多倍長計算手法

平成２５年度第3四半期

目次

1. はじめに
2. 32倍精度演算を基本とする多倍長演算
   2.1 除算
   2.2 平方根
   2.3 演算量
   2.4 ヒルベルト行列
3.量子モンテカルロ法による物性スペクトル計算
   3.1 32,64,96,128倍精度演算の計算結果
   3.2 32,64,96,128倍精度演算性能測定結果

1．はじめに

平成２５年第2四半期では
量子モンテカルロ法による物性スペクトル計算
で正しく計算できる範囲をieee754-2008データ
ン形式の32倍精度演算まで実施しました。
このデータ形式では表現可能な数値範囲の制限
から128倍精度演算まで可能であるので，
これを作成して正しく計算出来るパラメータ範囲
の限界を求めた。結果の検証は静的かつ規則的
なパスに対してG(β)=-1でチェックしている。

128倍精度演算の作成に当たっては,32倍精度
変数をつなげる方式を採用しました。
倍精度変数を4つつなげて8倍精度変数を作成
した場合,消失する有効ビット数は
(241-212)/241=0.12(12%)と大きいが
32倍精度変数を4つつなげた場合には
(4081-4036)/4081=0.011(1.1%)と
消失する有効ビット数の比率が少ない事と
4倍精度変数を複数個つなげた8倍精度演算
12倍精度演算,16倍精度演算がすでにあるので
作業効率の良さを考慮した事によります。

2.32倍精度演算を基本とする多倍長精度演算

一般に,ある精度を持つ変数をつなげた形式で多倍長演算を作成した場合に以下の問題が発生します。
演算精度の増加分だけ,変数の値の範囲が制限されます。
例えば,倍精度変数を4つつなげて,8倍精度変数と演算を定めると,倍精度変数では$10^{-308} \sim 10^{308}$の範囲で精度が保証されていますが,8倍精度変数では,$10^{-260} \sim 10^{260}$の範囲でしか精度が保証されません。倍精度変数を5つつなげて,10倍精度変数と演算を定めると,精度が保証される範囲は$10^{-244} \sim 10^{244}$となります。
これが*ieee*754形式で多倍長演算を定めたときに量子モンテカルロ法によるスペクトル計算で発生した問題です。
32倍精度変数を複数つなげて多倍長演算を定めるとこれ以外にも問題がでてきます。
*ieee*754の倍精度変数は指数部のビット数と仮数部のビット数が非常に良いバランスをとっています。*ieee*754 – 2008形式ではバランスを考えれば仮数部のビット数は$(52+1)*16$ビット$=848$ビットで27倍精度変数となり,32倍精度変数は仮数部が160ビット多いという事になります。
また,32倍精度変数を複数つなげると,精度が保証される範囲が1つ毎に10^{303}と非常に大きな制限が加わります。
ただし,32倍精度変数を4つつなげて128倍精度演算を定める場合と整数演算方式で128倍精度演算を作成する作業量は2 ~ 3桁の差があります。
このため,32倍精度変数を複数つなげる影響をみるために,除算,平方根やヒルベルト行列を使用した連立一次方程式の解に関して調査しました。

2.1 除算

逆数がアンダーフローが発生しない引数範囲

64**倍精度** 10^{4627} **未満**

96**倍精度** 10^{4323} **未満**

128**倍精度** 10^{4019} **未満**

で除算(逆数計算)途中で32**倍精度演算でアンダーフローの値になると**0**としています。**

整数演算方式　　+ *DD* 形式の除算の精度

基本となる演算精度は32倍精度

引数は10^{n}とし逆数10^{-n}を求める。

逆数計算　精度はm倍精度で表している。
32倍精度演算はn=4931まで32倍精度分の精度を持つ。

n	128倍精度	96倍精度	64倍精度
4931	0	0	0
4625	64	64	64
4626	64	64	64
4627	64	64	64
4628	32	32	32
4629	32	32	32
4321	96	96	64
4322	96	96	64
4323	96	96	64
4324	64	64	64
4325	64	64	64
4017	128	96	64
4018	128	96	64
4019	128	96	64
4020	96	96	64
4021	96	96	64

2.2 平方根

平方根の計算方式には以下の2つがある。

(1)逆数平方根計算方式

$$x_0 = \frac{1}{\sqrt{a}}\text{(基本演算精度)}$$

$$x_{n+1} = x_n + \frac{x_n(1 - ax_n^{\,2})}{2}$$

$$\sqrt{a} = a \times x_{last}$$

$x_n^{\,2} \fallingdotseq \frac{1}{a}$となるので，$a$の大きさに制限がつく。

$ax_n^{\,2} = (ax_n)x_n$として計算する必要がある。

(2)除算方式

$$x_0 = \sqrt{a}\text{(基本演算精度)}$$

$$x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n})$$

$$\sqrt{a} = x_{last}$$

この場合，aの大きさには新たな制限は加わらない。

整数演算方式　＋DD形式の平方根の精度

基本となる演算精度は32倍精度

引数は　$10^{\,n}$とし，平方根　$10^{\,\frac{n}{2}}$を求める。

平方根計算はすべて演算精度分の精度をもつ。

2.3　演算量

1.演算量は32倍精度演算を単位として
　計算している。
2.平方根計算は逆数平方根算出方式
　(旧方式) から除算を使用する方式 (新方式)
　に変更しています。
3.絶対値の大きい順に並べる処理では、IF
　文を用いない方式を取っています。

演算での注意事項

1.乗算で32倍精度変数を2つに分ける定数は
　$r = 2^{505} + 1$
2.乗算a×bではアンダーフローを防ぐため,
　a,bの最初の4倍精度変数の値が0と見なされれば
　$(a(8,1) = 0, b(8,1) = 0)$　$a \times b = 0$としている。
3.0.5をかける場合,$0.5 \times a(8,1)$,$0.5 \times a(8,2)$,
　$0.5 \times a(8,3)$,$0.5 \times a(8,4)$
　と最初の4倍精度のみに0.5をかける。

多倍長精度演算の演算量

32倍精度演算を単位とする。

演算	精度	add,sub32	mult32	div32	sqrt32
加減算	64	14	0	0	0
	96	45	0	0	0
	128	94	0	0	0
乗算	64	18	9	0	0
	96	72	24	0	0
	128	174	46	0	0
除算	64	38	9	2	0
	96	258	48	3	0
	128	858	138	4	0
平方根	64	120	18	4	1
(新)	96	651	96	6	1
	128	2950	414	12	1
平方根	64	182	63	1	1
(旧)	96	684	168	1	1
	128	2304	460	1	1

平方根計算では演算量で見ると新方式は乗算の実行回数が大きく削減されるので、実行時間では旧方式と同じか若干短くなります。

2.4 ヒルベルト行列

ヒルベルト行列の条件数

ヒルベルト行列$H_{ij} = \frac{1}{i+j-1}\ (1 \le i, j \le n)$

逆行列$H_{ji}^{-1} = (-1)^{i+j} \frac{(n+i-1)!(n+j-1)!}{(i+j-1)[(i-1)!(j-1)!]^2 (n-i)!(n-j)!}$

H_{ij}のノルムは$1+\frac{1}{2}+.....+\frac{1}{n}$　　nが大とすると,

$\ln(n)+\gamma$　　$\gamma = 0.57721566490153286060$　(オイラー数)

逆行列H_{ji}^{-1}のノルムを$|H_{ji}^{-1}|$の最大値とする。

またnは大として、スターリングの公式$n! \sim \sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}}$を使用する。

逆行列の絶対値Pは変形して,

$$P = \frac{(n+i)!(n+j)!i^2 j^2}{(n+i)(n+j)(i+j-1)[i!\,j!]^2 (n-i)!(n-j)!}$$

$i = an, j = bn$とすると,

$$P = \frac{a^2 b^2 n^2 ((1+a)n)!((1+b)n)!}{((a+b)n-1)(1+a)(1+b)[(an)!(bn)!]^2 ((1-a)n)!((1-b)n)!}$$

$$= \frac{ab\sqrt{\frac{1+a}{1-a}}\sqrt{\frac{1+b}{1-b}}}{(1+a)(1+b)((a+b)n-1)4\pi^2} (\frac{1+a}{1-a})^{(1-a)n} (\frac{1+b}{1-b})^{(1-b)n} (\frac{1+a}{a})^{2an} (\frac{1+b}{b})^{2bn}$$

Pはaとbに対して対称より,$a = b$とすると、

$$P = \frac{a^2}{(1-a^2)(2an-1)4\pi^2} (\frac{1+a}{1-a})^{2(1-a)n} (\frac{1+a}{a})^{4an}$$

$f(a) = \ln[(\frac{1+a}{1-a})^{(1-a)} (\frac{1+a}{a})^{2a}]$とすると$f'(a) = \ln(\frac{1-a^2}{a^2})$

$0.5 \le a \le 0.75$では$f(a)$は$a = \sqrt{0.5} = 0.7071$で最大となる.

$i = an, j = bn$が整数となる事から,$a = 0.7$とする.

すると$P_{\max} = \frac{0.0243}{1.4n-1} (\frac{17}{7})^{2.8n} (\frac{17}{3})^{0.6n} = \frac{0.0243}{1.4n-1} 10^{1.530979n}$

より詳細にすると以下の様になる。

$$a = 0.7 \qquad P = \frac{0.02433695097}{1.4n-1}\left(\frac{17}{3}\right)^{0.6n}\left(\frac{17}{7}\right)^{2.8n}$$

$$= 10^{1.530979068n-1.613733833-\log_{10}(1.4n-1)} \qquad (1)$$

$$a = \sqrt{0.5} = \frac{\sqrt{2}}{2} \qquad P = \frac{1}{(\sqrt{2}n-1)4\pi^2}\left(\frac{2+\sqrt{2}}{2-\sqrt{2}}\right)^{(2-\sqrt{2})n}(\sqrt{2}+1)^{2\sqrt{2}n}$$

$$= 10^{1.531102741n-1.596359739-\log_{10}(\sqrt{2}n-1)} \qquad (2)$$

$$(2)/(1) = 10^{0.000123673n+0.017374096+\log_{10}(1.4n-1)-\log_{10}(\sqrt{2}n-1)}$$

10^mのm

	(1)	(2)	(2)/(1)	条件数比率
$n = 200$	302.1364756	302.1741817	0.0377	1.091
$n = 400$	608.0304816	608.0929301	0.0624	1.155

ヒルベルト行列Hを使用した連立一次方程式Hx=bでは,bの要素に誤差があると条件数が大きいので結果に大きな影響がでます。

その影響をみるため$x = (1,1,\ldots,1)^T$となる様に,bを同じ精度の演算で計算して,$Hx = b$からxをもとめ,得たx^Tと$(1,1,\ldots,1)^T$の値を比較しました。(bには演算精度におおじた誤差が入る.)

これまでのヒルベルト行列実測結果

連立一次方程式は部分軸選択付きのLU分解法を使用しています。

このためbに全く誤差のない値(たとえばbの要素すべて1など)を入れると,得られた解xを使用してHxとbを比較するとほとんど誤差は出ません.

ヒルベルト行列最大誤差一覧表

次元数	倍精度	4倍精度	8倍精度
20	51.616	1.44E-07	2.13E-47
40	182.107	62.034	8.72E-16
60	1971.569	62.804	251.132
80	642.074	300.734	252.653
100	201.355	378.428	242.672

次元数	16倍精度	32倍精度
20	1.47E-99	1.44E-276
40	4.09E-86	4.30E-243
60	2.66E-60	1.36E-210
80	3.65E-31	1.08E-178
100	2.58E+00	1.41E-150

詳細なテスト結果

ヒルベルト行列の精度

結果のビット数が113以上は誤差なしとして0としている。
有効ビット数 32倍精度　1009,64倍精度 2018

演算精度	32倍精度				64倍精度		
	結果	条件数	条件数		結果	条件数	条件数
次元数	(ビット数)	(ビット数)	(定義式)	次元数	(ビット数)	(ビット数)	(定義式)
170	0	853	857	370	0	1870	1874
172	0	864	867	372	0	1880	1884
174	0	874	878	374	0	1890	1894
176	0	884	888	376	0	1900	1905
178	0	894	898	378	106	1910	1915
180	105	904	908	380	97	1920	1925
182	101	914	918	382	86	1931	1935
184	85	924	928	384	76	1941	1945
186	76	935	939	386	76	1951	1955
188	74	945	949	388	61	1961	1966
190	54	955	959	390	47	1971	1976
192	46	965	969	392	39	1981	1986
194	37	975	979	394	25	1992	1996
196	27	985	989	396	16	2002	2006
198	15	996	1000	398	5	2012	2016
200	6	1006	1010	400	−6	2022	2027
202	−8	1016	1020				
204	−5	1026	1030				
206	−7	1036	1040				
208	−6	1046	1050				
210	−7	1057	1061				

(注)マイナスのビット数は最初の1桁目より結果が異なる事を示す。

ヒルベルト行列Aを使用して,

$Ax = b$で$x = (1,1,\ldots,1)^T$となる様に

$b_i = \sum_{j=1}^{n} a_{ij}$として,$b$を定めますが条件数は

このbを作成する際の誤差に大きく影響します。

例えばAを32倍精度で求め,その他の演算を

全て64,96,128倍精度で行うと条件数の影響は

全くでなくなり、解$x = (1,1,\ldots,1)^T$が得られます。

簡単な例

$$A = \begin{pmatrix} 1 & 1-\varepsilon \\ 1-\varepsilon & 1-\varepsilon \end{pmatrix} \qquad A^{-1} = \begin{pmatrix} \frac{1}{\varepsilon} & -\frac{1}{\varepsilon} \\ -\frac{1}{\varepsilon} & \frac{1}{\varepsilon(1-\varepsilon)} \end{pmatrix}$$

条件数$= \dfrac{(2-\varepsilon)^2}{\varepsilon(1-\varepsilon)} \fallingdotseq \dfrac{4}{\varepsilon}$

解　$x = (1,1)^T$　となる様にbを定める場合εの値で

$b_0 = (2,2-2\varepsilon)^T, b_1 = (2-\varepsilon,2-2\varepsilon)$となる場合があります。

$Ax = b_0$の解は$x = (2,0)^T$, $Ax = b_1$の解は$x = (1,1)^T$

3.量子モンテカルロ法による物性スペクトル計算

結果のチェックとして,$L+1$個の計算結果の
絶対値を$E(i)\quad(i=0,1,.....,L)$としたとき、
$E(0)=1,E(L)=1,E(\frac{L}{2})$で絶対値が最小値をとる
という条件とした。
$n=20$とすると,

$$\frac{1}{E(\frac{L}{2})}=e^{P}(P=\sqrt{\frac{\beta U}{L}}\times\frac{L+1}{2})$$ となる。

$\beta<L,\beta U>L$の場合,
条件数は,$e^{P}(P=(\frac{\beta U}{L})^{\frac{\beta}{L}})$ となり注意を要する。
また$\beta>L$となると誤差が$O((\frac{\beta}{L})^{2})$から正しく計算
する条件はさらに厳しくなる。正しく計算されると

$$\frac{1}{E(\frac{L}{2})}=e^{P}(P=\sqrt{\frac{\beta U}{L}}\times\frac{L+1}{2})$$ となる。

3.1　32,64,96，128倍精度演算の結果の精度

N=20,L=448 結果一覧

32倍精度演算

β	u	実測値	理論値
250	10	3.12E−231	4.79E−231

64倍精度演算

β	u	実測値	理論値
500	10	0.202E−325	0.190E−325
600	10	0.193E−356	0.155E−356
700	5	2.45E−273	3.03E−273
800	5	4.13E−292	4.63E−292
900	3	3.01E−240	4.41E−240
1000	2	5.69E−207	9.90E−207

96倍精度演算

β	u	実測値	理論値
700	8	0.228E−344	0.194E−344
800	7	0.228E−344	0.194E−344
900	4	3.41E−277	4.13E−277
1000	3	3.63E−253	4.98E−253

128倍精度演算

β	u	実測値	理論値
700	9	0.312E−365	0.239E−365
800	7	0.227E−344	0.194E−344
800	8	0.407E−368	0.308E−368
900	4	3.41E−277	4.13E−277
900	5	0.961E−309	0.985E−309
1000	3	3.63E−253	4.98E−253
1000	4	4.13E−292	4.63E−292

実測値と理論値は良く一致している。！

$N = 30, L = 448$　での計算式

$$\frac{1}{\text{絶対値最小値}} = e^{P}\left(P = \left(\sqrt{xU} + \frac{1}{m}(\ln(x))^{m}\right) \times \frac{L+1}{2}\right)$$

$$x = \frac{\beta}{L} > 1 \qquad m = \frac{\beta - 200}{200}$$

N=30,L=448 の実行結果

64倍精度演算

β	u	実測値	理論値
500	10	0.630E-328	0.818E-328
600	10	0.438E-360	0.107E-360
700	5	2.39E-278	1.96E-278
800	5	1.21E-298	2.14E-298
900	3	1.42E-248	5.57E-248
1000	2	2.40E-217	7.29E-217

実測値と理論値は良く一致している。！

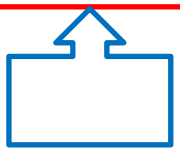
32倍精度と64倍精度演算の結果比較

$N=100, L=448$で正しく計算出来る

β, U

32倍精度演算	$\beta=250, U=10$
	$\beta=300, U=8$
64倍精度演算	$\beta=500, U=10$
	$\beta=600, U=10$

64倍精度と96倍精度,128倍精度演算結果で差がでるものは意外に少ない。



32倍精度変数を複数つなげる事による引数範囲の縮小化による。

この後差の出たケースの結果をしめしました。

$$N = 12, \beta = 1000, U = 10, L = 160$$

0.9750000000D+03 0.1185494857+131
0.9812500000D+03 0.9203004545+139
0.9875000000D+03 0.1006367236+149
0.9937500000D+03 0.1112616274+157
0.1000000000D+04 0.1036389481+166

64bai elapse= 118.255023000000 sec

0.9750000000D+03 -0.3077877772D-14
0.9812500000D+03 -0.8349313422D-11
0.9875000000D+03 -0.2264972890D-07
0.9937500000D+03 -0.6191951556D-04
0.1000000000D+04 -0.1000000000D+01

96bai elapse= 334.062215000000 sec

0.9750000000D+03 -0.3077877772D-14
0.9812500000D+03 -0.8349313422D-11
0.9875000000D+03 -0.2264972890D-07
0.9937500000D+03 -0.6191951556D-04
0.1000000000D+04 -0.1000000000D+01

128bai elapse= 652.552798000000 sec

$$N = 12, \beta = 1200, U = 10, L = 100$$

0.1152000000D+04 -0.1556030131D-19
0.1164000000D+04 -0.8901756069D-15
0.1176000000D+04 -0.5092527422D-10
0.1188000000D+04 -0.2913410882D-05
0.1200000000D+04 -0.1000000000D+01

96bai elapse= 194.011505000000 sec

0.1152000000D+04 -0.1556030131D-19
0.1164000000D+04 -0.8901756069D-15
0.1176000000D+04 -0.5092527422D-10
0.1188000000D+04 -0.2913410882D-05
0.1200000000D+04 -0.1000000000D+01

128bai elapse= 442.914668000000 sec

$$N = 8, \beta = 1500, U = 10, L = 96$$

0.1437500000D+04 0.2067081954+228
0.1453125000D+04 0.8661847662+246
0.1468750000D+04 0.3114430095+265
0.1484375000D+04 0.2193652260+284
0.1500000000D+04 0.5614536671+302

96bai elapse= 65.1950880000000 sec

0.1437500000D+04 -0.4821874620D-22
0.1453125000D+04 -0.1293888751D-16
0.1468750000D+04 -0.3471985966D-11
0.1484375000D+04 -0.9316632940D-06
0.1500000000D+04 -0.1005946110D+01

128bai elapse= 121.441539000000 sec

3.2 32,64,96,128倍精度演算性能測定結果

測定条件

N=20,L=448
X5570　-O2 -fp-model strict で実行

32倍精度	600秒	1
64倍精度	6000秒	10
96倍精度	17052秒	28.4
128倍精度	34407秒	57.3

(注) 96倍精度はもとになった6倍精度の減算に無駄な演算があったので修正しています。
6倍精度は,x5570,e5430,SR16000/M1,BG/Q,T2Kで性能測定をして、性能が向上したのを確認しています。