Non-archimedean analytic geometry: first steps

Vladimir Berkovich

When Dinesh Thakur asked me to write an introduction to this volume, I carelessly agreed. Later I started thinking that a short description of my journey to non-archimedean analytic geometry and of some of the circumstances accompanying it might be an entertaining complement to the notes of Matthew Baker and Brian Conrad. Since I had no other ideas, I've written what is presented below.

I start by briefly telling about myself. I was very lucky to be accepted to Moscow State University for undergraduate and, especially, for graduate studies in spite of the well-known Soviet policy of that time towards Jewish citizens. I finished studying in 1976, and got a Ph.D. the next year. (My supervisor was Professor Yuri Manin.) Getting an academic position would be too much luck, and the best thing I could hope for was the job of a computer programmer at a factory of agricultural machines and, later, at the institute of information in agriculture. As a result, I practically stopped doing mathematics, did not produce papers, and was considered by my colleagues as an outsider. It took me several years to become an expert in computers and nearby fields, and to learn to control my time. Gradually I started doing mathematics again, and my love for it blazed up with new force and became independent of surrounding circumstances. By the time my story begins, I was hungry for mathematics as never before.

Thus, my story begins one July evening of 1985 in a train in which I was returning to Moscow after having visited my numerous relatives in Gomel, Belarus. Instead of talking to people near me — my usual occupation during long train trips — I opened a book on classical functional analysis by Yuri Lyubich which my eldest brother Yakov had given me a couple of days before. The basic material of the book was familiar to me. Nevertheless, I was thrilled to read about it again and, suddenly, asked myself: what is the analog of all this over a non-archimedean field k? In particular, what is the spectrum of a bounded linear operator acting on a Banach space over k?

It did not take much time to find that, if one defines the spectrum in the same way as in the classical situation, it may be empty even if k is algebraically closed. Indeed, if K is a non-archimedean field larger than k, then the multiplication by any element of K which does not lie in k is an operator with empty spectrum. That such a larger field always exists is easily seen: one can take the completion of the field of rational functions in one variable over k with respect to the Gauss norm. I was very intrigued, and decided to understand what all this meant.

I knew that my fellow Manin student Misha Vishik had written a paper on non-archimedean spectral theory. The next day, I found the paper and started reading it. It turned out Vishik was studying bounded linear operators on a Banach space over k with the property that their resolvents are analytic on the complement of the spectrum defined in the usual way as a subset of k (the field was assumed to be algebraically closed). When I understood that, a very natural idea came to me. In the classical situation, the spectrum of an operator coincides with the complement of the analyticity set of its resolvent. Can one find out what the spectrum of a non-archimedean operator is by investigating a similar analyticity set of its resolvent? That such a resolvent takes values in the Banach algebra of bounded linear operators was not a problem. It was the notion of analyticity set that was not clear. But at least, one could try to investigate sets from a reasonable class at which the resolvent is analytic. For example, the resolvent is analytic at the complement of a closed (or open) disc with center at zero of a big enough radius.

At the beginning, I considered the so-called quasiconnected (and infraconnected) sets introduced by M. Krasner in 1940s, and I found a curious phenomenon whose slightly weakened form states the following. If the resolvent of a bounded linear operator is analytic at a standard set (i.e., the complement of a nonempty finite disjoint union of open discs in the projective line), then it is analytic at a strictly bigger standard set (i.e., all of the radii of the corresponding discs are strictly bigger or smaller). Of course, in the light of our present knowledge this phenomenon is completely clear since the standard sets being defined by non-strict inequalities are *closed* subsets of the *compact* projective line. But the analyticity set of the resolvent being the complement of the compact spectrum is an *open* set.

At that time I was not so smart to see the above. I considered the analyticity sets as strictly increasing families of finite disjoint unions of standard sets, and the spectra as strictly decreasing families of complementary sets of the same type. (A precise definition of complementary sets is given on p. 141 of my book.) The latter families can be viewed as filters of finite unions of standard sets. It turned out that one can easily describe the maximal elements in the family of filters, i.e., ultrafilters, and there are four types of them.

First of all, every element $a \in k$ defines an ultrafilter which is formed by the sets (finite unions of standard sets) that contain the point a, i.e., a base of this ultrafilter is formed by closed discs with center at a. Furthermore, every closed disc $E(a;r)$ of radius $r > 0$ with center at $a \in k$ defines an ultrafilter. If $r \in |k^*|$, a base of the corresponding ultrafilter $p(E(a;r))$ is formed by standard sets of the form $E(a;r')\backslash \bigcup_{i=1}^{n} D(a_i;r_i)$ with $r_i < r < r'$, $|a_i - a| \le r$ and $|a_i - a_j| = r$ for $1 \le i \ne j \le n$, where $D(a_i;r_i)$ is the open disc of radius r_i with center at a_i. If $r \notin |k^*|$, a base of the corresponding ultrafilter $p(E(a;r))$ is formed by the closed annuli $E(a;r')\backslash D(a;r'')$ with $r'' < r < r'$. Finally, if the field k is not maximally complete, then every family of nested discs with empty intersection is a base of an ultrafilter. (By the way, it is easy to see that there is a natural bijection between the set of ultrafilters and the set of nested families of closed discs.) The above four types of ultrafilters correspond to what are now known as points of types (1)-(4) of the affine line, and elements of the ultrafilters are precisely affinoid neighborhoods of those points.

In fact, as soon as I found the above description, I knew that the space of all ultrafilters must be considered as the affine line $\mathbf{A}^1$ over k. This space is endowed with a natural topology with respect to which it is locally compact: its basis consists of sets of ultrafilters which contain a given standard set. It is also endowed with a natural sheaf of local rings, the sheaf of analytic functions $\mathcal{O}_{\mathbf{A}^1}$. But my main reason to view the space $\mathbf{A}^1$ as the affine line was the fact that it provided an answer to the question on the spectrum of a bounded linear operator I posed to myself at the very beginning. Namely, the spectrum of such an operator is a nonempty compact subset of $\mathbf{A}^1$, and it coincides with the complement of the analyticity set of its resolvent. The field k is naturally embedded in the affine line as a dense subset, and the operators studied by Misha Vishik were precisely those with spectrum contained in k.

It was a pleasant exercise to extend Vishik's results to arbitrary operators, and it helped me to understand better the topological tree-like structure of the non-archimedean affine and projective lines, to get used to them, and to accept them as reality. During this work I met with Misha several times to tell him about the progress. At that time, he was the only person (besides my wife) who shared my excitement about all this. The usual reaction of my colleagues was simple indifference at best, and the quite understandable reason for that was nicely expressed by Professor Manin. When I told him about what I was doing, he observed that it is worthwhile to develop a general theory only having in mind a concrete problem. Of course, understanding what the spectrum of a non-archimedean operator should be was not a concrete problem. Had I followed this wise advice, I would have turned back to concrete problems I had in abundance in the area of computers, and would probably have become rich during the present age of the high tech boom since I was a really good programmer. Fortunately, I was already stupid enough to miss such an attractive opportunity, and I continued my exploration of the unknown new world revealed to me by a fluke.

My job occupied me five days per week from 8am till 5pm. It took me several years to learn to devote an hour or two to mathematics during working hours. Time free from my job belonged to my family, and when I was completely hooked on mathematics and an hour or two per day was not enough for it, I discovered an additional source of time. I learned to get up every day very early (often as early as at 2am), and thus extended the time for doing mathematics. At this time of the day, the world around me was quiet and fresh, nobody and nothing disturbed me, my head was clear, and I could plunge into another world to explore and describe it.

When I had finished writing everything I had in mind, I could look at it quietly and listen to an inner feeling that something was not satisfactory. I thought about this from time to time more and more often, but could not even express what tormented me. One day at the very end of 1985 all this obsessed me. I could not stop thinking about it at my job, and later at home. I did not go to bed early as usual. The right question and an immediate answer to it came early the next morning.

As I mentioned above, the affine line $\mathbf{A}^1$ is provided with a sheaf of rings $\mathcal{O}_{\mathbf{A}^1}$. Its stalk $\mathcal{O}_{\mathbf{A}^1,x}$ at a point x is a local ring with residue field $\kappa(x)$ provided with a valuation that extends that on k. If x is of type (1) (i.e., corresponds to an element of k), then $\mathcal{O}_{\mathbf{A}^1,x}$ is the algebra of convergent power series at that point, and so $\kappa(x) = k$. Otherwise, it is a field of infinite degree over k, and so it coincides with the non-complete field $\kappa(x)$. If $\mathcal{H}(x)$ denotes the completion of $\kappa(x)$, one gets a character $k[T] \to \mathcal{H}(x)$ over k. The question that came to me on that early morning was the following. What are all possible characters $k[T] \to K$ to non-archimedean fields K over k?

After the above question had been formulated, I already knew the answer: every such character goes through a character $k[T] \to \mathcal{H}(x)$ for a unique point $x \in \mathbf{A}^1$. The proof is very easy. Indeed, given a character $\chi : k[T] \to K$, consider the family of closed discs of the form $E(a; |\chi(T-a)|)$ with $a \in k$ and $\chi(T-a) \neq 0$. It is easy to see that it is a nested family of discs and, if x is the corresponding point of $\mathbf{A}^1$, the character χ goes through the character $k[T] \to \mathcal{H}(x)$.

Thus, the affine line $\mathbf{A}^1$ can be defined as the set of equivalence classes of characters $k[T] \to K$ to non-archimedean fields K over k or, equivalently, as the set of all multiplicative seminorms on $k[T]$ that extend the valuation on k. Wow, this definition was so simple and easily seen to be applicable in a much more general setting (e.g., for defining affine spaces of higher dimension). It also gave a clear idea how to define the non-archimedean analog of the Gelfand spectrum of a complex commutative Banach algebra. But the main thing I was struck by was the fact that this definition was also applicable in the classical situation and gave the corresponding classical objects. In this way, I was thrown into the new (for me) area of analytic geometry. It took me several days to calm down and to quietly look at what all that meant.

The above observation made it clear how to define analytic spaces over an arbitrary field k complete with respect to a nontrivial valuation (archimedean or not). First of all, one should start one step earlier and define the affine space $\mathbf{A}^n$ as the set of multiplicative seminorms on the ring of polynomials $k[T_1, \dots, T_n]$ that extend the valuation on k. The space $\mathbf{A}^n$ is endowed with the evident topology, and each point $x \in \mathbf{A}^n$ defines a character $k[T_1, \dots, T_n] \to \mathcal{H}(x) : f \mapsto f(x)$ to a complete valuation field $\mathcal{H}(x)$ over k so that the corresponding seminorm is the function $f \mapsto |f(x)|$. Furthermore, as we were taught by Krasner, an analytic function f on an open subset $\mathcal{U} \subset \mathbf{A}^n$ should be defined as a local limit of rational functions. The latter means that f is a map that takes each point $x \in \mathcal{U}$ to an element $f(x) \in \mathcal{H}(x)$ with the following property: one can find an open neighborhood $x \in \mathcal{U}' \subset \mathcal{U}$ such that, for every $\varepsilon > 0$, there exist polynomials $g, h \in k[T_1, \dots, T_n]$ with $h(x') \neq 0$ and $\left| f(x') - \frac{g(x')}{h(x')} \right| < \varepsilon$ for all points $x' \in \mathcal{U}'$. Finally, arbitrary analytic spaces are those locally ringed spaces which are locally isomorphic to a local model of the form $(X, \mathcal{O}_X)$, where X is the set of common zeros of a finite system of analytic functions $f_1, \dots, f_m$ on an open subset $\mathcal{U} \subset \mathbf{A}^n$ and $\mathcal{O}_X$ is the restriction of the quotient $\mathcal{O}_\mathcal{U}/\mathcal{J}$ by the subsheaf of ideals $\mathcal{J} \subset \mathcal{O}_\mathcal{U}$ generated by $f_1, \dots, f_m$. By the way, the spectrum $\mathcal{M}(\mathcal{A})$ of a commutative Banach k-algebra $\mathcal{A}$ should be defined as the space of all bounded multiplicative seminorms on A.

If $k = \mathbf{C}$, the affine space $\mathbf{A}^n$ is the maximal spectrum of the ring of polynomials $\mathbf{C}[T_1, \dots, T_n]$ (i.e., the vector space $\mathbf{C}^n$), analytic functions are local limits of polynomials, and the spectrum of a complex commutative Banach algebra is the Gelfand space of its maximal ideals. If $k = \mathbf{R}$, the above construction gives a new object: the real analytic affine space $\mathbf{A}^n$ is the maximal spectrum of the ring of polynomials $\mathbf{R}[T_1, \dots, T_n]$ (i.e., the quotient of $\mathbf{C}^n$ by the complex conjugation), and local limits of polynomials with real coefficients are not enough to define analytic functions.

The above definition of an analytic space was a lodestar in my journey, but I was unable to work with it directly. The difficulty was in establishing functional analytic properties of the analytic spaces, whereas establishing their geometric properties was much easier. Fortunately, the fundamental paper by John Tate on rigid analytic

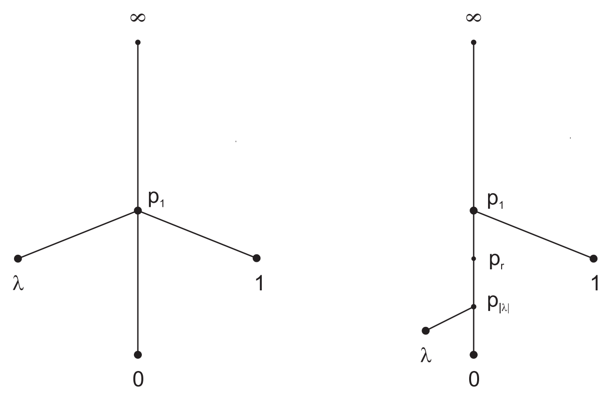

Figure 1

spaces was available since it was translated and published in the Soviet Union (even before it was published in the West). It was Tate's theory of affinoid algebras and affinoid domains that compensated for the lack of the usual complex analytic tools in the non-archimedean world. I studied Tate's work intensively and adjusted to the new framework, introducing a category of analytic spaces which eventually coincided with that given above. In the present framework, it is precisely the full subcategory of the category of analytic spaces consisting of the spaces without boundary. They are glued from the interiors of affinoid spaces (the interior of $\mathcal{M}(\mathcal{A})$ consists of the points x for which the corresponding character $\mathcal{A} \to \mathcal{H}(x)$ is a completely continuous operator), and include the analytifications of algebraic varieties. The new analytic spaces were applied to define the common spectrum of a finite family of commuting bounded linear operators, to develop holomorphic functional calculus, and to prove the Shilov idempotent theorem. The latter states that for any open and closed subset of the spectrum $\mathcal{M}(\mathcal{A})$ of a commutative Banach algebra $\mathcal{A}$ there exists a unique idempotent $e \in \mathcal{A}$ which is equal to 1 precisely at that subset.

At that time I found that the process of writing down of what I was starting to understand was very enjoyable and extremely helpful for better understanding. The need to express an idea forced me to concentrate on each small object or detail of reasoning. This concentration helped me see hidden and refined nuances which could change the whole picture, or to discover again and again a deep-rooted prejudice or wrong vision or simple stupidity.

The first typewritten text was finished in April of 1986, and I succeeded in passing it to Professor Barry Mazur, who knew me from my previous work. Later on, I was surprised to learn that my text had been accepted by the American Mathematical Society for publication as a book. But I was actually lucky that everybody at AMS immediately forgot about me finishing that book, and so I could continue to rewrite the text infinitely many times, gradually extending the framework of the new analytic spaces and investigating their amazing properties.

Tate's paper was still the only source of my knowledge in rigid analytic geometry, when I considered the following situation. Assume that the ground field k is algebraically closed and the characteristic of its residue field is not 2, and let $\mathcal{E}$ be an elliptic curve over k defined by the affine equation $y^2 = x(x-1)(x-\lambda)$ with

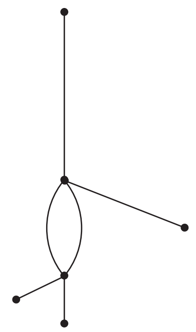

FIGURE 2

$0 < |\lambda| \le 1$ and $|\lambda - 1| = 1$. (The latter can be always achieved after a change of coordinates.) Recall that the projective line $\mathbf{P}^1$ has the property that any two different points can be connected by a unique path. Let us connect the points 0, 1, λ and ∞ in $\mathbf{P}^1$. We get one of the two graphs Γ_λ presented in Figure 1, which correspond to the cases $|\lambda| = 1$ and $|\lambda| < 1$. (For brevity the point $p(E(0;r))$ is denoted by p_r.) The complement of Γ_λ in $\mathbf{P}^1$ is a disjoint union of open discs of the form $D(a;r_a)$ with $a \in k\backslash\{0,1,\lambda\}$ and $r_a = \min\{|a|,|a-1|,|a-\lambda|\}$. Every such disc is glued to its boundary which is a point of Γ_λ, and Γ_λ is a strong deformation retraction of the whole projective line $\mathbf{P}^1$. Consider now the x-projection $\pi : \mathcal{E}^{\mathrm{an}} \to \mathbf{P}^1$ from the analytification $\mathcal{E}^{\mathrm{an}}$ of $\mathcal{E}$. Since the characteristic of the residue field of k is not two, the square root of each of the linear factors of $x(x-1)(x-\lambda)$ can be extracted at $D(a;r_a)$ and, therefore, the preimage of $\pi^{-1}(D(a;r_a))$ is a disjoint union of two open discs which are glued to their boundaries at the preimage $\pi^{-1}(\Gamma_\lambda)$. Thus, the latter is a strong deformation retraction of $\mathcal{E}^{\mathrm{an}}$. If $0 < r < |\lambda|$, then the square roots of $x - \lambda$ and $x - 1$ are extracted at the open annulus $D(0;r+\varepsilon)\backslash E(0;r-\varepsilon)$ with $0 < r-\varepsilon < r+\varepsilon < |\lambda|$, but the square root of x is not. This means that each point of the interval that connects 0 with $p_{|\lambda|}$ has a unique preimage in $\mathcal{E}^{\mathrm{an}}$. Similarly, each point from the intervals that connect 1 with p_1, λ with $p_{|\lambda|}$, and ∞ with p_1, has a unique preimage in $\mathcal{E}^{\mathrm{an}}$. In particular, if $|\lambda| = 1$, then $\pi^{-1}(\Gamma_\lambda) \xrightarrow{\sim} \Gamma_\lambda$. If now $|\lambda| < 1$, then the square roots of $x - 1$ and of the product $x(x - \lambda)$ are extracted at the open annulus $D(0;r+\varepsilon)\backslash E(0;r-\varepsilon)$ with $|\lambda| < r-\varepsilon < r+\varepsilon < 1$. This means that each point p_r with $|\lambda| < r < 1$ has two preimages in $\mathcal{E}^{\mathrm{an}}$, and the graph $\pi^{-1}(\Gamma_\lambda)$ has the form presented in Figure 2.

Thus, the analytic curve $\mathcal{E}^{\mathrm{an}}$ is contractible if $|\lambda| = 1$, and homotopy equivalent to a circle if $|\lambda| < 1$. It is well known that these two cases correspond to those when the modular invariant $j(\mathcal{E})$ is integral or not. But the latter case $|j(\mathcal{E})| > 1$ is precisely that of a Tate elliptic curve. Wow, such a curve is homotopy equivalent to a circle! I was always fascinated by Tate elliptic curves, but never understood the reason for which they admit uniformization. And here I had a very elementary explanation of this astonishing phenomenon discovered by Tate; it reminded me of the classical construction of the Riemann surface of an algebraic function.

Of course, all this strongly lifted up my spirit and eagerness in exploration of the new spaces. This was very timely since it distracted me from a serious health problem I had at that time, not to mention my job in a dull institution and the reality of a country in an advanced stage of decaying.

The great day of liberation for me and my family came in August, 1987, when we got out of the cesspool of the Soviet Union and arrived in the wonderful, sunny and crazy State of Israel. Although I was not so young, I was given an opportunity to renew my scientific career without also having to earn my family's living in some other way. I was again hungry for mathematics as never before, and ready for the next steps in my journey.