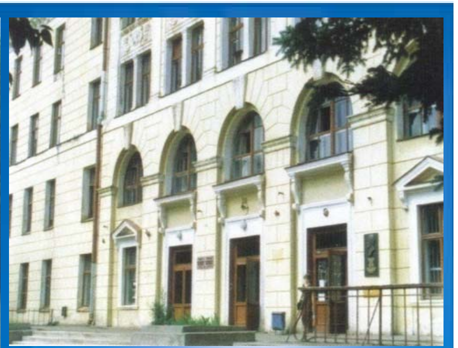


Нижегородский государственный университет
им. Н.И. Лобачевского
Национальный исследовательский университет
Институт информационных
технологий математики и механики

Решение задач глобальной оптимизации на графических ускорителях

К.А. Баркалов, И.Г. Лебедев

Russian Supercomputing Days

МОСКВА,
26-27 сентября 2016 г.

Решение задач глобальной оптимизации на графических ускорителях

Нижегородский государственный университет им. Н.И. Лобачевского

Институт информационных технологий, математики и механики

К.А. Баркалов,
И.Г. Лебедев

Содержание

Постановка задачи

$$\varphi(y^*) = \min\{\varphi(y) : y \in D\},$$

$$D = \{y \in R^N : a_i \le y_i \le b_i, 1 \le i \le N\}.$$

Предположение: выполнено условие Липшица

$$|\varphi(y_1) - \varphi(y_2)| \le L\|y_1 - y_2\|,\ \ y_1, y_2 \in D,\ \ 0 < L < \infty$$

Предположение липшицевости типично для многих других подходов (Ю.Г. Евтушенко, J. Pinter, D. Jones, системы глобальной оптимизации LGO, DIRECT, IOSO)

Трудоемкость задачи

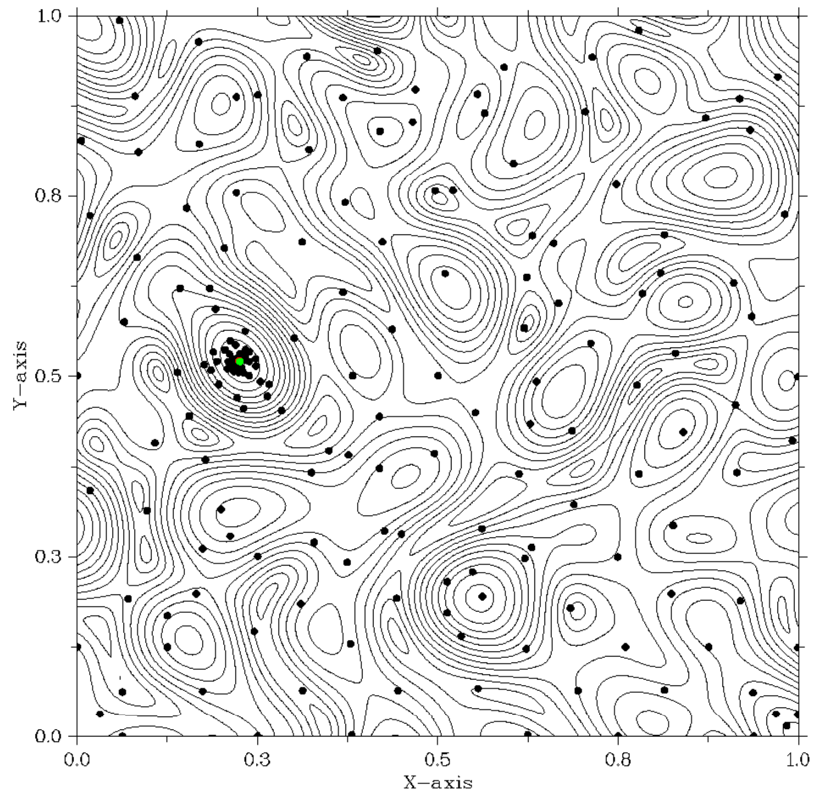
- Экспоненциальный рост затрат при увеличении размерности задачи.

Поиск решения

- Построение неравномерных адаптивных покрытий области поиска
- Методы ориентированы на построение существенно более плотной сетки только в окрестности глобально-оптимального решения задачи, чем вне этой окрестности.

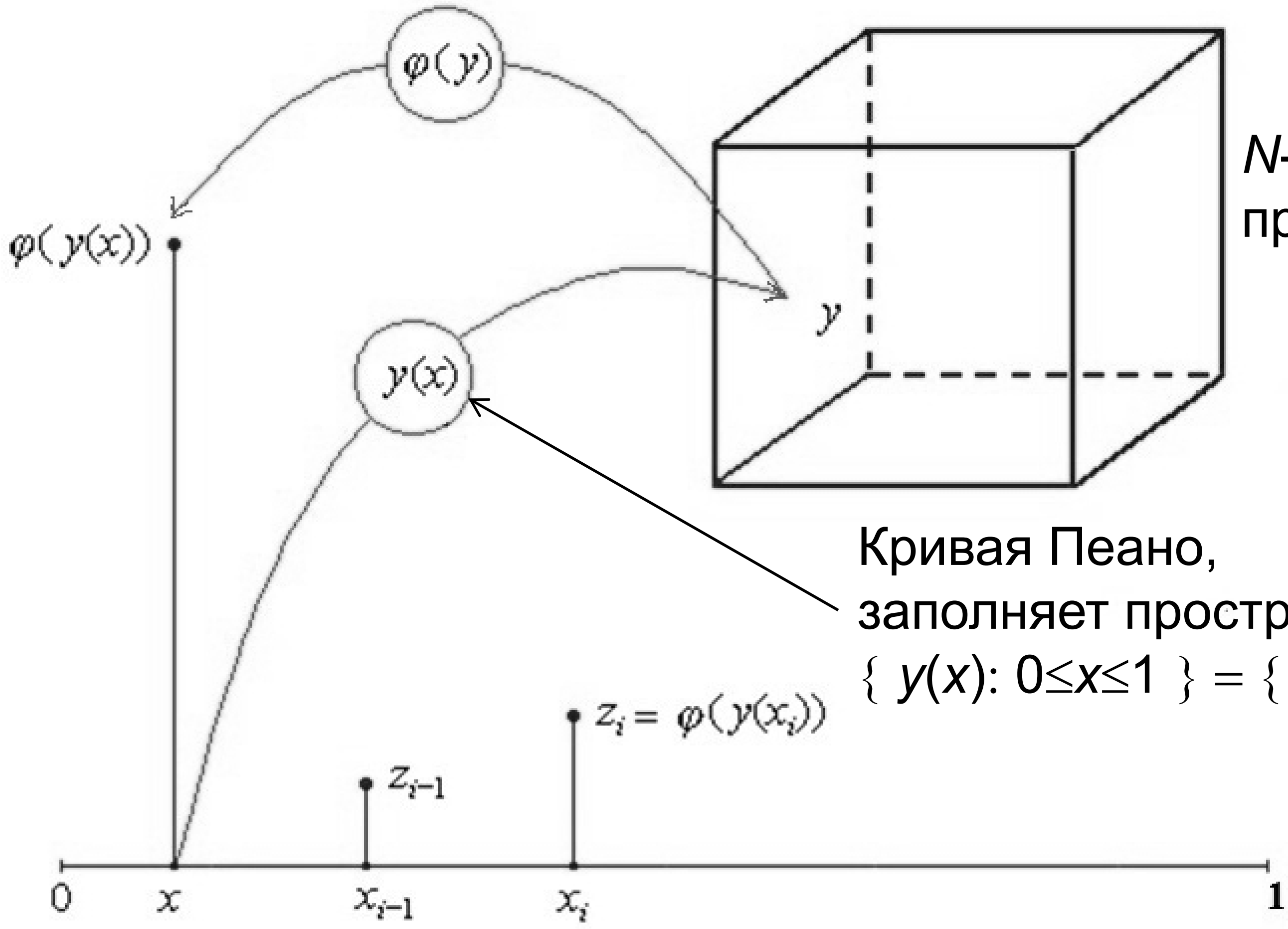
$$y^{k+1} = G_k\left(y^1, \ldots, y^k; Z^1, \ldots, Z^k\right)$$

- При точности $\varepsilon = 10^{-2}$ 128 испытаний



Информационно-статистический подход (Стронгин, Сергеев (2000))

Редукция размерности



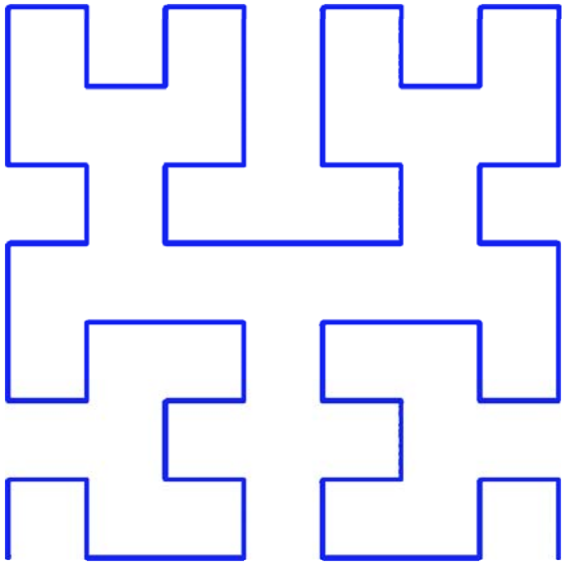
N-мерное пространство

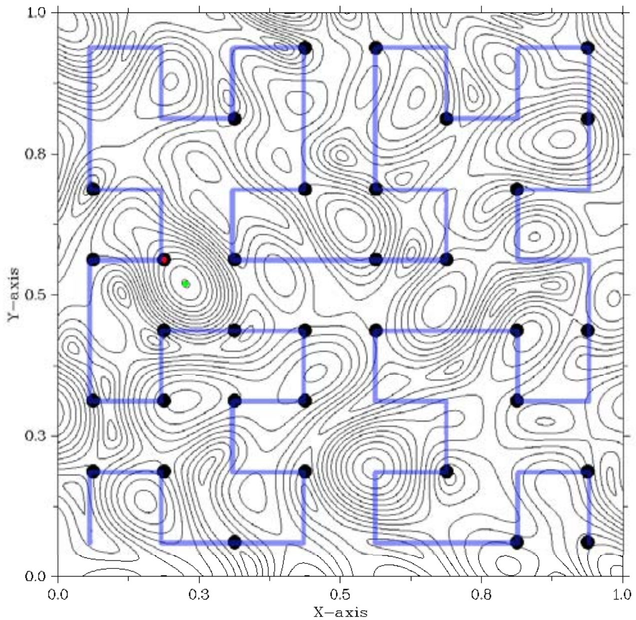
Кривая Пеано, заполняет пространство
$\{ y(x): 0 \le x \le 1 \} = \{ y \in R^N: y \in D \}$

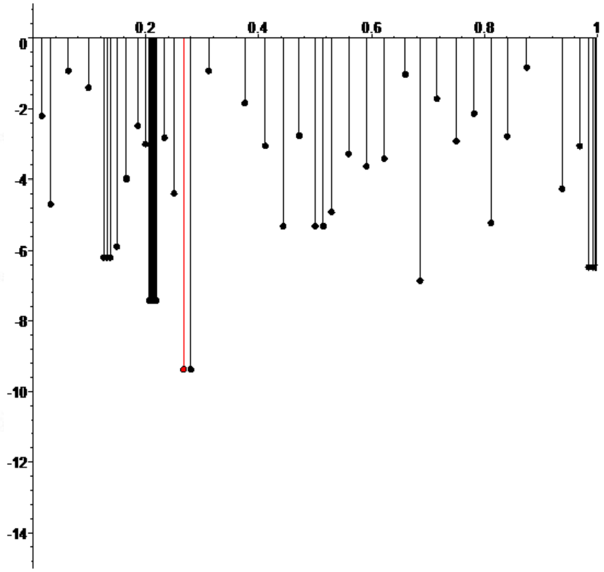
1-мерный интервал

$$\min \{ \varphi(y): y \in D \} = \min \{ \varphi(y(x)): x \in [0,1] \}$$

Редукция размерности







Численный метод построения кривой Пеано с заданной точностью рассмотрены Сергеевым, Стронгиным (2013)

Условие Липшица трансформируется в условие Гельдера: $\left|\varphi(y(x_1)) - \varphi(y(x_2))\right| \le 4L\sqrt{N}\left(\left|x_1 - x_2\right|\right)^{1/N}$

где x_1, $x_2 \in [0,1]$

Параллельный алгоритм глобального поиска

$0=x_0 \quad x_1 \quad x_2 \quad \ldots \quad x_{i-1} \quad x_i \quad \ldots \quad x_k=1$

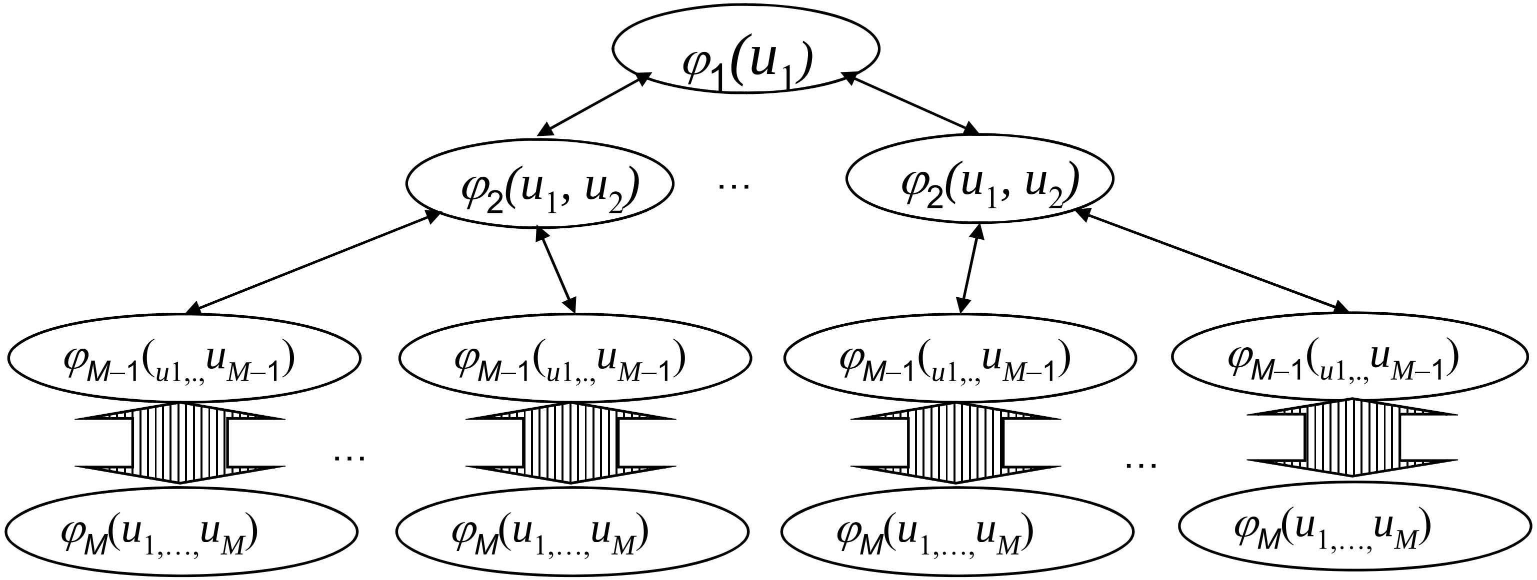
1. $0 = x_0 < \ldots < x_i < \ldots < x_k = 1$, $z_i = \varphi(x_i)$, $\mu = \max\left\{ \frac{|z_i - z_{i-1}|}{x_i - x_{i-1}}, \quad i = 1,\ldots,k \right\}$
2. Для каждого (x_{i-1}, x_i), $1 \le i \le k$, Δ_i – длина интервала,

   $r > 1$, – параметр метода. $R(i) = \Delta_i + \frac{(z_i - z_{i-1})^2}{r^2 \mu^2 \Delta_i} - 2\frac{(z_i + z_{i-1})}{r\,\mu}$,
3. Сортируем интервалы по убыванию характеристик, берем p интервалов $R(t_1) \ge R(t_2) \ge \ldots \ge R(t_p)$
4. **Проводим p испытаний параллельно**

   (x_{t_1-1}, x_{t_1}), (x_{t_2-1}, x_{t_2}),… (x_{t_p-1}, x_{t_p}) $\quad x^{k+1} = \frac{x_{t_i} + x_{t_i-1}}{2} - \frac{z_{t_i} - z_{t_i-1}}{2r\,\mu}$
5. Критерий остановки: $x_{t_i} - x_{t_i-1} \le \varepsilon$, $1 \le i \le p$

Сергеев, Гришагин, Стронгин (1997).

Блочная рекурсивная схема редукции размерности



$$u_1=(y_1,\dots,y_{N_1}),u_2=(y_{N_1+1},\dots,y_{N_2}),\dots u_M=(y_{N_{M-1}+1},\dots,y_N)$$

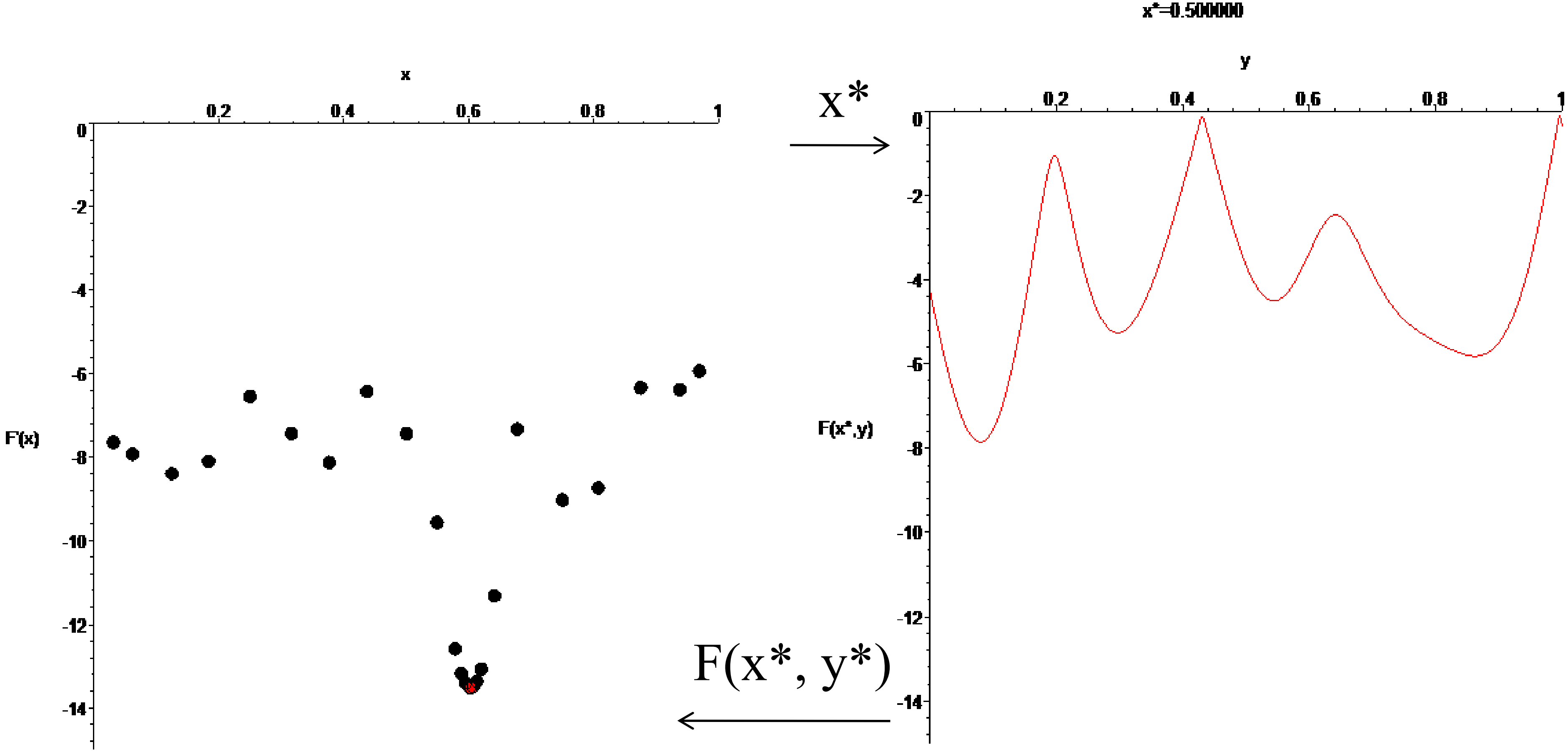
$$\varphi_i(u_1,\dots,u_i)=\min_{u_{i+1}\in D_{i+1}}\varphi_{i+1}(u_1,\dots,u_i,u_{i+1})$$

$$\varphi_{M-1}(u_1,\dots,u_{M-1})=\min_{u_M\in D_M}\varphi_M(u_1,\dots,u_M)=\min_{y_{N_{M-1}+1},\dots,y_N\in D_M}\varphi(y_1,\dots,y_N)$$

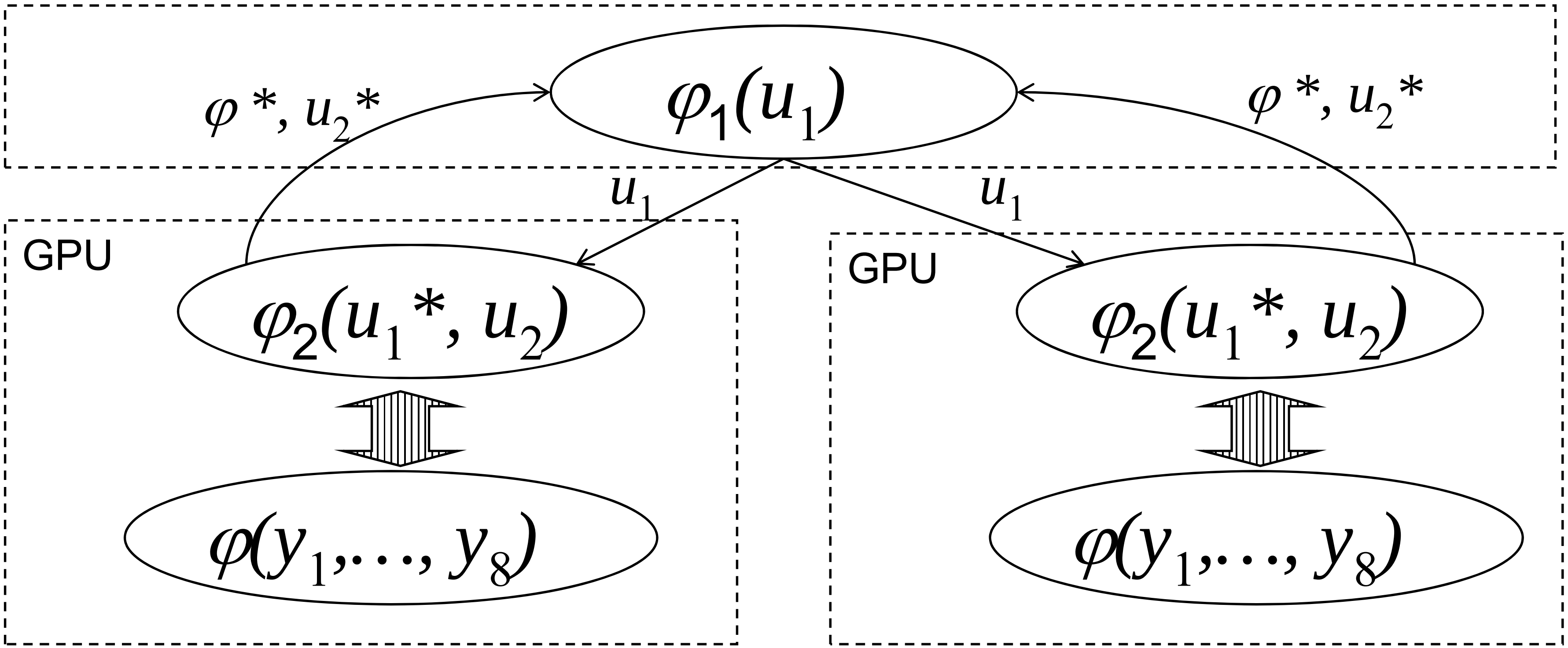
$$\min_{y\in D}\varphi(y)=\min_{u_1\in D_1}\left[\min_{u_2\in D_2}\left[\dots\left[\min_{u_M\in D_M}\varphi(y)\right]\right]\right]$$

Блочная рекурсивная схема редукции размерности

$$F'(x) = \min_{a_2 \le y \le b_2} F(x,y); \qquad \min_{x,y \in D} F(x,y) = \min_{a_1 \le x \le b_1} \left[F'(x) \right]$$



Организация параллельных вычислений

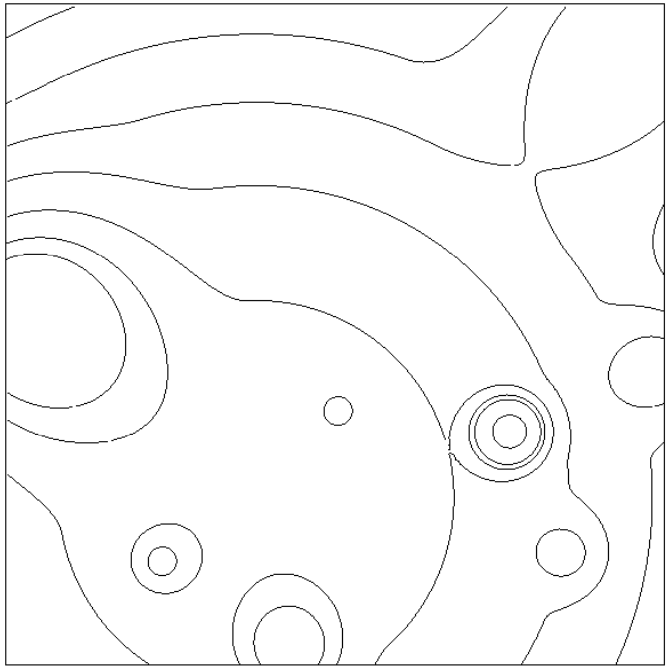


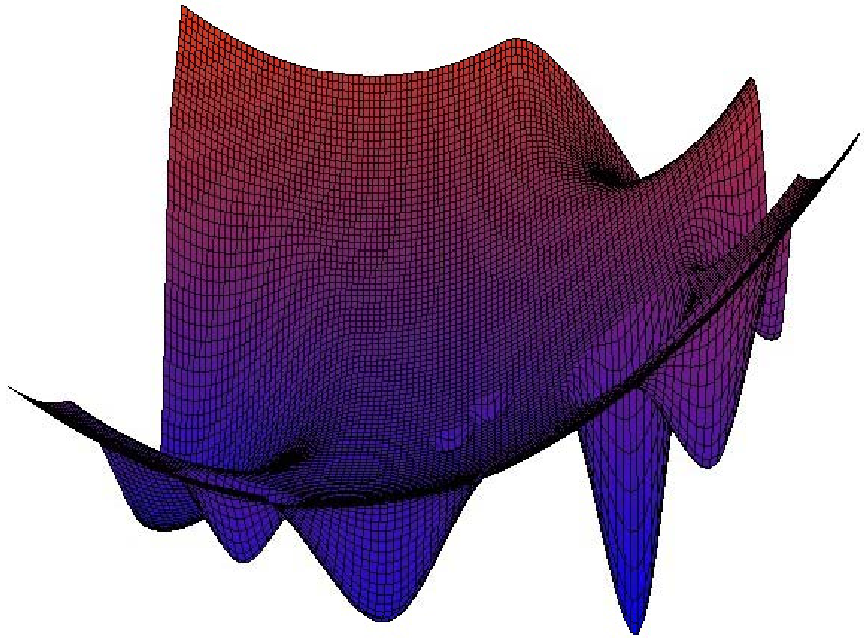
$$u_1 = (y_1, \dots, y_4), u_2 = (y_5, \dots, y_8)$$

$$\min_{y \in D} \varphi(y) = \min_{y_1, \dots, y_4} \left[\min_{y_5, \dots, y_8} \varphi(y) \right]$$

GKLS generator

GKLS генератор позволяет получать задачи многоэкстремальной оптимизации с заранее известными свойствами: количеством локальных минимумов, размерами их областей притяжения, точкой глобального минимума, значением функции в ней и т.п.





GKLS generator

- Вычислительные эксперименты проводились на кластере ННГУ им. Н.И. Лобачевского.
- Узел кластера располагает двумя 4-х ядерных процессора Intel Xeon L5630 2.13 GHz, 24 Gb RAM, две видео карты NVIDIA Tesla X2070.
- Каждая видео карта располагает 6 Gb памяти и 448 CUDA ядрами.
- Ограничение на время решение одной задачи – 3 минуты.

Сравнение с другими методами

Результаты сравнения трех последовательных алгоритмов – DIRECT [1], DIRECT*l* [2] и алгоритм глобального поиска (АГП)

N	Problem class	DIRECT	DIRECT*l*	АГП
4	*Simple*	>47282(4)	18983	11953
	Hard	>95708(7)	68754	25263
5	*Simple*	>16057 (1)	16758	15920
	Hard	>217215 (16)	>269064 (4)	>148342 (4)

Критерий остановки: $\|y^k - y^*\| \lesssim \delta \quad \delta = \|b-a\| \sqrt[N]{\varepsilon}$

Результаты работы первых двух алгоритмов приводятся по работе Sergeyev, Ya.D. Global search based on efficient diagonal partitions and a set of Lipschitz constants.

Решение на CPU

.

Время решения последовательного запуска (сек.)

p	*N*=2		*N*=3		*N*=4		*N*=5	
	Simple	*Hard*	*Simple*	*Hard*	*Simple*	*Hard*	*Simple*	*Hard*
1	0,02	0,05	0,03	0,08	0,19	0,57	0,24	3,8

Ускорение многопоточного запуска

p	*N*=2		*N*=3		*N*=4		*N*=5	
	Simple	*Hard*	*Simple*	*Hard*	*Simple*	*Hard*	*Simple*	*Hard*
2	7,4	6,5	1,1	1,01	1,35	1,32	0,68	1,2
4	1,16	0,63	1,33	1,12	1,53	1,37	0,99	1,29
8	6,87	0,9	1,36	1,33	1,49	1,61	0,62	2,22

Решение на GPU

Параллельный алгоритм глобального поиска для вычислений целевой функции используется CUDA. На узле два ускорителя по *p* нитей на каждом. Размер CUDA блока равен 32. Каждая нить вычисляет одну функцию.

Ускорение по времени относительно последовательного запуска

p	*N*=2		*N*=3		*N*=4		*N*=5	
	Simple	*Hard*	*Simple*	*Hard*	*Simple*	*Hard*	*Simple*	*Hard*
128	4,6	4,97	1,06	1,19	1,12	1,38	0,66	1,82
256	3,37	6,11	1,02	1,34	1,6	1,51	1,02	2,35
512	2,27	4,36	0,76	1,36	1,59	1,53	0,83	1,29

Решение на GPU

Параллельный алгоритм перебора, вычислений происходят на GPU с использованием CUDA. На узле два ускорителя, число нитей вычисляется исходя из шага сетки. Размер CUDA блока равен 32. Каждая нить вычисляет несколько значений функции исходя из ограничения на число блоков.

Шаг сетки 0.027, K_{max} = 30 000 000

- Ускорение по времени относительно последовательного запуска

N=2			*N*=3			*N*=4	
Simple	*Hard*		*Simple*	*Hard*		*Simple*	*Hard*
15,49	39,31		11,73	31,28		2,60	7,60

Гибридное решение

- Решение задач размерности 6 и 8 простого класса.
- AGP – последовательный алгоритм глобального поиска
- B-AGP – рекурсивной схемы редукции размерности , по одному процессу на уровне, с последовательным вычислением на нижним уровне.
- H-AGP – рекурсивной схемы редукции размерности , по одному процессу на уровне, с перебором реализованным на GPU на нижнем уровне. Шаг сетки 0,01.
- M-AGP – рекурсивной схемы редукции размерности , одни процесс на верхнем уровне и четыре на нижнем, используется четыре узла кластера. На нижнем уровне перебор реализованный на GPU. Шаг сетки 0,01.

Гибридное решение

Среднее время решения задачи большой размерности(сек.)

N	**AGP**	**B-AGP**	**H-AGP**	**M-AGP**
6	53,5(20)	4,4	1,1	0,4
8	72,6(19)	78,3	10,7	3,1

Ускорение по времени для решения задач большой размерности

N	**B-AGP**	**H-AGP**	**M-AGP**
6	12,1	48,4	133,7
8	0,9	6,8	23,4

Литература

- Gergel V.P., Sergeyev Ya.D. Sequential and parallel algorithms for global minimizing functions with Lipschitzian derivatives // Computers and Mathematics with Applications 1999. Vol. 37 No. 4-5. P. 163–179.
- Gergel V.P., Strongin R.G. Parallel computing for globally optimal decision making on cluster systems // Future Generation Computer Systems 2005. Vol. 21 No. 5. P. 673-678.
- Сысоев А.В. Баркалов К.А. Гергель В.П. Лебедев И.Г. Решение задач глобальной оптимизации на гетерогенных кластерных системах // Суперкомпьютерные дни в России: Труды международной конференции (28-29 сентября 2015 г., г. Москва).Москва: Изд-во МГУ. 2015. С. 411-419
- K.Barkalov, V. Gergel. Parallel global optimization on GPU // Journal of Global Optimization vol. 66 (1), 2016, pp. 3-20.

Контакты

- к.ф.-м.н., доц., доцент каф. МОСТ института ИТММ,
  Баркалов Константин Александрович
  konstantin.barkalov@itmm.unn.ru

- аспирант каф. МОСТ института ИТММ
  Лебедев Илья Генадьевич
  ilya.lebedev@itmm.unn.ru

Спасибо за внимание