattices [7][8], including the Banach function lattices such as $C(X)$ and $L^1(X, \mu)$. The idea of learning Banach lattice *homomorphisms* is further explained in the next Section 3.
Interestingly, the problem of learning Banach lattice homomorphisms by *positive* algorithms has a simple solution - the Korovkin theory [1][4]. Korovkin's striking result [13] was originally formulated for the identity operator in the lattice $C[0, 1]$ as follows: a sequence of positive linear operators on $C[0, 1]$ converges uniformly to the identity operator at *every* function in $C[0, 1]$ if it converges uniformly at each of the *three* functions $t \mapsto 1$, $t \mapsto t$, and $t \mapsto t^2$; see also [14]. In intuitive language, it is enough in this case to check that an algorithm learns three chosen concepts in order to conclude that it learns all other concepts as well. The theorem extends to the general uniform lattice $C(X)$ on a compact set as an addendum to the Stone-Weierstrass theorem, and there are abstract generalizations. The the main points of Korovkin theory are recalled in Section 4 following Schaefer [17]. Finer notions of Korovkin closure naturally lead to various "theories of convexity" of which linear convexity is a special case; these ideas are fundamental in analysis [9] but are not pursued here.
Section 5 then takes a brief look at the Korovkin theory in a randomised setting. The point here is to weaken the sufficient convergence requirement on the finite "testing set" so that ordinary stochastic analysis may be used to verify them. In particular, this can be easily done if the algorithm is monotone on the testing set with respect to the lattice order, since, by the Dini theorem for lattices, strong and weak convergence are the same for monotone sequences.

2 Banach lattices

Recall very briefly the main notions; see otherwise the standard reference [17].
Consider a real Banach space $(E, \|\cdot\|)$ with a closed conic subset $E_+ \subset E$ of *positive elements*, and a projection mapping ("taking the *positive part* of") $x \mapsto x_+$ of E onto E_+; one then writes $x \le y$ if $y - x \in E_+$, and puts $x_- := (-x)_+$ and $|x| := x_+ - x_-$, defining the *negative part* and the *absolute value* of x, respectively. The Banach space E is called a *Banach lattice* if its norm is monotone in the sense that $|x| \le |y|$ implies $\|x\| \le \|y\|$ for all x, y in E. Let, for example, X be a compact set and let $\mathcal{A}$ denote a σ-algebra of its subsets. The classical Banach space $C(X)$ of all real-valued continuous functions on X with the supremum norm $\|\cdot\|_\infty$ is naturally a Banach lattice, and so are, for $1 \le p \le \infty$, the Banach spaces $L^p(X, \mu)$ of all (equivalence classes of) $\mathcal{A}$-measurable real-valued functions with (finite) norm $\|\cdot\|_p$ defined by the p-th root of the integral $\int |f|^p \, d\mu$.
The notion of a Banach sublattice is defined in the obvious way, and a set $S \subset E$ is said to *generate* E if there is no proper Banach sublattice between S and E. Naturally, a Banach lattice generated by a finite set is said to be *finitely generated*. For example, by the Stone-Weierstrass theorem, the uniform lattice $C(X)$ on a compact set X in R^n is (finitely) generated by the constant function 1_X and the coordinate functions $x_1, ..., x_n$, and by the dense inclusion $C(X) \subset L^p(X, \mu)$, so are the lattices $L^p(X, \mu)$, $1 \le p < \infty$.
The Banach space dual E' of E is then also a Banach lattice with E'_+ consisting of functionals positive on E_+. Recall that the positive part U°_+ of the dual unit ball is compact in the weak* topology (ie the $\sigma(E', E)$-topology) and E_+ can be viewed as a set of positive continuous real-valued functions on U°_+. One now shows that there exists a unique closed subset S of U°_+ such

that every function in E_+ attains its maximum on S. This set is called the *Šilov boundary* of E_+ and is here denoted by $\check{S}(E_+)$. For example, he dual positive cone of the uniform function lattice $E = C(X)$ consists of the positive Borel measures on X, of which the point masses $\{\delta_x : x \in X\}$ constitute the Šilov boundary.
Further, the Banach space $\mathcal{L} = \mathcal{L}(E, F)$ of bounded linear maps between Banach lattices is naturally ordered by the positive cone $\mathcal{L}_+$ of *positive maps* mapping E_+ into F_+. A map $T \in \mathcal{L}(E, F)$ is a *lattice homomorphism* if it commutes with the absolute value: $T\,|x| = |Tx|$ for all $x \in E$; in particular, a lattice homomorphism is a positive map. For example, the set of the normalised real-valued homomorphisms on $C(X)$ consists of the evaluation functionals δ_x, $x \in X$, and thus coincides with the Šilov boundary of $C(X)_+$. Less trivially, all lattice homomorphisms $T : C(X) \to C(Y)$ preserving the constant functions are of the form $Tf = f \circ k$ for some continuous map $k : Y \to X$; see [17], Ch.III, Th.9.1, for example.

3 Models of learning

In application to models of learning, an element x of a Banach lattice such as $C(X)$ or $L^p(X, \mu)$ may be viewed as an encoding of knowledge; it may, for example, represent a behaviour, a concept, or a decision rule. In the lattices $L^p(X, \mu)$, it is instructive to recall, the lattice operations applied to the indicator functions of measurable sets, represent the Boolean operations on the sets. Furthermore, the lattice homomorphisms between such lattices extend the Boolean morphisms of their algebras of measurable sets, and, by results of the type just quoted, essentially arise from a transformation between their "instance spaces". Intuitively speaking, lattice homomorphisms preserve the structure of knowledge when merging instances into concepts.
A learning algorithm may then be represented by a sequence of positive operators of finite rank, as follows. Let E and F be real Banach lattices, and let $T : E \to F$ be a Banach lattice homomorphism. The goal is to approximate T by sequences T_n of operators of the form $R_n \circ S_n : E \to R^n \to F$ where S_n is thought to "sample" an element $x \in E$ which is then "approximately reconstructed" by R_n. If both R_n and S_n are positive (and therefore continuous!), the learning operator T_n has the form $x \mapsto \sum_{1 \leq k \leq n} \psi_{nk}(x) \cdot y_{nk}$ with $\psi_{nk} \in E'_+$ and $y_{nk} \in F_+$, $1 \leq k \leq n$. If, for example, T is the identity operator on $C(X)$ and the functionals ψ_{nk} are taken as evaluations, one has $T_n\, x(t) = \sum_{1 \leq k \leq n} x(t_{nk}) \cdot y_{nk}(t)$ for $t, t_{nk} \in X$, $1 \leq k \leq n$. Recall the classical Bernstein operators B_n in $C([0, 1])$, in which case $t_{nk} = \frac{k}{n}$ and $y_{nk}(t) = \binom{n}{k} t^{n-k}(1-t)^k$, $0 \leq k \leq n$.
It is thus natural to think of a (linear) theory of learning in Banach lattices as a theory of approximation of lattice homomorphisms by (linear) positive maps of finite rank. This is, in principle, classical Banach space theory, see [17], Ch.IV, for example. Empirical problems and constructivity considerations, however, lead to less classical questions, essentially concerning the rate of learning. For example, convergence rates in the operator norm are currently studied in terms of Kolmogorov (metric) entropy [2]. Other questions, such as those about Valiant learnability [19], combine weak topologies and stochasticity, and do not seem as yet to have found their way into approximation theory in Banach spaces. In neither of the cases, however, the role of lattice structure - usually present in applications - and the positivity of approximating operators, seem to have been fully explored.
Questions of convergence rates of learning are however not of concern in the present note. It will only be illustrated how lattice structure allows in certain cases to ascertain strong convergence of learning everywhere (or almost everywhere) by testing for weak convergence (or weak in measure) on a finite number of concepts.

4 Korovkin theory and Banach function lattices

For Banach lattices E and F denote by $\Lambda = \Omega(E, F)$ the set of all lattice homomorphisms $T : E \to F$, and let $\Omega = \Omega(E, F)$ denote the set of all equicontinuous sequences $(T_n)_{n\geq 1}$ of positive linear maps $T_n : E \to F$. A subset $M \subset E$ is called a *Korovkin family* for E if for all Banach lattices F, all T in $\Omega(E, F)$, and all $(T_n)_{n\geq 1}$ in $\Omega(E, F)$, the $\lim_n T_n x = Tx$ for $x \in M$ implies $\lim_n T_n x = Tx$ for $x \in E$. (Roughly: strong convergence on M implies strong convergence everywhere). For example, by the quoted result of Korovkin, the uniform lattice $C([0, 1])$ has a Korovkin family consisting of the three monomials $t \mapsto t^k$, $k = 0, 1, 2$. This result has an elementary proof, see the Section 7, and was applied by Korovkin to the Bernstein operators yielding a simple proof of the Weierstrass theorem for $C([0, 1])$.
An simple answer to the question: which Banach lattices have a finite Korovkin family? was given by Wolff [21] in the 70's: the finitely generated lattices. Such lattices can be identified as the so called Banach function lattices, studied in the 60's by Luxemburg and Zaanen, see [17]. Let $X = X_n$ be a compact set in R^n and let $\mathcal{M}$ be a vaguely compact set of positive Radon measures on X whose supports have a union dense in X. For any finite Borel function f on X define the seminorm $p_{\mathcal{M}}(f) = \sup_{\mu\in\mathcal{M}} \int |f|\, d\mu$, and let $B(X, \mathcal{M})$ be the vector lattice of all such functions f for which $p_{\mathcal{M}}(f)$ is finite. The lattice $B(X, \mathcal{M})$ is complete under the seminorm $p_{\mathcal{M}}$, and, hence, so is the closure $C(X)^-$ of the set $C(X)$ of all continuous functions in $B(X, \mathcal{M})$.

Definition 1 *The Banach lattice $C(X)^-/p_{\mathcal{M}}^{-1}(0)$ is called a* Banach function lattice, *and is denoted by $E = E(X, \mathcal{M}) = E(X_n, \mathcal{M})$.*

Notice, for example, that $C(X)$ and $L^1(X, d\mu)$ are both Banach function lattices corresponding to $\mathcal{M} = \{\delta_x : x \in X\}$ and $\mathcal{M} = \{\mu\}$, respectively.
By Wolff 's theorem, a Banach lattice has a finite Korovkin family if it is finitely generated. Hence, by the Stone-Weierstrass theorem, all function lattice spaces $E(X_n, \mathcal{M})$ have finite Korovkin families. The converse holds for Banach lattices with quasi-interior elements. Recall that a positive element $x \geq 0$ in a Banach lattice E is called a *quasi-interior element* of E if it distinguishes positive non-zero functionals from zero: $\psi(x) > 0$ for all $0 \neq \psi \in E'_+$. In the uniform lattice $C([0, 1])$, for example, the quasi-interior elements are the positive functions x bounded away from zero, $\inf_{0\leq t\leq 1} x(t) > 0$.

Theorem 1 (Schaefer [17]-Wolff [21]) *A Banach function lattice $E(X_n, \mathcal{M})$ has a system of n generators, and, consequently, a Korovkin family with $2n + 1$ elements. Conversely, any Banach lattice with quasi-interior positive elements which has a finite system of generators, or, equivalently, a finite Korovkin family, is isomorphic to a Banach function lattice $E(X, \mathcal{M})$.*

Note, as a corollary, that also each of the Banach lattices $L^p(X_n, \mu)$, $1 < p < \infty$, is isomorphic to a Banach function lattice, since, by the Stone-Weierstrass theorem, each is finitely generated, and each possesses the quasi-interior element 1_{X_n}. Banach function lattices may thus be presented with additional structure.

5 Randomising Korovkin

There are several directions in which one may pursue the Korovkin phenomenon. One such direction is to use a priori information, such as the monotonicity of a learning algorithm, to weaken the topology of convergence on a Korovkin family. An extention to a stochastic setting is then straightforward.

Theorem 2 *Let E be a Banach function lattice with a Korovkin family M, let F be Banach lattice, and let $T : E \to F$ be a lattice homomorphism. Denote by $\Omega = \Omega(E, F)$ the set of all equicontinuous sequences $\omega = (T_n)_{n\geq 1}$ of positive linear maps $E \to F$, writing $T_n = T_n(\omega)$ to indicate the n:th operator in the sequence ω. Let $\mathcal{A}$ be a σ-algebra on Ω with respect to which all the evaluations $\omega \mapsto \langle T_n(\omega)x, \psi\rangle$ for $x \in E$, $\psi \in \check{S}(F_+)$, $n \geq 1$, are measurable. Finally, let μ be a probability measure on Ω. Then, if for μ-almost all ω, and all $x \in M$ and $\psi \in \check{S}(F_+)$, the numerical sequences $\langle T_n(\omega)x, \psi\rangle$ are increasing and converge to $\langle Tx, \psi\rangle$ in the measure μ,*

$$\lim_{n\to\infty} \mu \{\omega : |\langle T_n(\omega)x, \psi\rangle - \langle Tx, \psi\rangle| > \varepsilon\} = 0 \text{ for all } \varepsilon > 0,$$

then, for μ-almost all ω, the operators $T_n = T_n(\omega)$ converge to T almost uniformly on E,

$$\lim_{n\to\infty} \sup_{x\in K} \|T_n x - Tx\|_F = 0 \text{ for all compact } K \subset E,$$

and $Tx = \sup_n T_n x$ for $x \in E$.

Proof. Observe first that for increasing sequences of real random variables the notions of convergence in measure and convergence almost everywhere coincide, hence the sequences $\langle T_n(\omega)x, \psi\rangle$ converge to $\langle Tx, \psi\rangle$ for $x \in M$, $\psi \in \check{S}(F_+)$, and μ-almost all ω. The proof now parallels that of Dini's theorem for Banach lattices, cf [17], Ch. II, Th. 5.9. For $x \in M$ and ω outside a zero set, the functions $\psi \mapsto \langle T_n(\omega)x, \psi\rangle$ are continuous on the $\sigma(E', E)$-compact space $\check{S}(F_+)$ and converge pointwise monotonously to the continuous function $\psi \mapsto \langle Tx, \psi\rangle$. By Dini's classical theorem, the convergence is uniform on $\check{S}(F_+)$. By the defining property of the Šilov boundary of F_+, this gives $\lim_{n\to\infty} \|T_n x - Tx\|_F = 0$ for $x \in M$, and therefore for all $x \in E$, considering M is a Korovkin family. The convergence is uniform on compact sets in E by the assumed equicontinuity of the sequence T_n. The last statement is due to the fact that, in any normed vector lattice, the cone of positive elements is a closed set. ■

Obviously, everything works equally well with the operators T_n *decreasing* on M instead.

It is not clear how to characterize the class of algorithms which are monotone on *some* Korovkin family, or how to look for such a Korovkin family for a given algorithm. In application, it may help to remember that if a family $f_1, ..., f_N$ separates points in X_n then it may be extended to a Korovkin family for $E(X_n, \mathcal{M})$ by adjoining the constant function 1_{X_n} and the squares $f_1^2, ..., f_N^2$.

It is instructive to exemplify Theorem 2 with the Bernstein operators in $C([0, 1])$, recall Section 3, randomised in the following way. Sample points t_j, $j \geq 1$, in $[0, 1]$ independently and according to a probability distribution μ, and, for each $0 \leq k \leq n$, replace the point $\frac{k}{n}$ in the Bernstein operator B_n by its closest neighbour in the sequence $t_1, ..., t_n$. It is easy to see that the random algorithm so defined satisfies the hypothesis of the theorem for the standard Korovkin family $\{t \mapsto 1,\ t \mapsto t,\ t \mapsto t^2\}$ provided μ gives positive mass to every proper subinterval of $[0, 1]$.

6 In conclusion

Reassert briefly some apparently open connections to other mathematical matter. The sampling operators in a learning algorithm are often the operators of restriction of function to sets, finite or not. The randomisation of such algorithms then naturally connects to Matheron's theory of random sets [15]. In general, however, one may have to stay within a theory of random operators [6], or in some cases of finite rank, random measures [11]. Rates of random convergence are obviously possible to define in the spirit of Kolmogorov metric entropy [2], also for non-linear operators, but I am not aware of current work in this direction. One would like, in particular, to understand Valiant learnability [19] in this setting. One may finally recall that some classical problems of analysis have been studied from the perspective of random learning [10], though the term may not have been used there. Also related is the work in algebraic complexity [3] where, however, connections with randomised techniques are only very recent.

7 Addendum

To demistify abstract Korovkin theory in lattices, recall briefly an explicit proof of Korovkin's original results [14].
First, in all generality, let $\mathcal{F} = \mathcal{F}(\mathcal{X})$ be the set of all real-valued functions on an abstract set X, let a be a point in X, and let $g_a \geq 0$ be a function vanishing at $x = a$. Denote by $1 = 1_X$ the function on X identically equal to one. Assume now that for every f in a certain subset $\mathcal{H}$ of $\mathcal{F}$, and for every $\varepsilon > 0$, there is a constant $C_\varepsilon(f)$ such that the modulus of the difference $f - f(a)$ is bounded by $\varepsilon + C_\varepsilon(f) \cdot g_a$. If μ is a positive linear functional on $\mathcal{F}$ then, obviously, the modulus of $\mu(f) - f(a)\mu(1)$ will be bounded by $\varepsilon\mu(1) + C_\varepsilon(f)\mu(g_a)$. Hence, if a sequence of positive linear functionals μ_k converges to the evaluation functional δ_a at the *two* functions 1 and g_a, it will also converge at *every* function in $\mathcal{H}$.
Specialisations are straightforward. If, for example, there is a metric d on X, one may put $g_a(x) = d(x, a)$, and let $\mathcal{H}$ be the set of all the bounded continuous real-valued functions on the metric space X. The constant $C_\varepsilon(f)$ may here be taken as the quotient between twice the sup norm of f and the radius of the largest ball around a in which $f - f(a)$ is bounded by ε. Further specialisation to the case $X \subset R_n$, this time with g_a as the square of the Euclidean distance to a, gives the original Korovkin theorems.

References

[1] F. Altomare and M. Campiti. *Korovkin-type approximation theory and its applications.* Studies in mathematics 17. Walter de Gruyter, 1994.

[2] C. Bernd and S. Irmtraud. *Entropy, compactness, and the approximation of operators.* Tracts in Mathematics 98. Cambridge University Press, 1990.

[3] P. Burgisser, M. Clausen, and M. A. Shokrollahi. *Algebraic complexity theory.* Grundlehren math. Viss. 315. Springer-Verlag, 1997.

[4] K. Donner. *Extension of positive operators and Korovkin theorems.* Lecture Notes in Mathematics 904. Springer-Verlag, 1982.

[5] P. Enflo. A counterexample to the approximation problem in Banach spaces. *Acta Mathematica*, 130:309–317, 1973.

[6] U. Grenander. *Abstract inference.* Wiley, 1981.

[7] A. Grothendieck. Produits tensoriels topologiques et espaces nucléaires. *Memoirs of the American Mathematical Society*, 16, 1955.

[8] A. Grothendieck. Une charactérisation vectorielle-métrique des espaces L^1. *Canadian Journal of Mathemaics*, 7:552–561, 1955.

[9] L. Hörmander. *Notions of convexity.* Progress in mathematics 127. Birkhäser, 1994.

[10] J. P. Kahane. *Some random series of functions, 2nd edition.* Cambridge University Press, 1985.

[11] O. Kallenberg. *Random measures.* Akademie-Verlag, 1986.

[12] J. L. Kelly and I. Namioka. *Linear Topological Spaces.* Springer-Verlag, 1963.

[13] P. P. Korovkin. On the convergence of linear positive operators in the space of continuous functions. *Dokl. Akad. Nauk SSSR (N.S.)*, 90:961–64, 1953.

[14] P. P. Korovkin. *Linear operators and approximation theory.* Hindustan Publishing Co., 1960 (russian original 1959).

[15] G. Matheron. *Random sets and integral geometry.* Wiley, 1975.

[16] G. C. Rota. The many lives of lattice theory. *Notices of the American Mathematical Society*, 44(11), 1997.

[17] H. H. Schaefer. *Banach Lattices and Positive Operators.* Springer-Verlag, 1974.

[18] D. S. Scott. Data types as lattices. *SIAM Jornal on Computing*, 5:522–587, 1976.

[19] L. G. Valiant. A theory of the learnable. *Comm. ACM*, 27(11):1134–1142, 1984.

[20] M. Vidyasagar. *A theory of learning and generalization.* Communications and control engineering. Springer-Verlag, 1997.

[21] M. Wolff. Über Korovkin-Sätze in lokalkonvexen Vektorverbänden. *Mathematische Annalen*, 204:49–56, 1973.