Оптимизация на основе гауссовских процессов

Иван Коноваленко

Московский физико-технический институт
Институт проблем передачи информации РАН
Научный руководитель:
к.ф.-м.н., доцент, зав. сект. ИППИ РАН
Бурнаев Евгений Владимирович

Москва, 2012

Введение

- В некоторых практических приложениях требуется оптимизировать функцию, расчёт которой является вычислительно сложным.
- Таким образом, возникает задача оптимизации, где важна сходимость к минимуму целевой функции по количеству обращений к ней.
- Подход, основанный на суррогатном моделировании, может ускорить процесс оптимизации.
- Данная работа имеет цель экспериментально сравнить такие методы, предложив их эффективные программные реализации.

Задача оптимизации

- Предположим, что задана функция $f(\mathbf{x}) : \mathbb{X} \to \mathbb{R}$, где $\mathbb{X} \subset \mathbb{R}^n$.
- Пусть $D_i = \left\{\left(\mathbf{x}_k, y_k = f(\mathbf{x}_k)\right)\right\}_{k=1}^{i}, i = 1, \ldots, N$, $D_0 = \emptyset$.
- Пусть также $\pi = \left(\pi_1, \ldots, \pi_N\right)$ – некоторое управление, т.е.

$$\pi_i : D_{i-1} \to \mathbb{X}.$$

- Таким образом наблюдается последовательность точек $\mathbf{x}_i = \pi_i(D_{i-1}), y_i = f(\mathbf{x}_i), i = 1, \ldots, N$.
- Ставится задача нахождения такого управления π, которое доставляет минимум целевой функции:

$$f(\mathbf{x}_N(\pi)) \to \min_{\pi \in \Pi},$$

где Π - некоторый класс управлений

Способ задания управления

- Если задана обучающая выборка $D_{i-1} = \left\{\left(\mathbf{x}_k, f(\mathbf{x}_k)\right)\right\}_{k=1}^{i-1}$, то можно построить аппроксимацию $\hat{f}_{i-1} = \hat{f}(\mathbf{x}|D_{i-1})$.
- Аппроксимация является приближением целевой функции и может быть использована для построения управления.
- Например, управление можно искать в форме максимизации некоторого критерия:

$$\mathbf{x}_i = \pi_i(D_{i-1}, \hat{f}_{i-1}) = \arg\max_{\mathbf{x}\in\mathbb{X}} I(\mathbf{x}|D_{i-1}, \hat{f}_{i-1}),$$

где $I(\mathbf{x}|D_{i-1}, \hat{f}_{i-1})$ - некоторый критерий выбора новой точки на основе аппроксимации $\hat{f}_{i-1}$.

Модельные предположения о зависимости

Пусть зависимость $y(\mathbf{x})$ порождена моделью:

$$y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon(\mathbf{x}),$$

где $f(\mathbf{x})$ — реализация случайного гауссовского поля, а $\varepsilon(\mathbf{x}) \sim \mathcal{N}(0, \sigma_1^2)$. Предположим, что ковариационная функция гауссовского поля $f(\mathbf{x})$:

$$K_0(\mathbf{x}, \mathbf{x}') = \sigma^2 \exp\left(- \sum_{i=1}^{p} \theta_i^2 (x_i - x_i')^2 \right),$$

Тогда ковариационная функция процесса:

$$K(\mathbf{x}, \mathbf{x}') = K_0(\mathbf{x}, \mathbf{x}') + \sigma_1^2 \delta(\mathbf{x}, \mathbf{x}'),$$

где $\delta(\mathbf{x}, \mathbf{x}')$ — символ Кронекера.

Условное распределение

- Условное распределение $y(\mathbf{x})$ имеет вид:

$$Law(y(\mathbf{x})|y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_n)) \sim \mathcal{N}(\hat{f}(\mathbf{x}), \hat{\sigma}^2(\mathbf{x}))$$

- Здесь условное математическое ожидание, используемое для прогноза, равно

$$\hat{f}(\mathbf{x}) = \mathsf{E}(y(\mathbf{x})|y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_n)) = \mathbf{k}(\mathbf{x})\mathrm{K}^{-1}\mathbf{y},$$

  причем $\mathbf{k}(\mathbf{x}) = \{K(\mathbf{x}, \mathbf{x}_i)\}_{i=1}^n$, матрица $\mathrm{K} = \{K(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.

- Условная дисперсия $\hat{\sigma}^2(\mathbf{x})$ равна

$$\hat{\sigma}^2(\mathbf{x}) = K_0(\mathbf{x}, \mathbf{x}) + \sigma_1^2 - \mathbf{k}(\mathbf{x})\mathrm{K}^{-1}\mathbf{k}(\mathbf{x})^T.$$

Суррогатные методы оптимизации (SBO)

Surrogate-Based Optimization (SBO) - метод оптимизации, базирующийся на аппроксимации целевой функции:

1. Дана начальная обучающая выборка D.
2. По выборке D строится аппроксимация целевой функции $\hat{f}(\mathbf{x}|D)$ и апостериорной дисперсии $\hat{\sigma}^2(\mathbf{x}|D)$.
3. Новая точка получается из максимизации критерия $I(\mathbf{x}|D)$:
4. Добавляем новую точку: $D := D \cup (\mathbf{x}_{new}, y(\mathbf{x}_{new}))$.
5. Если лимит точек не исчерпан, то переходим к шагу 2.
6. Возвращаем минимальное значение в выборке D: $(\mathbf{x}_{\min}, y_{\min} = y(\mathbf{x}_{\min}))$.

Далее рассмотрим конкретные виды критериев $I(\mathbf{x}|D)$.

Minimum of approximation

В соответствии с Minimum of approximation новая точка добавляется туда, где значение аппроксимации минимально.

$$\mathbf{x}_{new} = \arg\min_{\mathbf{x}} \hat{f}(\mathbf{x}|D). \tag{1}$$

Преимущества:

- Качество работы критерия сильно зависит от качества аппроксимации целевой функции.
- Критерий очень просто считается, т. к. аппроксиматор на основе гауссовских процессов легко считается.

Недостатки:

- Метод обладает локальным свойством: не обязательно заполняет $\mathbb{X}$ плотно. Таким образом, сходимость с ростом размера выборки не гарантирована.

Expected Improvement (EI)

$$EI(\mathbf{x}) = \mathbb{E}\big[\max(0, f_{\min} - y(x))\big].$$

Заметим, что в рассматриваей модели Гауссовского поля критерий может быть вычислен аналитически:

$$EI(\mathbf{x}) = \begin{cases} (f_{\min} - \hat{f}(\mathbf{x}))\Phi(z(\mathbf{x})) + \hat{\sigma}(\mathbf{x})\phi(z(\mathbf{x})), \ \hat{\sigma}(\mathbf{x}) > 0 \\ 0, \ \hat{\sigma}(\mathbf{x}) = 0 \end{cases}.$$

Здесь $\Phi(\cdot)$ и $\phi(\cdot)$ — функция и плотность стандартного нормального распределения, $z(\mathbf{x}) = \frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{\sigma}(\mathbf{x})}$, $f_{\min} = \min\limits_{i=1,\dots,k} y_i$.

Преимущества:

- Проявляет как локальные, так и глобальные свойства.

Недостатки:

- Сложность критерия и его оптимизации растёт с увеличением размера выборки.
- Вырождается в ноль в численном представлении.

Knowledge gradient

- $\hat{f}(\mathbf{u}|D \cup \{\mathbf{x}, y(\mathbf{x})\})$ — случайная величина, соответствующая значению аппроксимации в точке $\mathbf{u}$ при условии, что к выборке будет добавлена точка $\mathbf{x}$. Важно, что здесь величина $y(\mathbf{x})$ ещё не известна и является случайной.
- Новая точка теоретически находится по формуле:

$$\mathbf{x}_{new} = \arg\max_{\mathbf{x}\in\mathbb{X}} \mathbb{E}[\max_{\mathbf{u}\in\mathbb{X}} \hat{f}(\mathbf{u}|D \cup \{\mathbf{x}, y(\mathbf{x})\})] - \max_{\mathbf{u}\in\mathbb{X}} \hat{f}(\mathbf{u}|D).$$

Knowledge gradient

- На практике критерий аппроксимируется следующим образом:

$$\mathbf{x}_{new} = \arg\max_{\mathbf{x}\in\mathbb{X}} \mathbb{E}[\max_{j=1,\dots,k+1} \hat{f}(\mathbf{x}_j|D\cup\{\mathbf{x}, y(\mathbf{x})\})] - \max_{j=1,\dots,k} \hat{f}(\mathbf{x}_j|D), \tag{2}$$

  где k — размер выборки D.
- Так как при отсутствии шума Knowledge gradient совпадает с Expected Improvement, то он для этого случая наследует все преимущества и недостатки Expected Improvement.

Используемые классические методы оптимизации

Представляется интересным сравнить SBO методы с другими методами оптимизации. В данной работе мы рассмотрели два классических алгоритма глобальной оптимизации, а именно:

- Стохастический алгоритм, известный в литературе как метод имитации отжига;
- Детерминированный алгоритм глобальной оптимизации DIRECT.

Вырождение критерия EI

- В соответствии с критерием

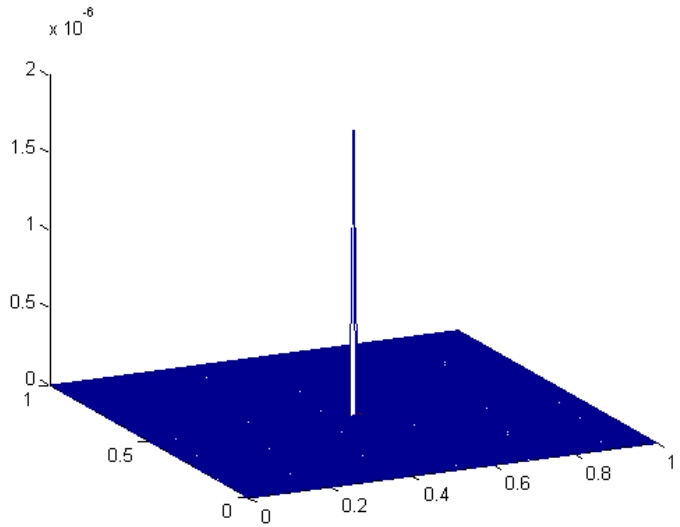
$$\mathbf{x}_{new} = \arg\max_{\mathbf{x}} \mathbb{E}[I] = \arg\max_{\mathbf{x}} \hat{\sigma}(\mathbf{x})(z(\mathbf{x})\Phi(z(\mathbf{x}))+\phi(z(\mathbf{x}))),$$

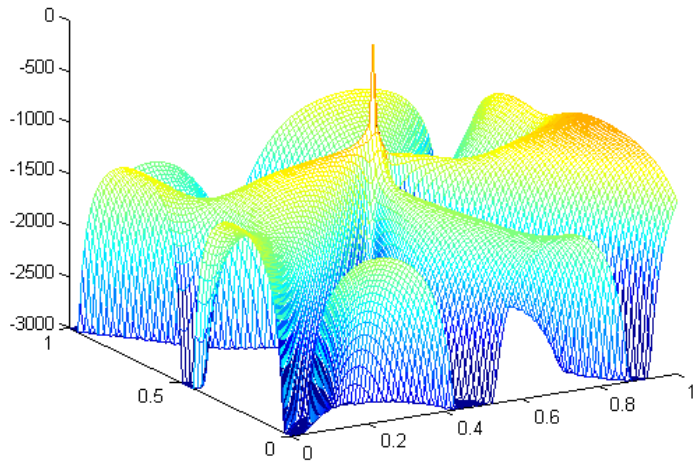
где $z(\mathbf{x}) = \frac{y_{\min} - \hat{f}(\mathbf{x})}{\hat{\sigma}(\mathbf{x})}$.

- Функция $\mathbb{E}[I]$ с учётом конечной машинной точности бывает равна нулю на большой части $\mathbb{X}$. По большей части это связано с малостью $\hat{\sigma}(\mathbf{x})$.
- Вместо $\mathbb{E}[I]$ будем использовать $\log(\mathbb{E}[I])$, которое можно корректно вычислить при $z(\mathbf{x}) > -38$. При $z(\mathbf{x}) \leq -38$ мы заменяем его на линейную по $z(\mathbf{x})$ функцию.

Вырождение критерия EI

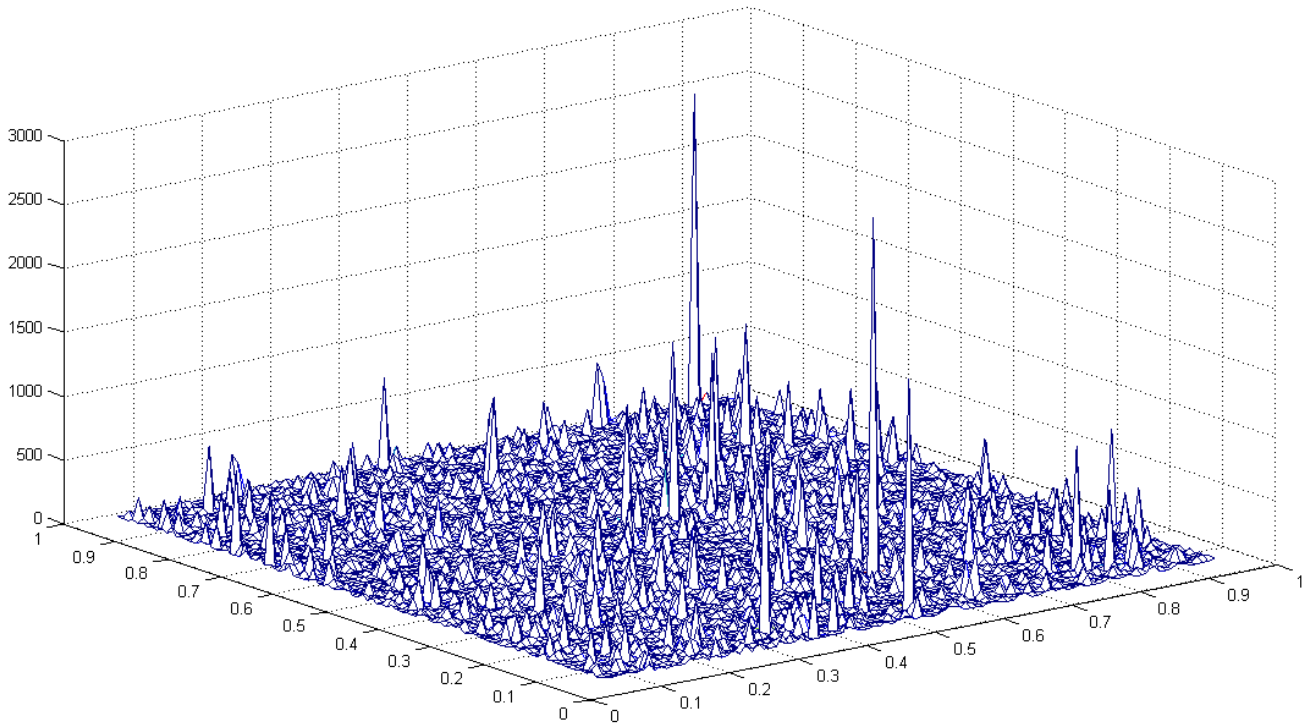
- Также проблемы вызывает случай, когда $\hat{\sigma}(\mathbf{x}|D) = 0$ в численном представлении. Мы без потерь заменяем $\hat{\sigma}$ на $\max(\hat{\sigma}, 10^{-100})$.
- Ниже приводится пример оптимизируемого критерия до и после замены для размерности равной двум:





Численный шум критерия KG

Наглядно проблему отражает следующий пример численного шума поверх критерия KG для размерности равной двум:

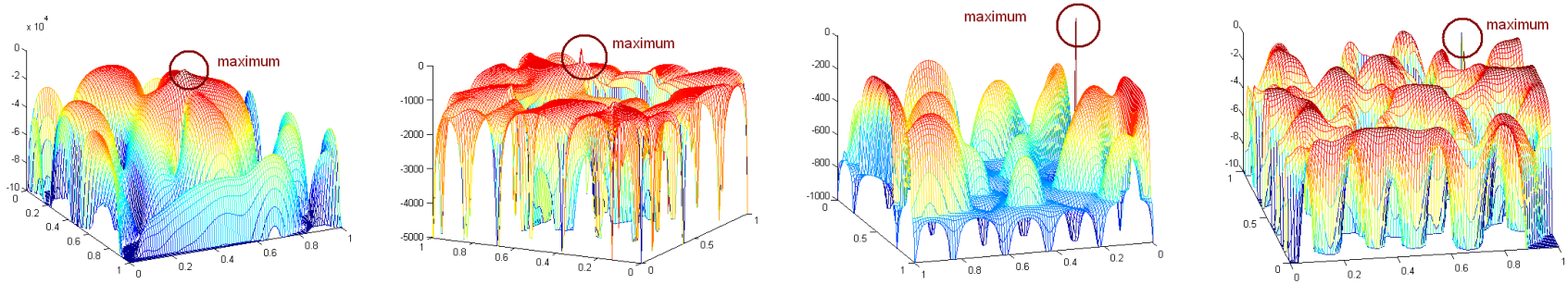


Численный шум критерия KG

- Главной причиной является плохая обусловленность некоторой матрицы S, которая подлежит обращению.
- Проблема решена стандартным образом: $S := S + diagonal(\lambda_{\min})$, где $\lambda_{\min}$ — такая минимально необходимая регуляризация, что число обусловленности матрицы S при ней не более 10^8: $Cond(S) \leq 10^8$.

Оптимизация критерия EI

Примеры критерия EI для $d = 2$:



- Видно, что критерий является сложным для оптимизации.
- Для выбора лучшего алгоритма оптимизации EI проводилось отдельное масштабное тестирование.
- Из работы алгоритма SBO было взято 72 реализации критерия EI для 3 размерностей, 6 целевых функций и 4 этапов работы SBO.

Оптимизация критерия EI

В ходе тестирования были опробованы следующие стандартные методы:

- Метод имитации отжига,
- DIRECT,
- Метод градиентного спуска,
- Случайный семплинг и метод гиперкубов,
- Методы, основанные на их комбинациях вышеизложенных.

Оптимизация критерия является подзадачей алгоритма SBO. В данном случае нас интересует сходимость по времени работы, а не по количеству обращений к критерию.

Экспериментально лучший алгоритм

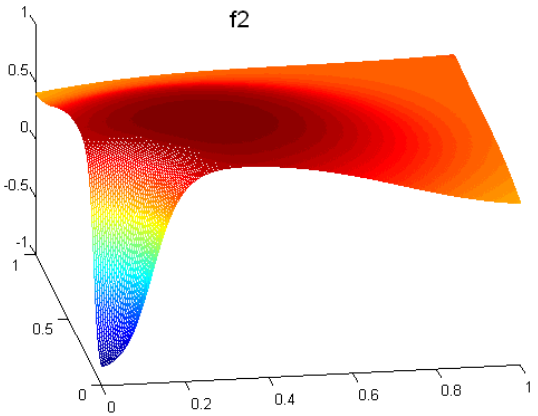
1. Генерируется выборка X_{rand} большого размера N состоящая из точек, равномерно и независимо распределённых по $\mathbb{X}$ и далее она сортируется по соответствующим значениям критерия в порядке убывания.
2. Из первой точки в выборке X_{rand} запускается градиентный алгоритм оптимизации, результат запоминается. Из выборки X_{rand} удаляются все точки, которые близки (что значит близки - параметр алгоритма) к первой точке (сама она тоже удаляется).
3. Предыдущий шаг повторяются небольшое число раз M. Алгоритм выдает лучший результат среди достигнутых градиентным спуском.

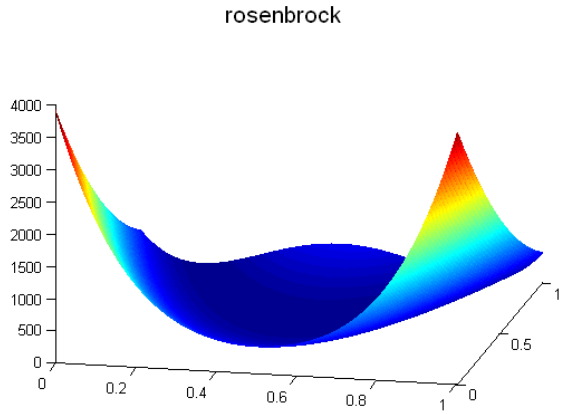
Постановка тестов

- Размерности $d \in \{2, 3, 4, 6, 10\}$.
- Для каждой тестовой функции методом латинских гиперкубов генерировалось 10 начальных обучающих выборок размером $2d + 3$ точек. Остальные точки ставились по общему алгоритму.
- Всего было использовано 50 обращений к целевой функции.
- Тестирование проводилось на данных без шума, поэтому KG и EI показали схожие результаты, и результаты для KG не приводятся.

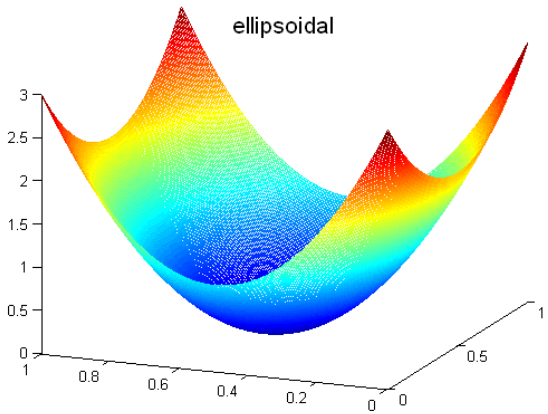
Используемые тестовые функции

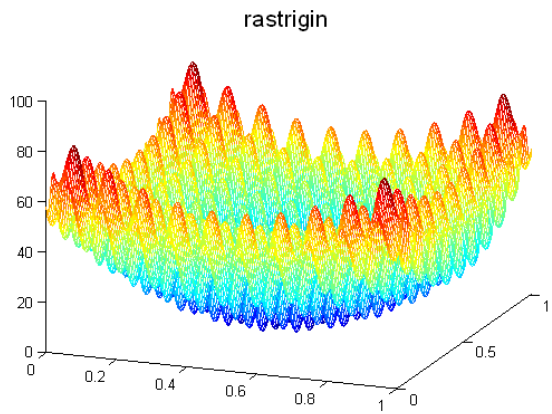
Для демонстрации экспериментальных результатов был использован набор из 24 разнообразных тестовых функций, которые применяются для тестирования задач оптимизации. Ниже представлено несколько примеров тестовых функций для $d = 2$.

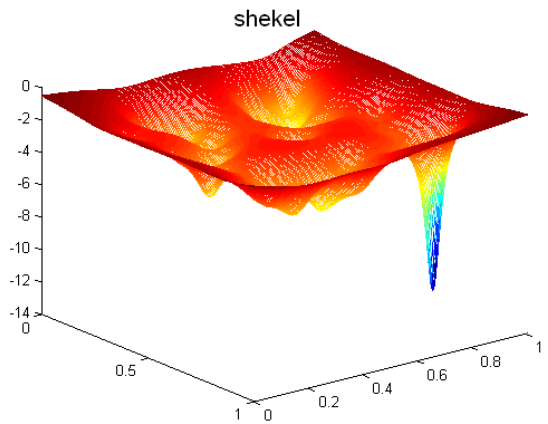


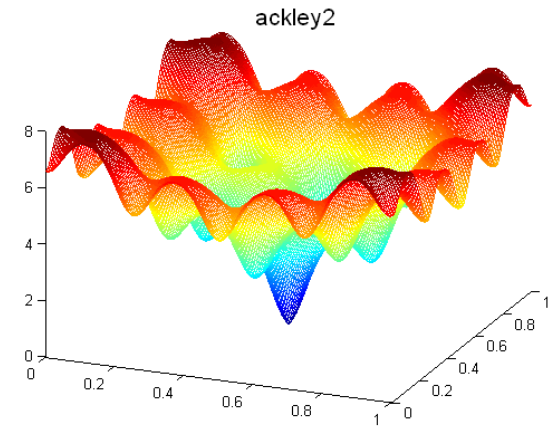


Используемые тестовые функции









Кривые Долан-Мора

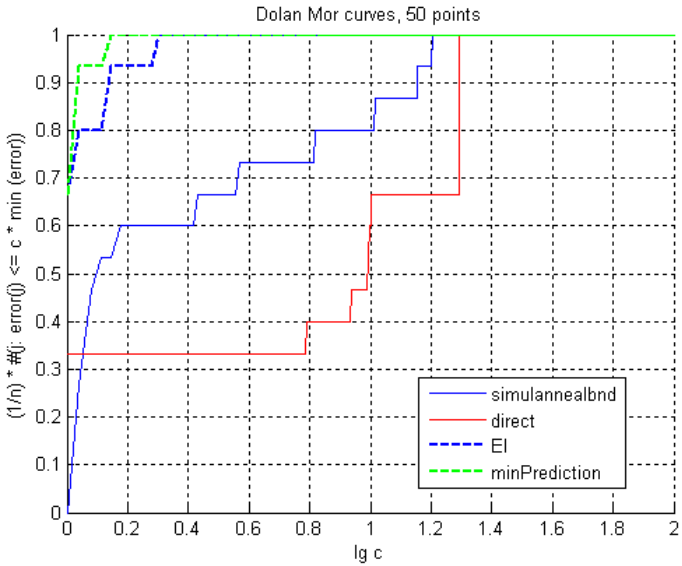


Рис. : Размерность $d = 2$

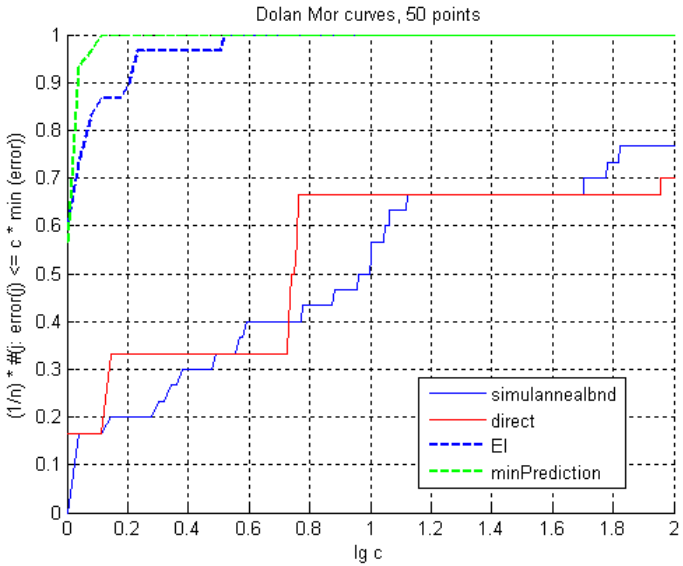


Рис. : Размерность $d = 3$

Кривые Долан-Мора

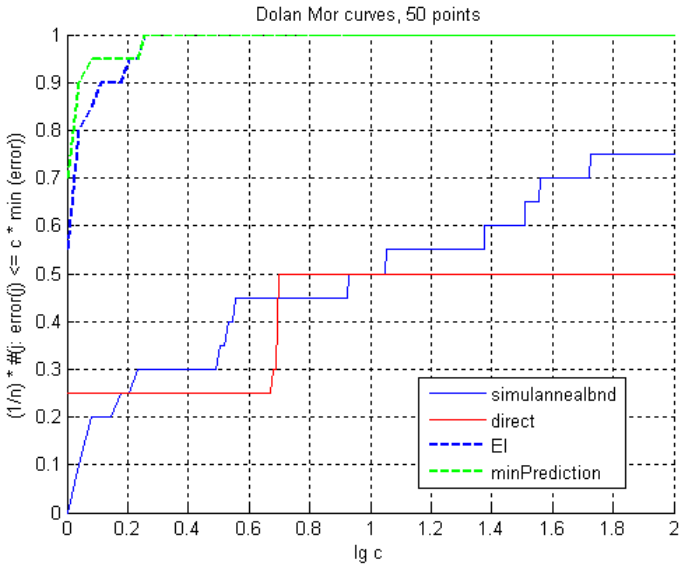


Рис. : Размерность $d = 4$

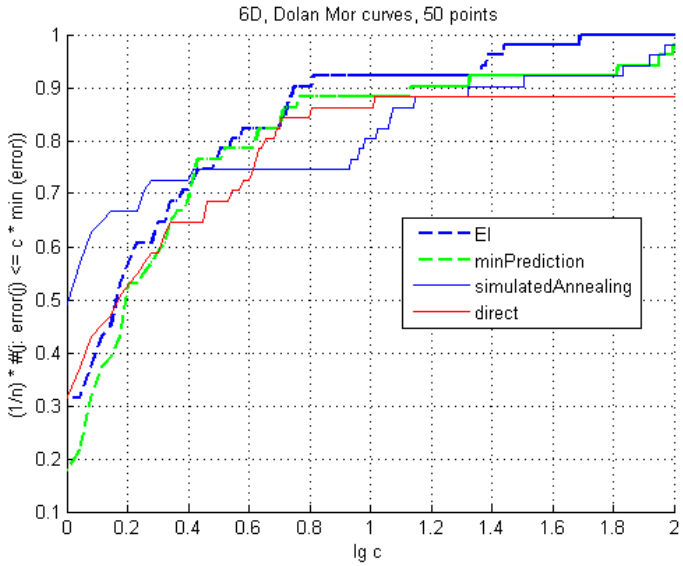


Рис. : Размерность $d = 6$

Кривые Долан-Мора

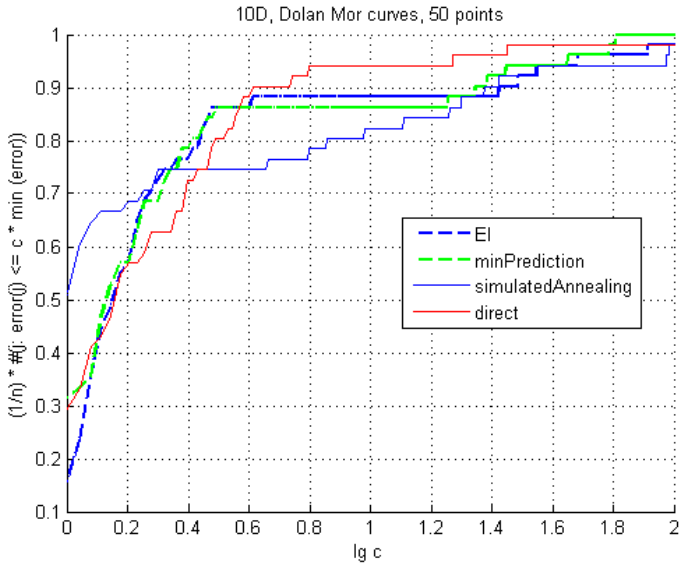


Рис. : Размерность $d = 10$

Выводы

- Методы SBO в целом показали преимущество над иными методами оптимизации.
- Время работы SBO методов значительно превосходит время работы других рассмотренных методов, если не учитывать время расчёта целевой функции.
- Интерес представляет сравнение методов SBO между собой. Отдельное тестирование показало, что метод EI может работать лучше, если качественнее оптимизировать его критерий.
- Качество оптимизации методами SBO главным образом зависит от качества аппроксимации целевой функции.

Выводы

- Вклад автора заключается в предложении эффективных численных реализаций алгоритмов и в проведении их сравнения.
- В качестве основных направлений дальнейшей работы предполагается построение хорошей теоретической постановки задачи SBO, которая позволит вывести новые критерии суррогатной оптимизации и исследовать их теоретические свойства.
- Также планируется исследование работы алгоритмов на зашумлённых и реальных данных.

Спасибо за внимание!