$J(1,\mathcal{F},F)<\infty$ $J(\delta,\mathcal{F},F)=\int_0^\delta \sup_Q \sqrt{\log N(\;)}\,d\varepsilon$

7 Vapnik-Červonenkis (VC) classes of sets/functions

Consider our canonical setting: $X_1, \ldots, X_n$ are i.i.d. P on some space $\mathcal{X}$. In this section we study classes of functions $\mathcal{F}$ (on $\mathcal{X}$) that satisfy certain *combinatorial restrictions*. These classes at first sight may seem have nothing to do with entropy numbers, but indeed will be shown to imply bounds on the covering numbers of the type

$$\sup_Q N(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \le K\left(\frac{1}{\epsilon}\right)^V, \qquad 0<\epsilon<1, \text{ some number } V>0,$$

where $\mathcal{F}$ is the underlying function class with envelope F, and K is a universal constant. Note that this has direct implications on the uniform entropy of such a class (see Definition 4.6) is of the order $\log(1/\epsilon)$ and hence the uniform entropy integral converges, and is of the order $\delta\log(1/\delta)$, as $\delta \downarrow 0$.

Classes of (indicator functions of) this type were first studied by Vapnik and Červonenkis in the 1970s, whence the name VC classes. There are many examples of VC classes, and more examples can be constructed by operations as unions and sums. Furthermore, one can combine VC classes in different sorts of ways (thereby, building larger classes of functions) to ensure that the resulting larger classes also satisfy the uniform entropy condition (though these larger classes may not necessarily be VC).

We first consider VC classes to sets. To motivate this study let us consider a boolean class of functions $\mathcal{F}$[55], i.e., every $f \in \mathcal{F}$ takes values in $\{0,1\}$. Thus,

$$\mathcal{F} = \{\mathbf{1}_C : C \in \mathcal{C}\},$$

where $\mathcal{C}$ is a collection of subsets of $\mathcal{X}$. This naturally leads to the study of $\mathcal{C}$.

Definition 7.1. *Let $\mathcal{C}$ be a collection of subsets of a set $\mathcal{X}$. Let $\{x_1,\ldots,x_n\} \subset \mathcal{X}$ be an arbitrary set of n points. Say that $\mathcal{C}$ picks out a certain subset A of $\{x_1,\ldots,x_n\}$ if A can be expressed as $C \cap \{x_1,\ldots,x_n\}$ for some $C \in \mathcal{C}$.*

$A=C\cap\{x_1,\cdots x_n\}$

The collection $\mathcal{C}$ is said to shatter $\{x_1,\ldots,x_n\}$ if each of its 2^n subsets can be picked out in this manner (note that an arbitrary set of n points possesses 2^n subsets).

Definition 7.2. *The VC dimension $V(\mathcal{C})$ of the class $\mathcal{C}$ is the largest n such that some set of size n is shattered by $\mathcal{C}$.*

Definition 7.3. *The VC index $\Delta_n(\mathcal{C}; x_1,\ldots,x_n)$ is defined as*

$$\Delta_n(\mathcal{C}; x_1,\ldots,x_n) = |\{C \cap \{x_1,\ldots,x_n\} : C \in \mathcal{C}\}|,$$

where $|A|$ denotes the cardinality of the set A. Thus,

$$V(\mathcal{C}) := \sup\Big\{n : \max_{x_1,\ldots,x_n\in\mathcal{X}} \Delta_n(\mathcal{C}; x_1,\ldots,x_n) = 2^n\Big\}.$$

[55] Boolean classes $\mathcal{F}$ arise in the problem of classification (where $\mathcal{F}$ can be taken to consist of all functions f of the form $I\{g(X) \neq Y\}$). They are also important for historical reasons: empirical process theory has its origins in the study of the function class $\mathcal{F} = \{\mathbf{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}\}$.

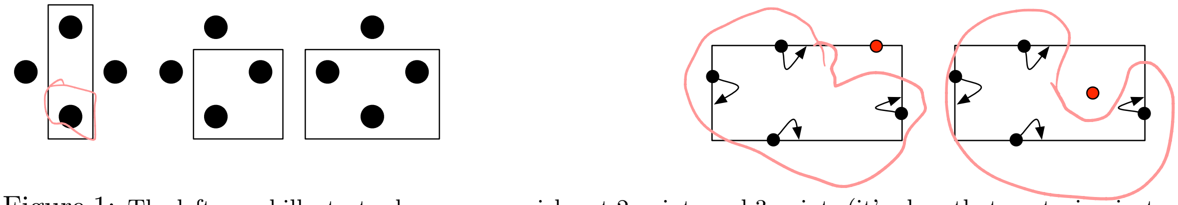


Figure 1: The left panel illustrates how we can pick out 2 points and 3 points (it's clear that capturing just 1 point and all 4 points are both trivial) thereby showing that there exist 4 points that can be shattered. The right panel illustrates that no set of 5 points can be shattered: the minimum enclosing rectangle that allows us to select all 5 points is defined by only four points — one for each edge. So, it is clear that the fifth point must lie either on an edge or on the inside of the rectangle thereby preventing us from selecting four points without the fifth.

Clearly, the more refined $\mathcal{C}$ is, higher the VC index. The VC dimension is infinite if $\mathcal{C}$ shatters sets of arbitrarily large size. It is immediate from the definition that $V(\mathcal{C}) \leq V$ if and only if no set of size $V + 1$[56] is shattered.

Example 7.4. *Let $\mathcal{X} = \mathbb{R}$ and define the collection of sets $\mathcal{C} := \{(-\infty, c] : c \in \mathbb{R}\}$. Consider any two point set $\{x_1, x_2\} \subset \mathbb{R}$, and assume without loss of generality, that $x_1 < x_2$. It is easy to verify that $\mathcal{C}$ can pick out the null set $\{\}$ and the sets $\{x_1\}$ and $\{x_1, x_2\}$ but cannot pick out $\{x_2\}$. Hence its VC dimension equals 1.*

$c < x_1$

$c \ \{x_1 \ x_2 \ c$

The collection of all cells $(a, b] \in \mathbb{R}$ shatters every two-point set but cannot pick out the subset consisting of the smallest and largest points of any set of three points. Thus its VC dimension equals 2.

$a \ x_1 \ x_2 \ x_3 \ b$ $\{x_1, x_3\}$

Remark 7.1. *With more effort, it can be seen that the VC ~~indices~~ dimension of the same type of sets in $\mathbb{R}^d$ are d and $2d$, respectively. For example, let $\mathcal{X} = \mathbb{R}^2$ and define*

$$\mathcal{C} = \{A \subset \mathcal{X} : A = [a, b] \times [c, d], \text{ for some } a, b, c, d \in \mathbb{R}\}.$$

Let us see what happens when $n = 4$. Draw a figure to see this when the points are not co-linear. We can show that there exists 4 points such that all the possible subsets of these four points are picked out by $\mathcal{C}$; see the left panel of Figure 7.1.

Now if we have $n = 5$ points things change a bit; see the right panel of Figure 7.1. If we have five points there is always one that stays "in the middle" of all the others, and thus the complement set cannot be picked out by $\mathcal{C}$. We immediately conclude that the VC dimension of $\mathcal{C}$ is 4.

A collection of measurable sets $\mathcal{C}$ is called a *VC class* if its dimension is finite. The main result of this section is the remarkable fact that the covering numbers of any VC class grow polynomially in $1/\epsilon$ as $\epsilon \to 0$, of order dependent on the dimension of the class.

[56] Some books define the *VC index* of the class $\mathcal{C}$ as the smallest n for which no set of size n is shattered by $\mathcal{C}$ (i.e., $V(\mathcal{C}) + 1$ in our notation).

Example 7.5. *Suppose that $\mathcal{X} = [0,1]$, and let $\mathcal{C}$ be the class of all finite subsets of $\mathcal{X}$. Let P be the uniform (Lebesgue) distribution on $[0,1]$. Clearly $V(\mathcal{C}) = \infty$ and $\mathcal{C}$ is not a VC class. Note that for any possible value of $\mathbb{P}_n$ we have $\mathbb{P}_n(A) = 1$ for $A = \{X_1, \ldots, X_n\}$ while $P(A) = 0$. Therefore $\|\mathbb{P}_n - P\|_{\mathcal{C}} = 1$ for all n, so $\mathcal{C}$ is not a Glivenko-Cantelli class for P.*

$\|\mathbb{P}_n - P\|_{\mathcal{C}} \to 0$ -

Exercise (HW3): Show that the class of all closed and convex sets in $\mathbb{R}^d$ does not have finite VC dimension (Hint: Consider a set of n points on the boundary of the unit ball).

Sauer's lemma[57] (also known as Sauer-Shelah-Vapnik-Červonenkis lemma), one of the fundamental results on VC dimension, states that the number $\Delta_n(\mathcal{C}; x_1, \ldots, x_n)$ of subsets picked out by a VC class $\mathcal{C}$, for $n \geq 1$, satisfies:

$$\max_{x_1,\ldots,x_n} \Delta_n(\mathcal{C}; x_1, \ldots, x_n) \leq \sum_{j=0}^{V(\mathcal{C})} \binom{n}{j}, \tag{73}$$

where we use the notation $\binom{n}{j} = 0$ if $j > n$. Observe that for $n \leq V(\mathcal{C})$, the right-hand side of the above display equals 2^n, i.e., the growth is exponential. However, it is easy to show[58] that for $n \geq V(\mathcal{C})$,

$$\sum_{j=0}^{V(\mathcal{C})} \binom{n}{j} \leq \left(\frac{ne}{V(\mathcal{C})}\right)^{V(\mathcal{C})}. \tag{74}$$

Consequently, the numbers on the left side grow polynomially (of order at most $O(n^{V(\mathcal{C})})$) rather than an exponential number. Intuitively this means that a finite VC index implies that $\mathcal{C}$ has an apparent simplistic structure.

7.1 VC classes of Boolean functions

The definition of VC dimension can be easily extended to a function class $\mathcal{F}$ in which every function f is binary-valued, taking the values $\{0,1\}$ (say). In this case, we define, for every

[57] See [van der Vaart and Wellner, 1996, pages 135–136] for a complete proof of the result.

[58] In the following we just give a proof of the right-hand inequality of (74). Note that with $Y \sim \text{Binomial}(n, 1/2)$,

$$\begin{aligned}
\sum_{j=0}^{V(\mathcal{C})} \binom{n}{j} &= 2^n \sum_{j=0}^{V(\mathcal{C})} \binom{n}{j} \left(\frac{1}{2}\right)^n = 2^n \mathbb{P}(Y \leq V(\mathcal{C})) \\
&\leq 2^n \mathbb{E}[r^{Y - V(\mathcal{C})}] \qquad \text{for } r \leq 1 \qquad \left[\text{as } \mathbf{1}\{Y - V(\mathcal{C}) \leq 0\} \leq r^{Y - V(\mathcal{C})} \text{ for } r \leq 1\right] \\
&= 2^n r^{-V(\mathcal{C})} \left(\frac{1}{2} + \frac{r}{2}\right)^n = r^{-V(\mathcal{C})}(1+r)^n \qquad \left[\text{as } \mathbb{E}[r^Y] = \sum_{j=0}^{n} r^j \binom{n}{j} \left(\frac{1}{2}\right)^n = \left(\frac{1}{2} + \frac{r}{2}\right)^n\right] \\
&= \left(\frac{n}{V(\mathcal{C})}\right)^{V(\mathcal{C})} \left(1 + \frac{V(\mathcal{C})}{n}\right)^n \qquad \text{by choosing } r = V(\mathcal{C})/n \\
&\leq \left(\frac{n}{V(\mathcal{C})}\right)^{V(\mathcal{C})} e^{V(\mathcal{C})}.
\end{aligned}$$

$E[1\{Y - V(C) \le 0\}]$.

$(1+r)^n$

$\left(1+\frac{k}{n}\right)^n \to e^k$.

$x_1, \ldots, x_n \in \mathcal{X}$,

$$\mathcal{F}(x_1, \ldots, x_n) := \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}. \tag{75}$$

As functions in $\mathcal{F}$ are Boolean, $\mathcal{F}(x_1, \ldots, x_n)$ is a subset of $\underline{\{0,1\}^n}$.

Definition 7.6. *Given such a function class $\mathcal{F}$ we say that the set $\{x_1, \ldots, x_n\}$ is shattered by $\mathcal{F}$ if*

$$\Delta_n(\mathcal{F}; x_1, \ldots, x_n) := |\mathcal{F}(x_1, \ldots, x_n)| = \boxed{2^n}.$$

The VC dimension $V(\mathcal{F})$ of $\mathcal{F}$ is defined as the largest integer n for which there is some collection $x_1, \ldots, x_n$ of n points that can be shattered by $\mathcal{F}$.

When $V(\mathcal{F})$ is finite, then $\mathcal{F}$ is said to be a VC class. $\frac{1}{n}\log N(\varepsilon, \mathcal{F}, L_1(P_n)) \to 0$

Example 7.7. *Let us revisit the Glivenko-Cantelli (GC) theorem (Theorem 3.5) when we have a binary-valued function class $\mathcal{F}$. In particular, suppose that $X_1, \ldots, X_n$ are i.i.d. P on $\mathcal{X}$. A natural question is how does one verify condition (8) in practice? We need an upper bound on $N(\epsilon, \mathcal{F}, L_1(\mathbb{P}_n))$. Recall that under $L_1(\mathbb{P}_n)$ the distance between f and g is measured by*

$$\|f - g\|_{L_1(\mathbb{P}_n)} := \underline{\frac{1}{n}\sum_{i=1}^{n} |f(X_i) - g(X_i)|.} \le \varepsilon \qquad \varepsilon \geq 1.$$

This notion of distance clearly only depends on the values of f and g at the data points $X_1, \ldots, X_n$. Therefore, the covering number of $\mathcal{F}$ in the $L_1(\mathbb{P}_n)$-norm should be bounded from above by the corresponding covering number of $\{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}$. It should be obvious that $N(\epsilon, \mathcal{F}, L_1(\mathbb{P}_n))$ is bounded from above by the cardinality of $\mathcal{F}(X_1, \ldots, X_n)$, i.e.,

$$\underline{N(\epsilon, \mathcal{F}, L_1(\mathbb{P}_n))} \le \underline{\overset{= 2^n}{|\mathcal{F}(X_1, \ldots, X_n)|}} \qquad \text{for every } \epsilon > 0.$$

This is in fact a very crude upper bound although it can be quite useful in practice. For example, in the classical GC theorem $\mathcal{F} := \{\mathbf{1}_{(-\infty,t]}(\cdot) : t \in \mathbb{R}\}$, and we can see that $|\mathcal{F}(X_1, \ldots, X_n)| \le (n+1)$. $2+1$ $n+1$ x_1 $\emptyset, \{x_1\}$ $\{x_1, x_3\}$

Since $\mathcal{F}(X_1, \ldots, X_n)$ is a subset of $\{0, 1\}^n$, its maximum cardinality is $\boxed{2^n}$. But if $\underline{\Delta_n(\mathcal{F}; X_1, \ldots, X_n)}$ is at the most a polynomial in n for every possible realization of $X_1, \ldots, X_n$, then $n^\alpha < 2^n$

$$\underline{\frac{1}{n}\log \Delta_n(\mathcal{F}; X_1, \ldots, X_n) \to 0} \qquad \text{as } n \to \infty \text{ a.s.} \tag{76}$$

which implies, by Theorem 3.5, that $\mathcal{F}$ is GC. Thus, if $\mathcal{F}$ is a boolean function class such that (76) holds, then $\mathcal{F}$ is P-GC.

Exercise (HW3): Consider the class of all two-sided intervals over the real line, i.e., $\mathcal{F} := \{\mathbf{1}_{(a,b]}(\cdot) : a < b \in \mathbb{R}\}$. Show that $\Delta_n(\mathcal{F}; X_1, \ldots, X_n) \le (n+1)^2$ a.s.

Exercise (HW3): For a scalar $t \in \mathbb{R}$, consider the function $f_t(x) := \mathbf{1}\{\sin(tx) \geq 0\}$, $x \in [-1, 1]$. Prove that the function class $\{f_t : [-1, 1] \to \mathbb{R} : t \in \mathbb{R}\}$ has infinite VC dimension (Note that this shows that VC dimension is not equivalent to the number of parameters in a function class).

7.2 Covering number bound for VC classes of sets

Theorem 7.8. *There exists a universal constant K such that for any VC class $\mathcal{C}$ of sets, any probability measure Q, any $r \geq 1$, and $0 < \epsilon < 1$,*

$$N(\epsilon, \mathcal{C}, L_r(Q)) \leq K\ V(\mathcal{C})(4e)^{V(\mathcal{C})} \left(\frac{1}{\epsilon}\right)^{rV(\mathcal{C})}. \tag{77}$$

Proof. See [van der Vaart and Wellner, 1996, Theorem 2.6.4]. □

In the following we will prove a slightly weaker version of the above result.

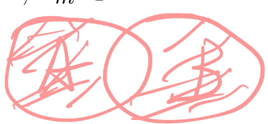
Theorem 7.9. *For any VC class $\mathcal{C}$ of sets, any $r \geq 1$, and $0 < \epsilon < 1$,*[59]

$$\sup_Q N(\epsilon, \mathcal{C}, L_r(Q)) \leq \left(\frac{c_1}{\epsilon}\right)^{rc_2V(\mathcal{C})} \tag{78}$$

Here c_1 and c_2 are universal positive constants and the supremum is over all probability measures Q on $\mathcal{X}$.

Proof. Fix $0 < \epsilon < 1$. Let $X_1, \ldots, X_n$ be i.i.d. Q. Let $m := D(\epsilon, \mathcal{C}, L_1(Q))$ be the ϵ-packing number for the collection $\mathcal{C}$ in the norm $L_1(Q)$. Thus, there exists $C_1, \ldots, C_m \in \mathcal{C}$ which satisfy

$A\triangle B=(A-B)\cup(B-A)$



$$Q|\mathbf{1}_{C_i} - \mathbf{1}_{C_j}| = Q(C_i \triangle C_j) > \epsilon, \qquad i \neq j.$$

Let $\mathcal{F} := \{\mathbf{1}_C : C \in \mathcal{C}\}$. We consider this function class view point as it is sometimes more natural than working with the collection of sets $\mathcal{C}$. Note that, $\{f_i \equiv \mathbf{1}_{C_i}\}_{i=1}^m$ is a set of m ϵ-separated functions in $\mathcal{F}$ in the $L_1(Q)$-metric, as, for $i \neq j$,

$$\epsilon < \int |f_i - f_j| dQ = Q\{f_i \neq f_j\} = Q(C_i \triangle C_j) = \mathbb{P}[X_1 \in C_i \triangle C_j].$$

By the above, we have

$(1+x) \leq \left(1+\frac{x}{n}\right)^n$

$\downarrow$

e^x

$>\epsilon$

$$\mathbb{P}[f_i(X_1) = f_j(X_1)] = 1 - \mathbb{P}[f_i(X_1) \neq f_j(X_1)] = 1 - \mathbb{P}[X_1 \in C_i \triangle C_j] < 1 - \epsilon \leq e^{-\epsilon}.$$

By the independence of $X_1, \ldots, X_n$ we deduce then that for every $k \geq 1$,

$$\mathbb{P}[f_i(X_1) = f_j(X_1), \ldots, f_i(X_k) = f_j(X_k)] \leq e^{-k\epsilon}.$$

In words, this means that the probability that f_i and f_j agree on every $X_1, \ldots, X_k$ is at most $e^{-k\epsilon}$. By the union bound, we have

$\frac{m(m-1)}{2}$

$$\mathbb{P}\left[(f_i(X_1), \ldots, f_i(X_k)) = (f_j(X_1), \ldots, f_j(X_k)) \text{ for some } 1 \leq i < j \leq m\right] \leq \binom{m}{2} e^{-k\epsilon} \leq \frac{m^2}{2} e^{-k\epsilon}.$$

Recalling that $\mathcal{F}(x_1, \ldots, x_k) = \{(f(x_1), \ldots, f(x_k)) : f \in \mathcal{F}\}$, this immediately gives

$$\mathbb{P}[|\mathcal{F}(X_1, \ldots, X_k)| \geq m] \geq 1 - \frac{m^2}{2} e^{-k\epsilon}.$$

$\frac{m^2}{2} e^{-\frac{2\log m}{\epsilon}\cdot 8}$

$\geq \frac{1}{2}$

$\leq$

[59] Note that $N(\epsilon, \mathcal{C}, L_r(Q)) = 1$ for all $\epsilon \geq 1$.

Thus if we take $k := \left\lceil \frac{2\log m}{\epsilon} \right\rceil \geq \frac{2\log m}{\epsilon}$, then, $\mathbb{P}[|\mathcal{F}(X_1, \ldots, X_k)| \geq m] \geq 1/2$. Thus for the choice of k above, there exists a subset $\{z_1, \ldots, z_k\}$ of cardinality k such that $|\mathcal{F}(z_1, \ldots, z_k)| \geq m$. We now apply the Sauer-Shelah-VC lemma and deduce that

$$m \leq |\mathcal{F}(z_1, \ldots, z_k)| \leq \max_{x_1, \ldots, x_k} \Delta_k(\mathcal{C}; x_1, \ldots, x_k) \leq \sum_{j=1}^{V(\mathcal{C})} \binom{k}{j}. \leq \left(\frac{k}{V(C)}\right)^{V(C)} e^{V(C)} \tag{79}$$

We now split into two cases depending on whether $k \leq V(\mathcal{C})$ or $k \geq V(\mathcal{C})$.

$$\sum_{j=1}^{V(C)} \binom{k}{j} = 2^{V(C)}$$

Case 1: $k \leq V(\mathcal{C})$. Here (79) gives

$$N(\epsilon, \mathcal{C}, L_1(Q)) \leq D(\epsilon, \mathcal{C}, L_1(Q)) = m \leq 2^{V(\mathcal{C})} \leq \left(\frac{2}{\epsilon}\right)^{V(\mathcal{C})}, 0 < \varepsilon < 1$$

which proves (78).

Case 2: $k \geq V(\mathcal{C})$. Here (79) gives

$$N(\epsilon, \mathcal{C}, L_1(Q)) = m \leq \left(\frac{ke}{V(\mathcal{C})}\right)^{V(\mathcal{C})},$$

so that using the choice of k which satisfies $k \leq \frac{4\log m}{\epsilon}$,

$$m^{1/V(\mathcal{C})} \leq \frac{ke}{V(\mathcal{C})} \leq \frac{4e}{V(\mathcal{C})\epsilon} \log m = \frac{8e}{\epsilon} \log m^{1/(2V(\mathcal{C}))} \leq \frac{8e}{\epsilon} m^{1/(2V(\mathcal{C}))},$$

where we have used $\log x \leq x$. This immediately gives

$$N(\epsilon, \mathcal{C}, L_1(Q)) \leq D(\epsilon, \mathcal{C}, L_1(Q)) = m \leq \left(\frac{8e}{\epsilon}\right)^{2V(\mathcal{C})},$$

which completes the proof of the result for $r = 1$.

For $L_r(Q)$ with $r > 1$, note that

$$\|\mathbf{1}_C - \mathbf{1}_D\|_{L_1(Q)} = Q(C \triangle D) = \|\mathbf{1}_C - \mathbf{1}_D\|^r_{L_r(Q)},$$

so that

$$N(\epsilon, \mathcal{C}, L_r(Q)) = N(\epsilon^r, \mathcal{C}, L_1(Q)) \leq \left(c_1 \epsilon^{-r}\right)^{c_2 V(\mathcal{C})}. \quad \left(\frac{c_1}{\varepsilon}\right)^{-r c_2 V(C)}$$

This completes the proof. □

<span style="color:red">Exercise (HW3)</span>: Suppose $\mathcal{F}$ is a Boolean class of functions with VC dimension $V(\mathcal{F})$. Then, for some constant $C > 0$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |(\mathbb{P}_n - P)f|\right] \leq C\sqrt{\frac{V(\mathcal{F})}{n}}.$$

Suppose $X_1, \ldots, X_n$ are i.i.d. real-valued observations having a common cdf F. Apply this result to obtain a high probability upper bound on $\sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)|$, i.e., show that

$$\sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \leq \frac{C}{\sqrt{n}} + \sqrt{\frac{2}{n} \log \frac{1}{\alpha}}$$

with probability at least $1 - \alpha$ (for $\alpha \in (0, 1)$).