Сравнительный анализ процедур оптимизации на основе гауссовских процессов

Евгений Бурнаев[1,2,4], Максим Панов[1,2,4], Даниил Кононенко[1,2,3], Иван Коноваленко[2,3]
1. Институт Проблем Передачи Информации,
127994, г.Москва, ГСП-4, Большой Каретный переулок, 19, стр.1
2. DATADVANCE,
105064, г.Москва, Садово-Черногрязская улица, 13/3
3. МФТИ,
141700, г.Долгопрудный, Институтский переулок, 9
4. PreMoLab,
141700, г.Долгопрудный, Институтский переулок, 9
{evgeny.burnaev,maxim.panov,daniil.kononenko,ivan.konovalenko}@datadvance.net

Аннотация

В некоторых практических приложениях требуется оптимизировать функцию, расчёт одного значения которой может занимать значительное время.

Одним из способов оптимизации таких функций является оптимизация на основе суррогатных моделей, основная идея которой состоит в построении аппроксимации (суррогатной модели) целевой функции и дальнейшем ее использовании при оптимизации. В данной работе в качестве суррогатной модели рассматривается стохастическая модель гауссовского процесса.

В работе рассмотрено несколько методов суррогатной оптимизации и проведено их сравнение с классическими методами оптимизации на большом количестве тестовых функций различных размерностей.

1. Введение

В современной инженерной практике широкое распространение получили методы, основанные на построении так называемых суррогатных моделей (моделей на основе данных). Одной из основных задач, которые решаются на основе таких моделей, является задача суррогатной оптимизации [1] (SBO - Surrogate-Based Optimization). Главной подзадачей SBO является аппроксимация неизвестной зависимости по данным [2, 3]. Наиболее популярная модель для построения аппроксиматоров, основанная на гауссовских процессах [4, 5, 6], используется в большом количестве разнообразных прикладных задач, включая концептуальное проектирование [7], структурную оптимизацию [8], многокритериальную оптимизацию при проектировании [9], конструирование в аэрокосмической [10] и автомобильной отраслях [11]. Цель данной работы состоит в том, чтобы провести сравнительный анализ процедур суррогатной оптимизации и классических методов глобальной оптимизации.

Статья устроена следующим образом:

- в разделе 2 даётся инженерная постановка задачи;
- в разделе 3 рассказывается про построение аппроксиматора на основе гауссовских процессов;
- в разделах 4 и 5 описываются методы оптимизации, используемые в тестировании;
- в разделах 6 и 7 рассказывается о результатах сравнительного анализа.

2. Инженерная постановка задачи

Задача оптимизации может быть поставлена следующим образом. Пусть

$$y = f(x) \tag{1}$$

заданная целевая функция, где $x \in \mathbb{X} \subset \mathbb{R}^d$ - вектор-строка, $y \in \mathbb{R}$. На практике $f(x)$ может быть зашумлена, что сильно усложняет оценку градиента, а значит и работу методов, использующих градиент.

$$D = (X, Y) = \{x_i, y_i = f(x_i)\}_{i=1}^{k} \tag{2}$$

выборка k точек из $\mathbb{X}$ и соответствующих значений целевой функции, матрица $k \times (d+1)$.

Получение значений целевой функции может быть вычислительно сложной задачей. В других случаях, для получения значения функции необходимо поставить дорогой либо продолжительный физический эксперимент. Таким образом, необходимо получить наилучшее значение целевой функции за небольшое, фиксированное число обращений к ней.

Более формально, к $f(x)$ можно обратиться заданное число раз n, выбрав набор точек $X^* = \{x_i^*\}_{i=1}^n$. Тогда $Y^* = f(X^*) = \{y_i^* = f(x_i^*)\}_{i=1}^n$ полученный набор значений целевой функции. Наш лучший результат - это минимальный из них $y_{min}^* = \min\limits_{i=1,\dots,n} y_i^*$. Задача состоит в том, чтобы выбирать X^*, минимизируя y_{min}^*, т.е.:

$$y_{min}^* = \min_{i=1,\dots,n} y_i^* \to \min_{X^*}. \tag{3}$$

Важно, что для выбора следующей точки можно пользоваться значениями целевой функции в предыдущих точках и модельными предположениями о функции, описанными в следующем разделе.

3. Аппроксиматор на основе гауссовских процессов

3.1. Гауссовские процессы

Гауссовский процесс является одним из возможных способов задания распределения на пространстве функций. Гауссовский процесс $f(x)$ полностью определяется своей функцией среднего $m(x) = \mathbb{E}[f(x)]$ и ковариационной функцией $cov(y,y') = k(x,x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]$. Если положить функцию среднего нулевой: $m(x) = \mathbb{E}[f(x)] = 0$, а ковариационную функцию считать известной, то функция апостериорного (для заданной обучающей выборки) среднего значения гауссовского процесса в точках контрольной выборки X_* выглядит следующим образом [12]:

$$\hat{f}(X^*) = K_* K^{-1} Y, \tag{4}$$

где $K_* = K(X^*, X) = \left[k(x_i^*, x_j), i = 1,\dots,N_*; j = 1,\dots,N\right]$, $K = K(X,X) = \left[k(x_i, x_j), i, j = 1,\dots,N\right]$.

В типичных, более реалистичных ситуациях при моделировании мы не имеем доступа непосредственно к значениям функции, а наблюдаем их только в зашумленном виде:

$$y(x) = f(x) + \varepsilon(x), \tag{5}$$

где шум $\varepsilon(x)$ моделируется независимыми одинаково распределенными нормальными случайными величинами с нулевым средним и дисперсией $\tilde{\sigma}^2$. В таком случае наблюдения $y(x)$ будут гауссовским процессом с нулевым средним и ковариационной функцией $cov(y(x), y(x')) = k(x,x') + \tilde{\sigma}^2$. Таким образом, функция апостериорного (для заданной обучающей выборки) среднего значения гауссовского процесса $f(x)$ в точках контрольной выборки X^* принимает вид:

$$\hat{f}(X^*) = K_* \left(K + \tilde{\sigma}^2 I\right)^{-1} Y, \tag{6}$$

где I - единичная матрица размера $N \times N$. Это выражение используется в качестве аппроксимации.

Заметим, что наличие в формуле (6) дисперсии шума $\tilde{\sigma}^2$ фактически приводит к регуляризации, что позволяет улучшить обобщающую способность аппроксиматора. При этом апостериорная ковариационная функция гауссовского процесса в точках контрольной выборки имеет вид:

$$\mathbb{V}\left[\hat{f}(X^*)\right] = K(X^*, X^*) + \tilde{\sigma}^2 I_* - K_* \left(K + \tilde{\sigma}^2 I\right)^{-1} K_*^T \tag{7}$$

где $K(X^*, X^*) = \left[k(x_i^*, x_j^*), i, j = 1,\dots,N_*\right]$, I_* - единичная матрица размера $N_* \times N_*$.

Дисперсии гауссовского процесса в точках контрольной выборки могут быть использованы как оценки ожидаемой ошибки аппроксимации в этих точках. Нет необходимости вычислять по формуле (7) всю матрицу $\mathbb{V}\left[\hat{f}(X^*)\right]$, достаточно вычислить только элементы ее главной диагонали $\hat{\sigma}^2(X^*) = \hat{\sigma}^2(X^*|D) = diag\left(\mathbb{V}\left[\hat{f}(X^*)\right]\right)$, которые и являются искомыми апостериорными дисперсиями.

При работе с реальными данными ковариационная функция породившего их гауссовского процесса, как правило, не известна, поэтому необходимо уметь идентифицировать ее по данным.

3.2. Нахождение параметров гауссовского процесса

Предположим, что ковариационная функция гауссовского процесса является членом некоторого параметрического семейства $k(x,x') = k(x,x'|a)$, где $a \in \mathbb{R}^K$ - вектор параметров ковариационной функции. Семейство $k(x,x'|a)$ обычно берется из класса так называемых стационарных ковариационных функций, т. е. функций, значение которых зависит только от разности значений аргументов $k(x,x'|a) = k(x - x'|a)$. Значение параметра a предлагается восстанавливать по обучающей выборке D_{learn}, исходя из принципа максимума правдоподобия. Для этого выпишем логарифм правдоподобия гауссовского процесса в точках обучающей выборки [12]:

$$\log\ p(Y|X,a,\tilde{\sigma}) = -\frac{1}{2} Y^T (K + \tilde{\sigma}^2 I_N)^{-1} Y - \\ - \frac{1}{2} \log|K + \tilde{\sigma}^2 I| - \frac{n}{2} \log 2\pi, \tag{8}$$

где $|K + \tilde{\sigma}^2 I|$ - детерминант матрицы $K + \tilde{\sigma}^2 I$.

Кроме параметров a ковариационной функции параметром функционала (8) является также значение дисперсии шума наблюдений $\tilde{\sigma}^2$, которое также

можно настраивать по обучающей выборке. Таким образом, нахождение оптимальных значений параметров сводится к отысканию максимума правдоподобия по параметрам:

$$\log\ p(Y|X,a,\tilde{\sigma}) \to \max_{a,\tilde{\sigma}}. \tag{9}$$

Выбор конкретного семейства ковариационных функций $k(x,x'|a)$ обычно продиктован соображениями удобства, а также априорными представлениями о свойствах аппроксимируемой зависимости. В данной работе мы используем ковариационные функции вида $k(x-\tilde{x}|a) = \sigma^2 \exp\left\{-\sum_{i=1}^{d} \theta_i^2 |x_i - \tilde{x}_i|^2\right\}$, где параметры $a = \left\{\theta_i, i = 1,\ldots,d; \sigma\right\}$ настраиваются по обучающей выборке при решении задачи (9).

4. Суррогатные методы оптимизации (SBO)

Surrogate-Based Optimization (SBO) - метод оптимизации, базирующийся на аппроксимации целевой функции. Ниже подробно описан алгоритм работы SBO:

1. Для работы аппроксиматора необходима начальная обучающая выборка D. В данной работе выборка имеет размер $2d+3$ точек и генерируется методом латинских гиперкубов.
2. По выборке D строится аппроксимация целевой функции $\hat{f}(x|D)$ и апостериорной дисперсии $\hat{\sigma}^2(x|D)$.
3. Новая точка получается из максимизации критерия $C(x|D)$, основанного на $\hat{f}(x|D)$ и $\hat{\sigma}^2(x|D)$
$$x_{new} = \arg\max_{x\in\mathbb{X}} C(x|D). \tag{10}$$
4. Новая точка добавляется к выборке: $D := D \cup (x_{new}, f(x_{new}))$.
5. Если лимит точек не исчерпан, переходим к шагу 2. Иначе переходим к шагу 6.
6. Алгоритм возвращает в качестве выхода минимальное значение функции в выборке D и соответствующую ему точку выборки: $(x_{min}, y_{min} = f(x_{min}))$.

В следующих разделах рассмотрим конкретные виды критериев $C(x|D)$.

4.1. Expected Improvement (EI)

Expected Improvement [13] (ожидаемое улучшение) – популярный алгоритм, совмещающий в себе глобальность оптимизации и концентрацию выборки точек в тех областях $\mathbb{X}$, где аппроксимация имеет меньшие значения. Имеет простую интуитивную интерпретацию. Новая точка ставится так, чтобы максимизировать математическое ожидание улучшения, где улучшение - это зависящая от $x \in \mathbb{X}$ случайная величина, соответствующая значению, на которое функция в точке x меньше достигнутого ранее минимума. Ниже дано математическое изложение метода.

Пусть $y_{min} = \min\limits_{i=1,\ldots,k} y_i$ - это минимальное значение функции в выборке. Введем понятия улучшения, как случайной величины следующего вида:

$$I = \max(0, y_{min} - Y),$$

где

$$Y = Y(x|D) \sim \mathbb{N}(\hat{f}(x|D), \hat{\sigma}^2(x|D)).$$

Обозначим

$$t = t(x|D) = \frac{y_{min} - \hat{f}(x|D)}{\hat{\sigma}(x|D)}.$$

Новая точка добавляется так, чтобы максимизировать математическое ожидание улучшения:

$$\mathbb{E}(I) = \hat{\sigma}(x|D)(t\Phi(t) + \phi(t)),$$

где $\Phi(\cdot)$ - функция распределения стандартного нормального распределения, $\phi(\cdot)$ - функция плотности стандартного нормального распределения.

Тогда критерий выбора новой точки x_{new} принимает вид:

$$x_{new} = \arg\max_{x} \hat{\sigma}(x|D)(t\Phi(t) + \phi(t)). \tag{11}$$

Остановимся подробнее на особенностях работы критерия Expected Improvement. Преимущества:

- Теоретические свойства и эмпирические результаты работы алгоритма указывают на то, что критерий плотно заполняет множество $\mathbb{X}$ при стремлении размера выборки к бесконечности.
- Новые точки чаще попадают туда, где аппроксимация имеет малое значение.

Недостатки:

- Сложность критерия и его оптимизации растёт с увеличением размера выборки.
- Есть следующая численная проблема при использовании критерия. При малых значениях $\hat{\sigma}(x|D)$ критерий $\mathbb{E}(I)$ бывает настолько близок к нулю, что в численном представлении $\mathbb{E}(I) = 0$. Это может происходить на большей части $\mathbb{X}$ и мешать оптимизации критерия. Проблема решена заменой $\mathbb{E}(I)$ на $\log(\mathbb{E}(I))$ с последующей непрерывной аппроксимацией критерия в областях малых значений $\hat{\sigma}(x|D)$.

4.2. Knowledge gradient

Данный метод является обобщением Expected Improvement на случай, когда значения целевой функции известны с точностью до аддитивного независимого в каждой точке гауссовского шума, который имеет нормальное распределение с нулевым средним и постоянной дисперсией λ.

Пусть $\mu(u|x_{new}=x)$ - случайная величина, соответствующая значению аппроксимации в точке u при условии, что $x_{new}=x$. Величина $f(x_{new})$ ещё неизвестна.

Новая точка находится как [14]:

$$x_{new} = \arg\max_{x\in\mathbb{X}} \mathbb{E}[\max_{j=1,\dots,k+1} \mu(x_j|x_{new}=x)] - \max_{j=1,\dots,k} \hat{f}(x_j|D), \tag{12}$$

где k - размер выборки D.

Исходя из модельных предположений, описанных в разделе 3, можно найти $\mu(x_j|x_{new}=x), j=1,\dots,k+1$. Заметим, что $x_{new}=x_{k+1}$. Приведём рабочие формулы без вывода.

Обозначим

$$\Sigma^0 = \left[k(x_i,x_j), i,j=1,\dots,k\right],$$

$$S = \Sigma^0 + diagonal(\lambda),$$

$$\overline{\Sigma}^0 = \left[k(x_i,x_j), i,j=1,\dots,k+1\right],$$

$$K = \overline{\Sigma}^0 \begin{bmatrix} I_k \\ - \\ \vec{0}^T \end{bmatrix} S^{-1},$$

$$\overline{\Sigma}^k = (I_{k+1} - K[I_k|\vec{0}])\overline{\Sigma}^0,$$

где I_k - единичная матрица, $\vec{0}$ - нулевой вектор-столбец размера k.

Тогда

$$\begin{bmatrix} \mu(x_1|x_{new}=x) \\ \mu(x_2|x_{new}=x) \\ \vdots \\ \mu(x_{k+1}|x_{new}=x) \end{bmatrix} = \begin{bmatrix} \hat{f}(x_1|D) \\ \hat{f}(x_2|D) \\ \vdots \\ \hat{f}(x_{k+1}|D) \end{bmatrix} + \breve{\sigma}(\overline{\Sigma}^k)Z, \tag{13}$$

где $\breve{\sigma}(\Sigma) = \frac{\Sigma e_x}{\sqrt{\lambda + e_x^T \Sigma e_x}}$, $Z \sim \mathcal{N}(0,1)$. Здесь e_x - вектор-столбец из нолей размера $k+1$, но с единицей в последней компоненте.

Так как при отсутствии шума Knowledge gradient совпадает с Expected Improvement, то он для этого случая наследует все преимущества и недостатки Expected Improvement. Критерий требует численного подсчета [14].

4.3. Minimum of approximation (minPrediction)

В соответствии с этим методом новая точка добавляется туда, где значение аппроксимации минимально.

$$x_{new} = \arg\min_x \hat{f}(x|D). \tag{14}$$

Использование minPrediction может быть особенно полезно в том случае, когда аппроксимация хорошо приближает целевую функцию. Тогда минимум аппроксимации и минимум целевой функции близки. Остановимся подробнее на особенностях работы критерия Minimum of approximation. Преимущества:

- Качество работы критерия сильно зависит от качества аппроксимации целевой функции. Если качество аппроксимации хорошее, то критерий за счет своих локальных свойств позволяет добиться хороших результатов. Для сравнения, Expected Improvement может даже в случае хорошей аппроксимации ставить точки в области, которые далеки от оптимума.
- Критерий очень просто считается, т. к. аппроксиматор на основе гауссовских процессов легко считается.

Недостатки:

- Метод локальный: генерируемая им последовательность точек не обязательно заполняет $\mathbb{X}$ плотно. Таким образом, сходимость с ростом размера выборки не гарантирована.

5. Классические методы оптимизации, используемые при тестировании

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 [18];
- Детерминированный алгоритм глобальной оптимизации DIRECT [19].

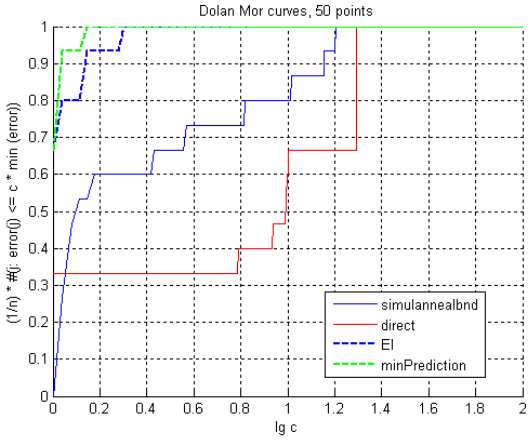
6. Экспериментальные результаты

Тестирование проходило на размерностях $d \in \{2,3,4,6,10\}$. Для демонстрации экспериментальных результатов был использован большой набор разнообразных тестовых функций, которые применяются для тестирования задач оптимизации [15, 16]. Всего тестирование проводилось на 24 различных функциях, для каждой из которых методом латинских

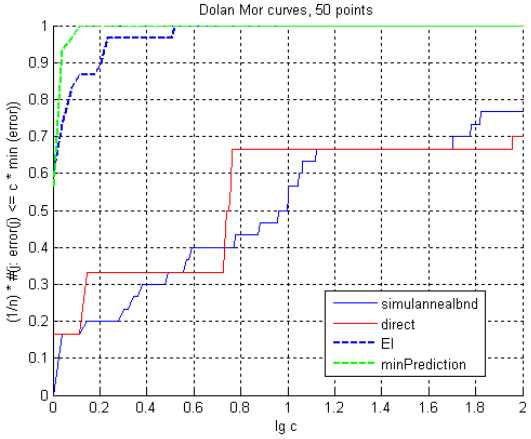
гиперкубов генерировалось 10 начальных обучающих выборках размером $2d+3$ точек. Это минимальный необходимый для работы аппроксиматора размер выборки. Остальные точки ставились по общему алгоритму. Всего было использовано 50 обращений к целевой функции. Критерий оптимизировался методом имитации отжига с количеством обращений к критерию: $500+200d$. Тестирование проводилось на данных без шума, поэтому KG и EI показали схожие результаты, и результаты для KG не приводятся.

Для удобства результаты представлены в виде кривых Долан-Мора [17]. Чем выше кривая находится на графике, тем выше качество работы соответствующего алгоритма. Результаты разбиты по размерностям.

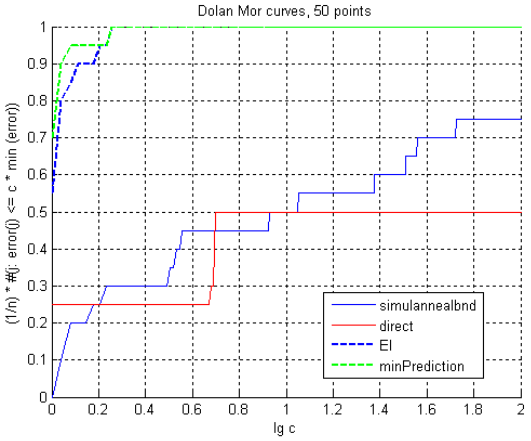
6.1. Кривые Долан-Мора, 2D



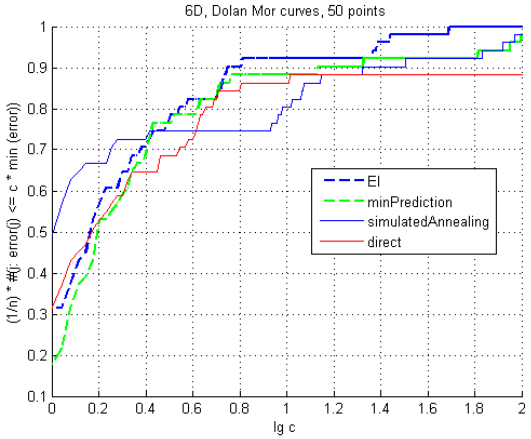
6.2. Кривые Долан-Мора, 3D



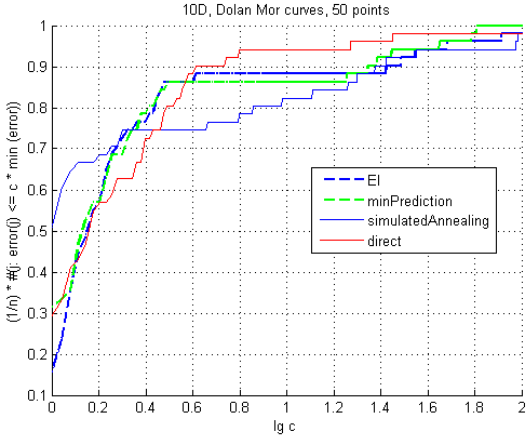
6.3. Кривые Долан-Мора, 4D



6.4. Кривые Долан-Мора, 6D



6.5. Кривые Долан-Мора, 10D



7. Выводы

На основании проведенного исследования можно сделать следующие основные выводы:

- Методы SBO в целом показали преимущество над другими методами оптимизации в смысле скорости сходимости.
- Время работы SBO методов значительно превосходит время работы других рассмотренных методов, поэтому первые могут выигрывать по времени только тогда, когда целевая функция является очень тяжёлой для расчёта.
- 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.

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.

Работа выполнена при поддержке Лаборатории структурных методов анализа данных в предсказательном моделировании, МФТИ, грант правительства РФ дог. 11.G34.31.0073.

Список литературы

[1] Jones, D. R., Schonlau, M. and Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. Journal of Global Optimization, 13(4), 455 - 492.

[2] Бернштейн А.В., Бурнаев Е.В., Кулешов А.П. Интеллектуальный анализ данных в метамоделировании. // Труды 17 Всероссийского Семинара "Нейроинформатика и ее приложения к Анализу Данных Красноярск, 2009 – с. 23-28.

[3] Forrester A., Sobester A., Keane A. Engineering Design via Surrogate Modelling. A Practical Guide. – Wiley, 2008. – 238 p.

[4] Giunta A., Watson L. T.A Comparison of Approximation Modeling Technique: Polynomial Versus Interpolating Models. // 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Vol. 1, AIAA, Reston, VA, 1998 – pp. 392–404.

[5] Simpson T. W., Booker A. J., Ghosh S., Giunta A., Koch P. N., Yang, R. J. Approximation Methods in Multidisciplinary Analysis and Optimization: A Panel Discussion. // Structural and Multidisciplinary Optimization, Vol. 27, No. 5, 2004 – pp. 302–313.

[6] Batill S. M., Renaud J. E., Gu X., Modeling and Simulation Uncertainty in Multidisciplinary Design Optimization // AIAA Paper 2000-4803, Sept. 2000.

[7] Pacheco J. E., Amon C. H., Finger S. Bayesian Surrogates Applied to Conceptual Stages of the Engineering Design Process // Journal of Mechanical Design, Vol. 125, No. 4, 2003 – pp. 664–672.

[8] Booker A. J., Dennis J. E., Frank P. D., Serafini D. B. Torczon, V., Trosset, M. A Rigorous Framework for Optimization of Expensive Functions by Surrogates // Structural Optimization, Vol. 17, No. 1, 1999, – pp. 1–13.

[9] Koch P. N., Wujek B. A., Golovidov O., Simpson, T. W. Facilitating Probabilistic Multidisciplinary Design Optimization Using Kriging Approximation Models // AIAA Paper 2002-5415, Sept. 2002.

[10] Simpson T. W., Maurey T. M., Korte J. J., and Mistree F. Kriging Metamodels for Global Approximation in Simulation-Based Multidisciplinary Design Optimization // AIAA Journal, Vol. 39, No. 12, 2001 – pp. 2233–2241.

[11] Yang R. J., Wang N., Tho C. H., Bobineau J. P., and Wang B. P. Metamodeling Development for Vehicle Frontal Impact Simulation // American Society of Mechanical Engineers, ASME Design Engineering Technical Conf.—Design Automation Conf., DETC2001/DAC-21012, Sept. 2001.

[12] Rasmussen C.E., Williams C.K.I. Gaussian Processes for Machine Learning. – the MIT Press, 2006.

[13] Jones R. R., Sconlau M., Welch W. J. Efficient Global Optimization of Expensive Black-Box Functions // Journal of Global Optimization 13: 455–492, 1998.

[14] Scott W., Frazier P., Powell W. The Correlated Knowledge Gradient for Simulation Optimization of Continuous Parameters Using Gaussian Process Regression

[15] GDR MASCOT-NUM Toy Functions benchmark. http://gdr-mascotnum.math.cnrs.fr/data2/benchmarks/jm.pdf

[16] Lappeenranta University of Technology: evoluationary computation pages - the function testbed. http://www.it.lut.fi/ip/evo/functions/functions.html

[17] Dolan E. D., More J. J. Benchmarking optimization software with performance profiles // Mathematical Programmming, Ser. A 91, 2002 – pp. 201–213

[18] Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. Optimization by Simulated Annealing // Science, New Series, Vol. 220, No. 4598, 1983 – pp. 671-680.

[19] Jones, D.R., Perttunen C.D., and Stuckman B. E. Lipschitzian optimization without the Lipschitz constant // Journal of Optimization Theory and Application 79 (1), 1993 – pp. 157–181.