МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ 2017 год, том 29, номер 3, стр. 42-50

УТОЧНЕНИЕ ПРЕЦИЗИОННЫХ АППРОКСИМАЦИЙ ФУНКЦИЙ ФЕРМИ-ДИРАКА ЦЕЛОГО ИНДЕКСА

© *2017 г.* ***Н.Н. Калиткин, С.А. Колганов***

Институт прикладной математики им. М.В. Келдыша РАН, Москва
Национальный исследовательский университет «МИЭТ», Зеленоград
e-mail: kalitkin@imamod.ru, mkandds2012@gmail.com

Работа поддержана грантом РНФ 16-11-10001.

Функции Ферми-Дирака целого индекса играют важную роль в задачах электронного переноса в плотной среде. Ранее для их быстрого вычисления были построены аппроксимации, использующие отношение многочленов. Они позволили получить относительную точность $\sim 2\cdot10^{-16}$ для индексов k=1, 2, 3. В данной работе использована библиотека boost::multiprecision языка C++, позволяющая проводить вычисления с произвольным числом знаков. Точность ранее полученных формул доведена до $\sim 5\cdot10^{-18}$ и построена аналогичная формула для индекса k=4. Показано также, что несложные глобальные формулы, состоящие из небольшого числа слагаемых, разумно описывают порядок величины функций при любых значениях аргумента и могут быть использованы для оценок.

Ключевые слова: функции Ферми-Дирака, прецизионные аппроксимации, рациональная аппроксимация, оценочные глобальные аппроксимации.

CORRECTION OF THE PRECISION APPROXIMATIONS OF THE FERMI-DIRAK FUNCTIONS OF INTEGER INDEX

N.N. Kalitkin, S.A. Kolganov

Keldysh Institute of Applied Mathematics of Rus. Acad. of Sci., Moscow
National Research University of Electronic Technology, Zelenograd
e-mail: kalitkin@imamod.ru , mkandds2012@gmail.com

Fermi-Dirac functions of integer index are widely used in problems of electronic transport in dense substances. Polynomial approximations were constructed for its quick computation. Such coefficients are founded for functions of index 1, 2, 3, which provide ratio error $2\cdot10^{-16}$ with 9 free parametrs. In this work we used C++ boost::multiprecision library, which allows to calculate with free number of digits. Precision of previously obtained formulas brought to $\sim5\cdot10^{-18}$ and the same formula has been built for the index k=4. It is also shown that simple global formulas, consisting of small number of parameters, reasonably describe the order of value of the functions for all values of the argument and can be used for estimations.

Key words: Fermi-Dirac functions, precision approximations, rational approximation, estimated global approximations.

1. Функции Ферми-Дирака. Функции Ферми-Дирака возникают в задачах квантовой механики при описании свойств вещества, обусловленных поведением электронов

(или других фермионов). При достаточно высоких плотностях или низких температурах они являются различными моментами фермиевского распределения и сводятся к следующим интегралам:

$$I_k(x) = \int_0^\infty \frac{t^k \mathrm{d}t}{1+\exp(t-x)}, \qquad x \in (-\infty;+\infty). \tag{1}$$

Индекс k принимает целые значения для нечетных моментов и полуцелые значения – для четных. В физических задачах нужны только целые и полуцелые индексы, хотя в математической теории этих функций рассматриваются произвольные k.

Укажем физические величины, соответствующие различным индексам. Плотности электронов соответствует $k = 1/2$, кинетической энергии – $k = 3/2$, электронной проводимости – $k = 1$, электронной теплопроводности – $k = 2$, электронной вязкости – $k = 3$. В ряде приложений возникают меньшие индексы (например, $k = -1/2$ и даже $k = -3/2$); но более высокие индексы до сих пор не требовались.

Аппроксимациям функций Ферми-Дирака полуцелого индекса посвящена обширная литература [1-7]. Для функций целых индексов $k = 1$, 2, 3 лишь недавно были построены прецизионные аппроксимации [8] и развита методика построения таких аппроксимаций. При этом в значительной части диапазона значений аргумента удалось добиться относительной точности $\sim 2\cdot 10^{-16}$, что соответствует погрешности 64-битовых вычислений. Однако более детальный анализ показал, что при значениях $x < -5$ ошибка начинает возрастать и при $x \sim -10$ увеличивается на ~ 4 порядка. Поэтому работа была повторена с использованием библиотеки boost::multiprecision языка C++, позволяющей вычислять с произвольным числом знаков. Был также разработан специальный прием вычисления вспомогательного аргумента. Все это позволило уменьшить погрешность до $\sim 5\cdot 10^{-18}$ во всем диапазоне значений x. Дополнительно была построена аппроксимация для индекса $k = 4$.

Кроме того, была исследована погрешность простейших глобальных аппроксимаций, описывающих весь диапазон $-\infty < x < +\infty$ единой гладкой формулой с небольшим числом коэффициентов. Такие аппроксимации дают грубую оценку функций Ферми-Дирака, причем их можно применять при произвольных нецелых индексах.

2. Вспомогательный аргумент. Интегралы (1) не берутся в элементарных функциях. Их асимптотики на двух концах вещественной оси имеют качественно различный вид: $I_k(x) \approx \Gamma(k+1)e^x$ при $x \to -\infty$, $I_k(x) \approx x^{k+1}/(k+1)$ при $x \to +\infty$. Это всегда сильно затрудняло построение хороших аппроксимаций. Однако в [8] был предложен перспективный подход. Существует единственный случай $k = 0$, когда интеграл (1) точно берется:

$$I_0(x) \equiv y(x) = \ln\left(1+e^x\right), \qquad 0 < y < +\infty. \tag{2}$$

Нетрудно видеть, что $I_k \approx \Gamma(k+1)\cdot y$ при $y \to 0$, $I_k \approx x^{k+1}/(k+1)$ при $y \to +\infty$. На обоих концах диапазона y асимптотики оказываются степенными, то есть однотипными.

Поэтому удобно взять величину y за вспомогательный аргумент и вычислять функции других индексов как функции этого аргумента. Это кардинально облегчает построение хороших аппроксимаций.

Однако при вычислениях на пределе точности компьютера возникает одна тонкость. Когда $x \to -\infty$, то $e^x \to 0$, и при вычислении $\ln(1+e^x)$ по стандартным компьютерным процедурам возможна большая потеря значащих цифр. Поэтому целесообразно использовать специальную процедуру вычислений:

$$y(x) = \sum_{n=0}^{\infty} \frac{2}{2n+1} \left(e^x / (2+e^x) \right)^{2n+1} \quad \text{при} \quad x \le 0; \tag{3}$$

ряд следует оборвать на числе членов, нужном для достижения заданной точности, а для суммирования целесообразно использовать схему Горнера. При $x>0$ вычисления $y(x)$ целесообразно проводить по стандартным компьютерным процедурам. Такой способ вычисления позволил ликвидировать возрастание погрешности, имевшееся в [8] при $x<-5$.

3. Прецизионная аппроксимация. Для диапазона $-\infty < x \le 0$ (соответственно $0 \le y \le \ln 2$) в [8] были предложены аппроксимирующие формулы, включающие отношения многочленов от y:

$$I_k(x) \approx \Gamma(k+1)\, y \left(\frac{\sum_{n=0}^{N+1} a_n y^n}{\sum_{m=0}^{N} b_m y^m} \right)^k, \quad a_0 = 1, \quad b_0 = 1, \quad x \le 0. \tag{4}$$

Для вычисления коэффициентов a_n, b_m при выбранном N был разработан специальный алгоритм, приближенно обеспечивающий чебышевский альтернанс для экстремумов относительной погрешности (очевидно, для аппроксимации функций (1) важна не абсолютная погрешность, а относительная). При этом относительная погрешность составляла $d \sim 10^{-6}$ при $N=1$, $d \sim 10^{-10}$ при $N=2$, $d \sim 10^{-14}$ при $N=3$. Следовало ожидать, что мы получим $d \sim 10^{-18}$ при $N=4$, но вычисления с 64-битовыми числами в принципе не могли обеспечить анализ ошибок $d \sim 10^{-16}$.

Чтобы полностью закрыть этот вопрос, мы провели расчеты с 25 десятичными знаками. Вдобавок, была использована формула (3) для промежуточного аргумента, что также несколько изменило далекие знаки коэффициентов. Окончательные значения коэффициентов для $N=4$ приведены в табл.1 с 18 десятичными знаками после запятой. В таблицу включены также коэффициенты для функций индекса $k=4$. Эти коэффициенты обеспечивают относительную погрешность $d \sim 5 \cdot 10^{-18}$ при $x \le 0$. Пример профиля относительной погрешности приведен на рис.1. Это гладкая кривая, у которой абсолютные величины экстремумов приблизительно одинаковые, а их знаки чередуются.

Видно, что коэффициенты быстро убывают с увеличением номера; это свидетельствует о быстрой сходимости рядов в числителе и знаменателе, то есть об удачном вы-

боре вида аппроксимации. Заметим, что в табл.1 для функции индекса $k = 4$ два коэффициента b_1 и b_3 оказались отрицательными. Обычно появление отрицательных коэффициентов считается нежелательным: в принципе они могут привести к обращению в нуль знаменателя (если это b_m) или числителя (если это a_n). Однако полученные коэффициенты настолько малы по модулю, что знаменатель всегда остается положительным. Так что отрицательность оказывается несущественной.

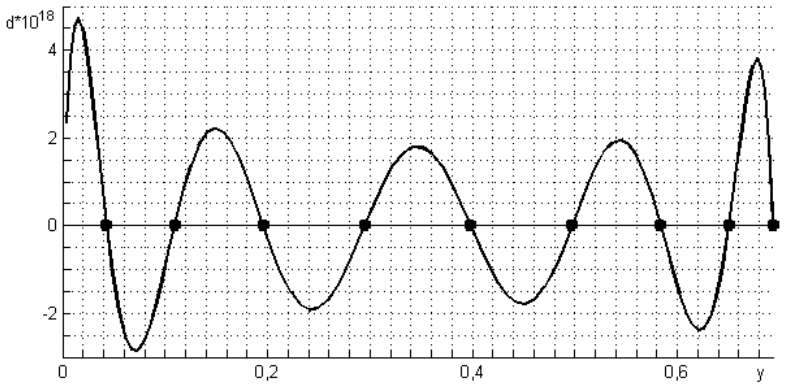

Рис.1. Профиль погрешности для $k = 2$.

Напомним, что при $x > 0$ вычисление функций целого индекса сводится к вычислению функций отрицательного аргумента по следующей формуле:

$$I_k(x) = \frac{x^{k+1}}{k+1}\left[1 + \frac{\Gamma(k+2)}{\Gamma(k)}\frac{\pi^2}{6}x^{-2} + \frac{\Gamma(k+2)}{\Gamma(k-2)}\frac{7\pi^4}{360}x^{-4} + \ldots\right] + (-1)^k I_k(-x), \quad k = 0,1,2,\ldots. \quad (5)$$

Число слагаемых в квадратных скобках конечно. Оно выбирается так, чтобы после умножения на x^{k+1} младшая степень многочлена была 1 при четном k и 0 при нечетном. Заметим также, что $I_1(0) = \pi^2/12,\ I_3(0) = 7\pi^4/120$.

Функцию отрицательного аргумента в правых частях следует вычислять по аппроксимациям (4). Поскольку $I_k(x)$ быстро возрастают с увеличением аргумента, то относительная погрешность при $x > 0$ будет гораздо меньше, чем при $x<0$; поэтому коэффициенты из табл.1 повсюду обеспечат относительную точность не хуже $\sim 5\cdot 10^{-18}$.

При вычислении коэффициентов по алгоритму [8] использовалось требование точной аппроксимации $I_k(x)$ при $x = 0$. Это обеспечивает непрерывность при переходе к положительному аргументу по (5) в точке $x = 0$.

Замечание. Обратим внимание вычислителей на одну тонкость использования библиотеки boost::multiprecision C++. В формулах могут использоваться численные коэффициенты, в том числе целочисленные. Целочисленные коэффициенты точно переводятся в двоичный код, поэтому в обычных расчетах нет необходимости указывать на ординарную или двойную точность их представления. Однако в boost::multiprecision C++ они по умолчанию воспринимаются как 64-битовые числа. Поэтому для получения по-

вышенного числа знаков надо описывать все коэффициенты, включая целочисленные как числа этой системы. Если этого не сделать, вы получите результаты только с 16 достоверными знаками.

Таблица 1. Коэффициенты a_n, b_m для $N = 4$.

k / a_n, b_m	1	2
a_1	0.271511313821436278	0.226381636434069856
a_2	0.056266123806058763	0.053368433557479886
a_3	0.006742074046934569	0.006290475634079521
a_4	0.000516950515533321	0.000502322827445298
a_5	0.000019477183676577	0.000018937967508806
b_1	0.021511313821435284	0.038881636434069113
b_2	0.023110517572972142	0.024304399874277445
b_3	0.000366908157736541	0.000629098532643319
b_4	0.000061042440873272	0.000065701816194546

Таблица 1. (продолжение).

k / a_n, b_m	3	4
a_1	0.158348214538045596	0.056014879123090215
a_2	0.046064514990930811	0.035111795789180087
a_3	0.004886137910884147	0.002183438694367233
a_4	0.000433673330597152	0.000246486152552295
a_5	0.000017343561379589	0.000009222817788667
b_1	0.012514881204710761	–0.061172620876911286
b_2	0.026669340700092963	0.027996854281614683
b_3	0.000328543109454736	–0.000751214829430754
b_4	0.000082091078789006	0.000086068074714292

4. Глобальные аппроксимации. Для несложных оценочных расчетов физикам полезно иметь простые формулы, непрерывно и гладко описывающие $I_k(x)$ во всем диапазоне $-\infty < x < +\infty$. Для успешного применения такие формулы должны описывать главные члены обеих асимптотик функций. Физически это обеспечивает переход фермионов в идеальный газ при высоких температурах, и в полностью вырожденный газ при низких температурах. Напишем несколько таких несложных формул.

Двучленная формула была предложена в [8]. Запишем ее в несколько более удобной форме:

$$I_k(x) \approx \frac{y}{(k+1)} \{[\Gamma(k+2)]^{1/k} + y\}^k. \qquad (6)$$

Эта формула не содержит подгоночных параметров. На рис.2. показаны погрешности двучленной формулы (6) для разных индексов. Видно, что для $k > 0$ эта формула всюду завышает значение функции. Погрешность её значительна. Максимальное завышение

составляет 1.3 раза для $k=1$ и доходит до 2.9 раза для $k=4$. Поэтому формула пригодна только для очень грубых оценок.

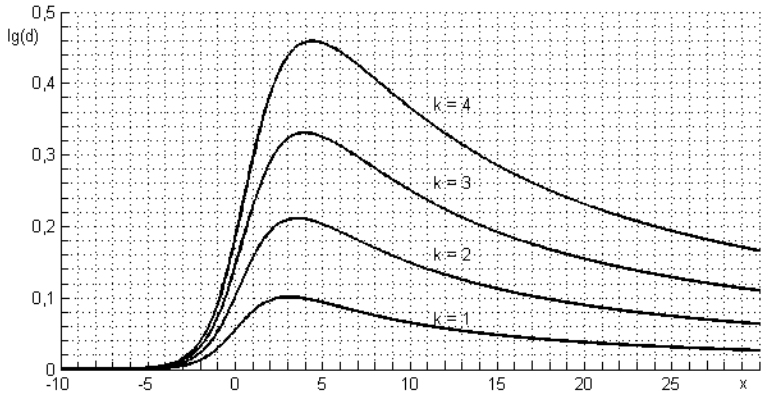

Рис. 2. Профили погрешности формул (6); на кривых указаны k.

Трехчленная формула была предложена в [9-10]. Ее также запишем в несколько удобной форме:

$$I_k(x) \approx \frac{y}{k+1}\{[\Gamma(k+2)]^{3/k}(1+cy)+y^3\}^{k/3}; \qquad (7)$$

коэффициент c подбирается так, чтобы минимизировать относительную погрешность аппроксимации. Формула построена так, что при $y \to 0$ она точно передает главный член левой асимптотики и порядок малости следующего члена разложения (но не точный коэффициент в этом члене). При $y \to +\infty$ она точно передает главный член асимптотики и порядок малости следующего члена. Это улучшает качественное поведение аппроксимации.

Формула (7) оказалась существенно лучше. Наличие подгоночного параметра c позволяет сделать профиль погрешности знакопеременным. Подбирая c из условия чебышевского альтернанса, мы минимизируем погрешность. Соответствующие оптимальные значения c приведены в табл.2; в ней приведены также значения для полуцелых индексов k, взятые из [9, 10]. Видно, что коэффициенты c монотонно убывают с ростом k. В табл.2 приведены также погрешности полученных формул в процентах; они тем больше, чем сильнее отличается k от нуля. Эти погрешности много меньше, чем для формулы (6). Видно, что формулу (7) можно рекомендовать для неплохих оценочных расчетов.

Таблица 2. Коэффициенты и погрешности формулы (7).

k	c	$d_{\max}(\%)$
$-1/2$	1.62	0.7
$1/2$	1.18	0.8
1	1.01	1.6
$3/2$	0.87	2.6
2	0.77	3.2
3	0.60	4.8
4	0.48	6.4

На рис.3 показан профиль погрешности формулы (7) для $k = 2$. Вид профиля подтверждает выполнение чебышевского альтернанса.

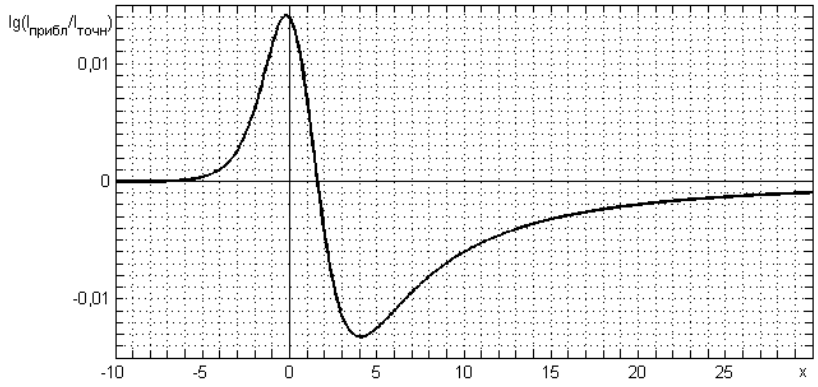

Рис.3. Профиль погрешности формулы (7) для $k = 2$.

Пятичленная формула дает еще более хорошие результаты. Запишем ее так, чтобы члены с первого по третий выглядели как разложение по степеням y при $y \to 0$, а члены с третьего по пятый – как разложение по степеням y^{-2} при $y \to \infty$:

$$I_k(x) \approx \frac{y}{k+1}\{\Gamma(k+2)^{6/k}(1 + c_1 y + c_2 y^2) + c_3 y^4 + y^6\}^{k/6}. \qquad (8)$$

Выбирая коэффициенты из различных соображений, рассмотрим три варианта этой формулы.

а) Выберем коэффициенты c_1 и c_3 так, чтобы правильно передать вторые члены левой и правой асимптотик:

$$c_1 = 3 \cdot (1 - 2^{-k}) / k, \qquad c_3 = \pi^2 (k+1); \qquad (9)$$

коэффициент c_2 оставим в качестве подгоночного. Такая формула обеспечивает описание не только пределов идеального и вырожденного ферми-газа, но и ближайшей поправки при уменьшении идеальности либо снятии вырождения. Особенно важен коэффициент c_3, поскольку он позволяет описать тепловые свойства почти вырожденного газа (например, теплоемкость или проводимость при малых температурах).

Один подгоночный коэффициент c_2 может обеспечить лишь один нуль погрешности. Подбираем c_2 из условия чебышевского альтернанса. Это позволяет минимизировать погрешность. В табл.3 приведены оптимальные значения c_2, а также погрешности полученных формул в процентах. Полученные погрешности 1-2% втрое лучше, чем для формулы (7). Тем самым эта формула предпочтительнее для оценочных расчетов.

б) Передача второго члена асимптотики при высоких температурах обычно не столь важна. Поэтому можно коэффициенты c_1 и c_2 сделать свободными параметрами, а коэффициент c_3 сохранить согласно (9). Это позволяет ввести второй нуль в график

погрешности, а коэффициенты c_1 и c_2 подобрать из условия чебышевского альтернанса. Значения этих коэффициентов и соответствующие им значения максимальных погрешностей приведены в табл.4. Видно, что при этом достигается точность 0.5–1 %. Это вдвое лучше, чем в табл.3.

Таблица 3. Коэффициенты и погрешности формулы (8), вариант а).

k	c_2	$d_{\max}(\%)$
1	1.73	0.9
2	0.98	1.3
3	0.62	1.8
4	0.42	1.9

Таблица 4. Коэффициенты и погрешности формулы (8), вариант б) .

k	c_1	c_2	$d_{\max}(\%)$
1	1.15	1.99	0.45
2	0.93	1.11	0.60
3	0.75	0.69	0.70
4	0.60	0.47	0.80

в) Пусть важна минимальная погрешность, а вторым членом правой асимптотики можно пожертвовать. Тогда все три коэффициента c_1, c_2, c_3 можно использовать как подгоночные и выбирать из условия чебышевского альтернанса. График погрешности при этом будет иметь три нуля. Соответствующие значения коэффициентов и погрешностей приведены в табл.5. Видно, что погрешности еще вдвое уменьшаются по сравнению с табл. 4 и составляют всего 0.2 – 0.5 %. Это превосходная точность для столь простых формул.

Таблица 5. Коэффициенты и погрешности формулы (8), вариант в).

k	c_1	c_2	c_3	$d_{\max}(\%)$
1	1.28	1.78	21.50	0.20
2	0.99	1.02	31.42	0.30
3	0.78	0.65	41.45	0.45
4	0.63	0.44	51.59	0.50

Отметим, что подобранные значения c_3 оказались близкими к теоретическим значениям (9). Это свидетельствует об удачном выборе аппроксимации.

По-видимому, все предложенные формулы будут хорошо работать и для полуцелых индексов, вплоть до $k = -3/2$.

СПИСОК ЛИТЕРАТУРЫ

1. *E.C. Stoner, J. McDougall.* The computation of Fermi-Dirac functions. Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 1938, 237(773): 67_104.
2. *Jr H.C. Thacher, W.J. Cody.* Rational Chebyshev approximations for Fermi-Dirac integrals of orders -1/2, 1/2 and 3/2 // Mathematics of Computation, 1967, p.30–40.
3. *M. Lundstrom, R. Kim.* Notes on Fermi-Dirac integrals. Arxiv preprint arXiv:0811.0116, 2008.
4. *Н.Н. Калиткин.* О вычислении функций Ферми-Дирака // Журнал вычислительной математики и математической физики, 1968, 8(1):173-175;
   N.N. Kalitkin. About computation of functions the Fermi-Dirac // Magazine of computational mathematics and mathematical physics, 1968, 8(1):173-175.

5. *L.D. Cloutman.* Numerical evaluation of the Fermi-Dirac integrals // The Astrophysical Journal Supplement Series, 71:677, 1989.
6. *M. Goano.* Algorithm 745: computation of the complete and incomplete Fermi-Dirac integral. ACM Transactions on Mathematical Software (TOMS), 1995, 21(3):221–232.
7. *A.J. MacLeod.* Algorithm 779: Fermi-Dirac functions of order-1/2, 1/2, 3/2, 5/2. ACM Transactions on Mathematical Software (TOMS), 1998, 24(1):1–12.
8. *Н.Н. Калиткин, С.А. Колганов.* Прецизионные аппроксимации функций Ферми–Дирака целого индекса // Матем. моделирование, 2016, т.28, №3, с.23–32;
   N.N. Kalitkin, S.A. Kolganov. Pretsizionnye approksimatsii funktsii Fermi–Diraka tselogo indeksa // Matem. modelirovanie, 2016, t.28, №3, s.23–32
9. *Н.Н. Калиткин, И.В. Ритус.* Гладкие аппроксимации функций Ферми–Дирака. – М.: ИПМ им. М. В. Келдыша, 1981, препринт №72, 9с.;
   N.N. Kalitkin, I.V. Ritus. Smooth approximations of functions the Fermi-Dirac // Magazine of computational mathematics and mathematical physics. – M.: IPM of M.V. Keldysh, 1981, preprint №72, 9c.
10. *Н.Н. Калиткин, И.В. Ритус.* Гладкая аппроксимация функций Ферми–Дирака // Журнал вычислительной математики и математической физики, 1986, т.26, №3, с.461-464.
    N.N. Kalitkin, I.V. Ritus. Smooth approximation of Fermi–Dirac functions // USSR Computational Mathematics and Mathematical Physics, 1986, 26:2, 87–89.

Поступила в редакцию 04.05.2016