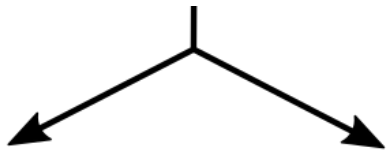
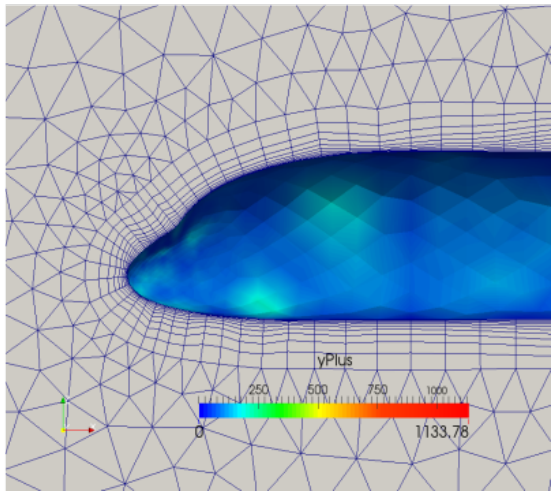
Минимаксный подход к моделированию данных разной точности

Евгений Бурнаев

(совместная работа
с А.А. Зайцевым)

Сколтех
ИППИ РАН
ВШЭ



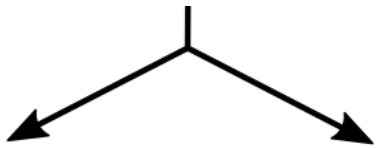


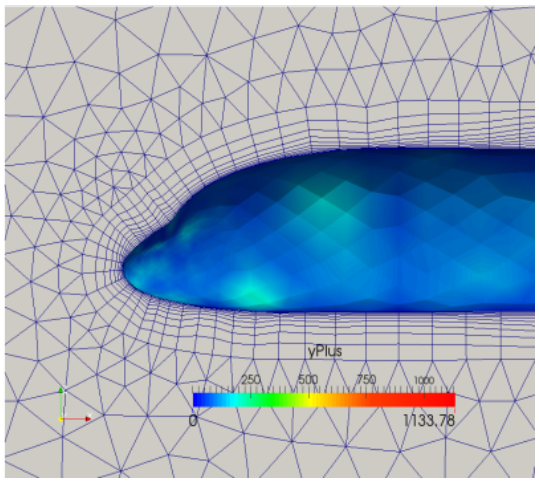
1. Задача выбора оптимального отношения размеров выборок
2. Регрессия на основе гауссовских процессов
3. Минимаксная ошибка интерполяции для регрессии на основе гауссовских процессов
4. Минимаксная ошибка интерполяции для данных разной точности
5. Оптимальное соотношение между размерами выборок данных разной точности
6. План экспериментов

Задача моделирования на данных в инженерном проектирования

- Выбрать план экспериментов $D = \{\mathbf{x}_i\}_{i=1}^n$.
- Подсчитать характеристики объектов для выбранного плана экспериментов $\mathbf{u} = \{u(\mathbf{x}_i)\}_{i=1}^n$.
- Построить модель $\widehat{u}(\mathbf{x}) \approx u(\mathbf{x})$ по выборке данных $S = (D, \mathbf{u})$.

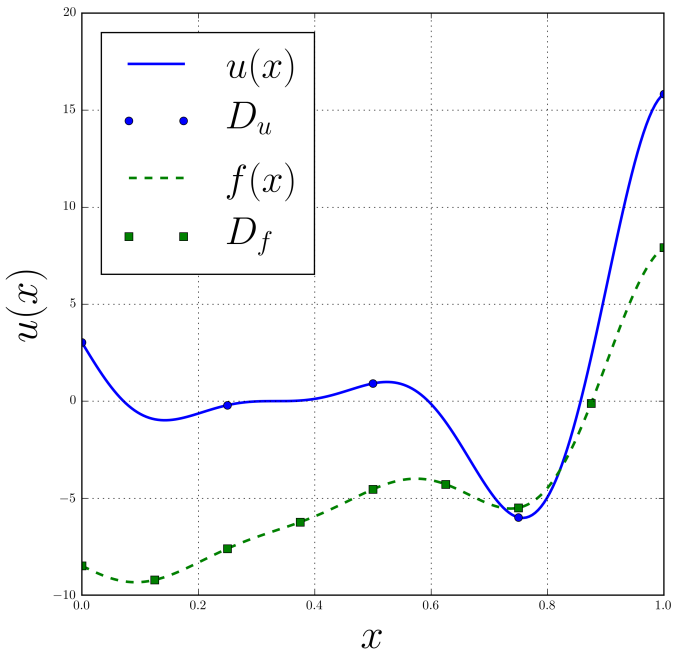
Грубая функция $f(\mathbf{x})$ и точная функция $u(\mathbf{x})$ моделируют один и тот же физический процесс, но с разной точностью





Задача регрессии для данных разной точности

- Задана выборка данных, порожденных грубой функцией, $S_f = (D_f, \mathbf{f}) = \left\{\mathbf{x}_i^f, f(\mathbf{x}_i^f)\right\}_{i=1}^{n_f}$, и выборка данных, порожденный точной функцией $S_u = (D_u, \mathbf{u}) = \{\mathbf{x}_i^u, u(\mathbf{x}_i^u)\}_{i=1}^{n_u}$ с
- $\mathbf{x}_i^f, \mathbf{x}_i^u \in \mathbb{R}^d$, $f(\mathbf{x}), u(\mathbf{x}) \in \mathbb{R}$.



Мы строим модель $\widehat{u}(\mathbf{x}) \approx u(\mathbf{x})$ точной функции, используя S_f и S_u.

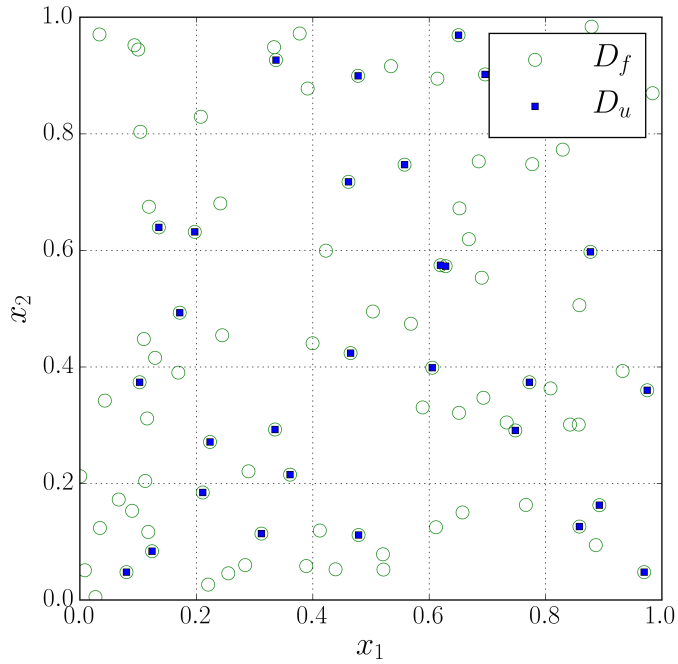
Инженеры решают задачу планирования эксперимента

- Стоимость вычисления точной функции стоит c, грубой — 1:
- Нужно выбрать размеры выборок точных данных n_u и грубых данных n_f, чтобы уложижться в заданный бюджет

$$\Lambda = cn_u + n_f.$$

- и минимизировать ошибку интерполяции

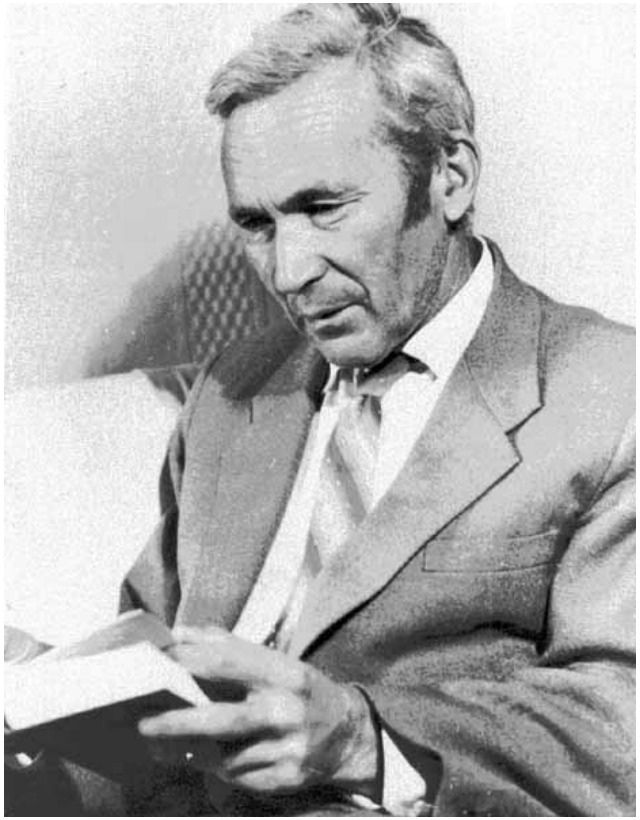
$$\int_{[0,1]^d} (u(\mathbf{x}) - \widehat{u}(\mathbf{x}))^2 d\mathbf{x}.$$



Задачу будет решать следующим образом:

1. Найти минимаксную ошибку для регрессии на основе гауссовских процессов для данных одной точности в случае $d \geq 1$.
2. Найти минимаксную ошибку интерполяции для данных разной точности.
3. Получить соотношение между размерами выборок разной точности, минимизирующее минимаксную ошибку интерполяции.

Колмогоров и Винер получили ошибку интерполяции в точке для $d = 1$



А.Н. Колмогоров

Н.Винер

Штайн написал книгу, посвященную ошибке интерполяции в точке

- Ибрагимов, И.А и Розанов, Ю.А. *Гауссовские случайные процессы*, 1970
- Stein, M.L. *Interpolation of spatial data: some theory for kriging*, 1997

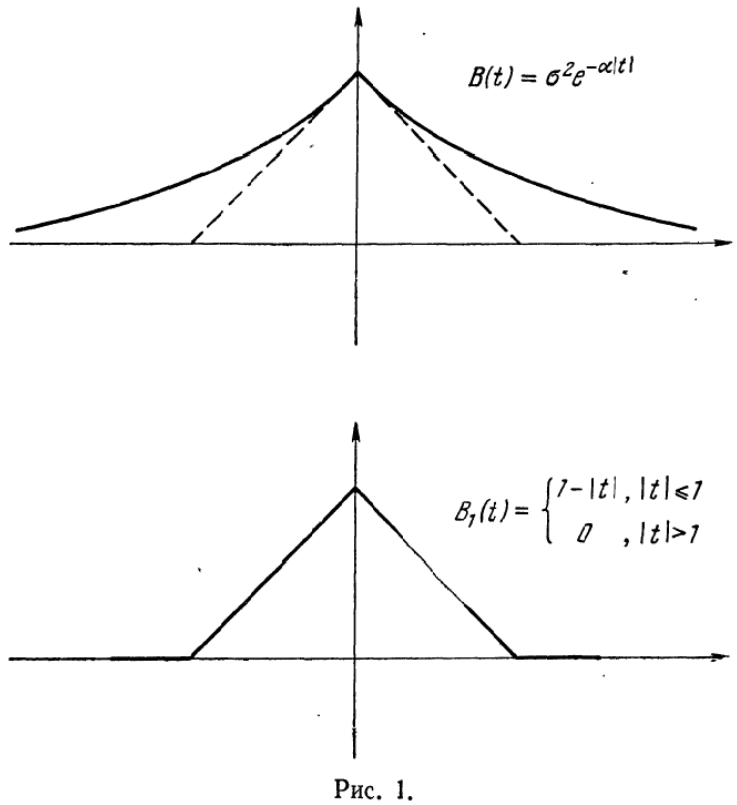


Рис. 1.

Современные работы рассматривают математическое ожидание ошибки на области

- Голубев, Г.К. и Крымова, Е.А. *Об интерполяции гладких процессов и функций.* Проблемы передачи информации. Т. 49, № 2. 2013
- Van der Vaart, A. *Information rates of nonparametric Gaussian process methods.* JMLR. V. 12, 2011

Когда использование данных разной точности лучше использования данных одной точности?

- Zhang, H. and Cai, W. *When doesn't cokriging outperform kriging?* Statistical Science. V. 30. No 2. 2015.

- *В одномерном случае* $d = 1$
- *Для квадратичной экспоненциальной функции*
- *Для выборки на бесконечной сетке*

отношение математических ожиданий ошибок при использовании только точных данных и точных и грубых данных

$$1 - \frac{1}{r^2},$$

где r *— коэффициент корреляции между грубой и точной функцией.*

Регрессия на основе гауссовских процессов

- Предположим, что функция $f(\mathbf{x})$ — реализация гауссовского процесса
- Пусть среднее равно нулю, а ковариационная функция

$$\operatorname{cov}(f(\mathbf{x}), f(\mathbf{x}')) = R(\mathbf{x}, \mathbf{x}').$$

- Тогда апостериорное распределение в точке $\mathbf{x}$ нормальное:

$$f(\mathbf{x})|S \sim \mathcal{N}(\widetilde{f}(\mathbf{x}), \widetilde{v}^2(\mathbf{x})).$$

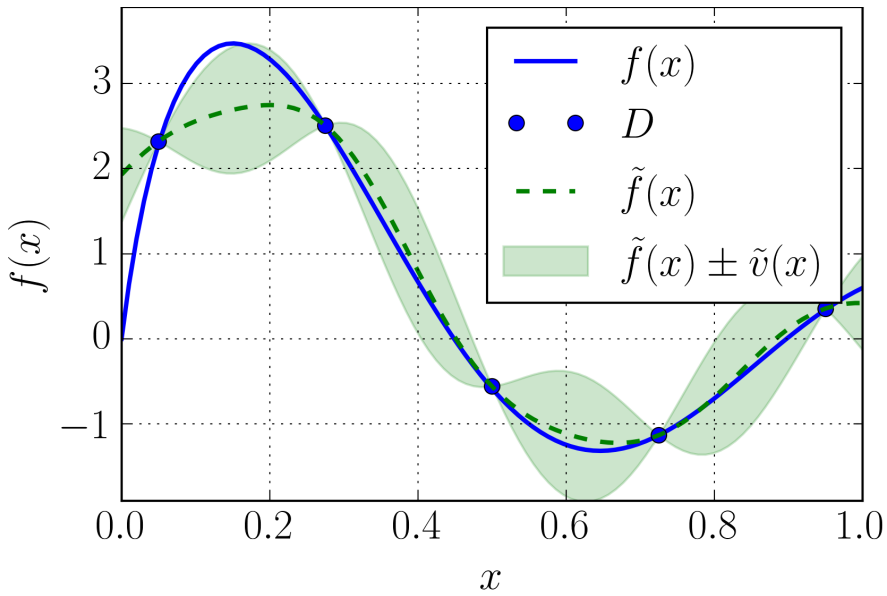


Figure: Регрессия на основе гауссовских процессов для $\mathbf{x} \in \mathbb{R}$

Наилучшая несмещенная линейная оценка

- Такая модель соответствует несмещенной оценке с минимальной дисперсией $\widetilde{f}(\mathbf{x})$ и записывается явно [Колмогоров, 1941]:

$$\widetilde{f}(\mathbf{x}) = \sum_{i=1}^{n} w_i f(\mathbf{x}_i) = \mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1}\mathbf{f},$$

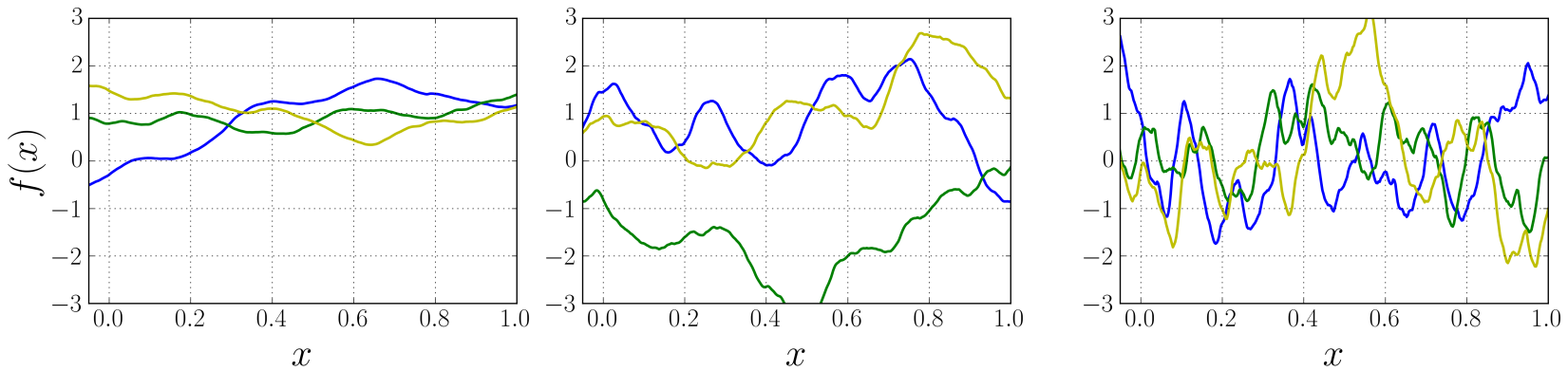
  где $\mathbf{r}(\mathbf{x}) = \{R(\mathbf{x}, \mathbf{x}_i)\}_{i=1}^n$, $\mathbf{R} = \{R(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$.
- Дисперсия оценки имеет вид:

$$\widetilde{v}^2(\mathbf{x}) = \mathbb{E}(\widetilde{f}(\mathbf{x}) - f(\mathbf{x}))^2 = R(\mathbf{x}, \mathbf{x}) - \mathbf{r}(\mathbf{x})^\top \mathbf{R}^{-1}\mathbf{r}(\mathbf{x}).$$

Реализации гауссовских процессов для ковариационной функции Матерна с $\nu = \frac{3}{2}$

Ковариационная функция Матерна для $\nu = \frac{3}{2}$:

$$R(x, x') = (1 + \sqrt{3}\theta|x - x'|)\exp(-\sqrt{3}\theta|x - x'|).$$

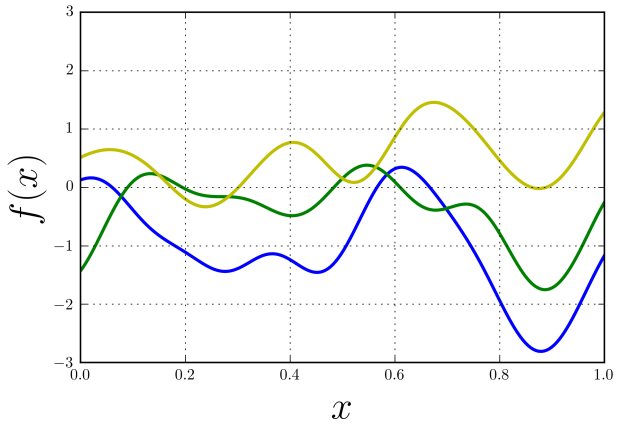


(a) $\theta = 2$ (b) $\theta = 6$ (c) $\theta = 20$

Модель регрессии на основе гауссовских процессов

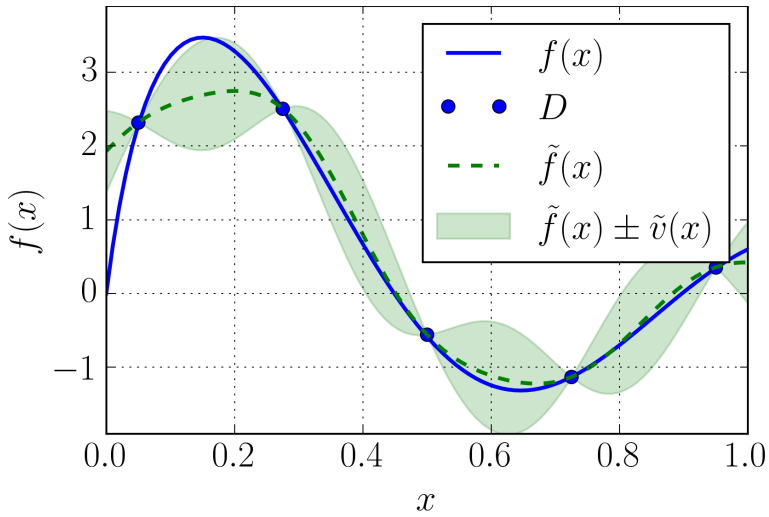
- $f(\mathbf{x})$ — реализация гауссовского процесса;
- гауссовский процесс стационарный: его ковариационная функция

$$\operatorname{cov}(f(\mathbf{x}), f(\mathbf{x}')) = R(\mathbf{x} - \mathbf{x}');$$



- спектральная плотность

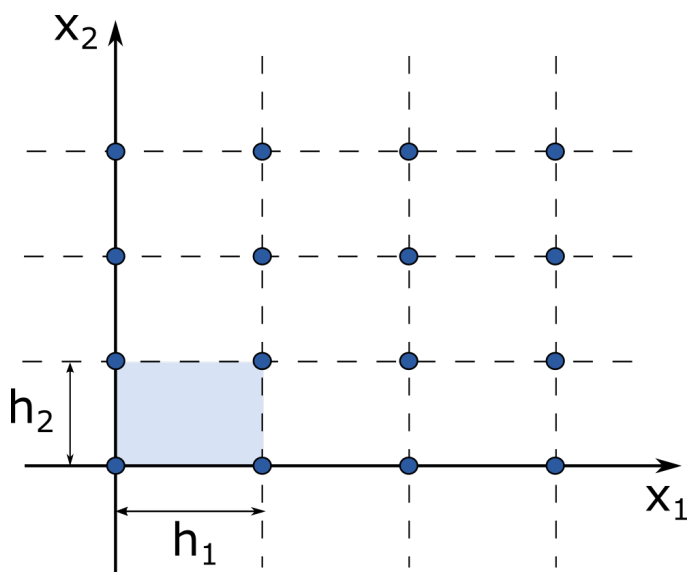
$$F(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} e^{2\pi \mathrm{i} \boldsymbol{\omega}^{\mathrm{T}} \mathbf{x}} R(\mathbf{x}) d\mathbf{x}.$$



Задача интерполяции

- Нам известны значения реализации $f(\cdot)$ в точках сетки $D_H = \{\mathbf{x} : \mathbf{x} = H\mathbf{k}, \mathbf{k} \in \mathbb{Z}^d\}$, $H = \mathrm{diag}(h_1, \ldots, h_d)$.
- Будем рассматривтаь $\widetilde{f}(\mathbf{x})$ вида:

$$\widetilde{f}(\mathbf{x}) = \mu(\Omega_H) \sum_{\mathbf{x}' \in D_H} K(\mathbf{x} - \mathbf{x}') f(\mathbf{x}'),$$



где $K(\mathbf{x} - \mathbf{x}')$ — симметричное ядро, $\Omega_H = [0, h_1] \times \ldots, [0, h_d]$, $\mu(\Omega_H) = \prod_{i=1}^{d} h_i$.

Ошибка интерполяции $f(\mathbf{x})$ аппроксимацией $\widetilde{f}(\mathbf{x})$:

$$\sigma_H^2(\widetilde{f}, F) = \frac{1}{\mu(\Omega_H)} \int_{\Omega_H} \mathbb{E}\left(\widetilde{f}(\mathbf{x}) - f(\mathbf{x})\right)^2 d\mathbf{x}.$$

Получено явное выражение для ошибки интерполяции для $d \geq 1$

Теорема

Для стационарного гауссовского процесса $f(\mathbf{x})$ со спектральной плотностью $F(\boldsymbol{\omega})$, и наблюдениями на D_H ошибка интерполяции имеет вид:

$$\sigma_H^2(\widetilde{f}, F) = \int_{\mathbb{R}^d} F(\boldsymbol{\omega}) \left[\left(1 - \widehat{K}(\boldsymbol{\omega})\right)^2 + \sum_{\substack{\mathbf{k} \in D_{H-1} \\ \mathbf{k} \neq \mathbf{0}}} \widehat{K}^2 (\boldsymbol{\omega} + \mathbf{k}) \right] d\boldsymbol{\omega}.$$

Для интерполяции минимизирующей $\sigma_H^2(\widetilde{f}, F)$ $K(\mathbf{x})$ такое, что его преобразование Фурье $\widehat{K}(\boldsymbol{\omega})$ имеет вид:

$$\widehat{K}(\boldsymbol{\omega}) = \frac{F(\boldsymbol{\omega})}{\sum_{\mathbf{k} \in D_{H-1} \setminus \{\mathbf{0}\}} F(\boldsymbol{\omega} + \mathbf{k})}.$$

Теперь можем считать ошибку интерполяции для конкретных ковариационных функций

Следствие

Для гауссовского процесса на $\mathbb{R}$

- *с экспоненциальной ковариацинной функцией* $R_\theta(x) = \sqrt{\frac{\pi}{2}}\exp\left(-\theta|x|\right)$ $(F_\theta(\omega) = \frac{\theta}{\theta^2+\omega^2})$*:*

$$\sigma_h^2(\widetilde{f}, F_\theta) \approx \frac{2}{3}\pi^2\theta h + O((\theta h)^2),\ \theta h \to 0\,.$$

- *с квадратичной экспоненциальной ковариационной функцией* $R_\theta(x) = \sqrt{2\pi}\exp\left(-2\pi^2\theta x^2\right)$ $\left(F_\theta(\omega) = \frac{1}{\sqrt{\theta}}\exp\left(-\frac{\omega^2}{2\theta}\right)\right)$ *для* $\theta h^2 \to 0$*:*

$$\frac{4}{3}\sqrt{\theta}h\exp\left(-\frac{1}{8\theta h^2}\right) \le \sigma_h^2(\widetilde{f}, F_\theta) \le 7\sqrt{\theta}h\exp\left(-\frac{1}{8\theta h^2}\right).$$

Минимаксная ошибка интерполяции характеризует качество интерполяции достаточно гладких функций

- Мы не знаем ковариационную функцию.
- Рассмотрим множество спектральных плотностей $F(\boldsymbol{\omega})$ гауссовских полей $\mathcal{F}(L, \boldsymbol{\lambda})$, $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_d\}$:

$$\mathcal{F}(L, \boldsymbol{\lambda}) = \left\{ F : \mathbb{E} \sum_{i=1}^{d} \lambda_i^2 \left(\frac{\partial f_F(\mathbf{x})}{\partial x_i} \right)^2 \leq L, \mathbf{x} \in \mathbb{R}^d \right\},$$

где $f(\mathbf{x}) = f_F(\mathbf{x})$ — гауссовский случайный процесс со спектральной плотностью F.

- Мы хотим найти **минимаксную ошибку интерполяции** $R^h(L)$ **в многомерном случае**

$$R^H(L, \boldsymbol{\lambda}) = \inf_{\widetilde{f}} \sup_{F \in \mathcal{F}(L, \boldsymbol{\lambda})} \sigma_H^2(\widetilde{f}, F).$$

Можно явно выписать минимаксную ошибку интерполяции

Теорема

Для множества гауссовских полей $\mathcal{F}(L, \boldsymbol{\lambda})$ на $\mathbb{R}^d$ минимаксная ошибка интерполяции имеет вид

$$R^H(L, \Lambda) = \frac{L}{2\pi^2} \max_{i \in \{1,\dots,d\}} \left(\frac{h_i}{\lambda_i}\right)^2.$$

При этом минимаксная интерполяция имеет вид $\widetilde{f}(\mathbf{x}) = \mu(\Omega_H) \sum_{\mathbf{x}' \in D_H} K(\mathbf{x} - \mathbf{x}') f(\mathbf{x}')$, где $K(\mathbf{x})$ — симметричное ядро, преобразование Фурье которого $\widehat{K}(\boldsymbol{\omega})$ имеет явный вид:

$$\widehat{K}(\boldsymbol{\omega}) = \begin{cases} 1 - \sqrt{\sum_{i=1}^d h_i^2 \omega_i^2}, & \sum_{i=1}^d h_i^2 \omega_i^2 \le 1, \\ 0, & \sum_{i=1}^d h_i^2 \omega_i^2 > 1. \end{cases}$$

Доказательство. Схема

- Мы получим оценку сверху и снизу вида $\frac{L}{2\pi^2}\max_{i\in\{1,\dots,d\}}\left(\frac{h_i}{\lambda_i}\right)^2$ на $R^H(L,\boldsymbol{\lambda})$.
- Введем

$$\Phi(F,\widehat{K})=\int_{\mathbb{R}^d}F(\boldsymbol{\omega})\left[(1-\widehat{K}(\boldsymbol{\omega}))^2+\sum_{\mathbf{x}\in D_{H-1}\setminus\{\mathbf{0}\}}\widehat{K}^2(\boldsymbol{\omega}+\mathbf{x})\right]d\boldsymbol{\omega},$$

- Для $F\in\mathcal{F}(L,\boldsymbol{\lambda})$, некоторого $\widehat{K}$:

$$\min_{\widehat{K}}\Phi(F,\widehat{K})\le R^H(L,\boldsymbol{\lambda})\le\max_{F\in\mathcal{F}(L,\boldsymbol{\lambda})}\Phi(F,\widehat{K}).$$

Доказательство. Оценка снизу

- Из доказанной выше теоремы:

$$\min_{\widehat{K}} \Phi(F, \widehat{K}) = \int_{\mathbb{R}^d} F(\boldsymbol{\omega}) \frac{\sum_{\mathbf{x} \in D_{H-1} \setminus \{\mathbf{0}\}} F(\boldsymbol{\omega} + \mathbf{x})}{\sum_{\mathbf{x} \in D_{H-1}} F(\boldsymbol{\omega} + \mathbf{x})} d\boldsymbol{\omega}.$$

- Рассмотрим

$$F_\varepsilon(\boldsymbol{\omega}) = \begin{cases} \frac{A_\varepsilon}{(2\varepsilon)^d}, & \exists \mathbf{s} \in U_h : \|\boldsymbol{\omega} - \mathbf{s}\|_\infty \le \varepsilon, \\ 0, & \text{иначе,} \end{cases}$$

$U_h = \left\{ \left(0, 0, \ldots, \frac{1}{2h_j}, \ldots, 0\right), \left(0, 0, \ldots, -\frac{1}{2h_j}, \ldots, 0\right) \right\}$, A_ε выбирается исходя из того, что $F_\varepsilon \in \mathcal{F}(L, \boldsymbol{\lambda})$:

$$(2\pi)^2 \int_{\mathbb{R}^d} F(\boldsymbol{\omega}) \sum_{i=1}^{d} \lambda_i^2 \omega_i^2 d\boldsymbol{\omega} = L.$$

- При $\varepsilon \to 0$ выполнено, что

$\min_{\widehat{K}} \Phi(F, \widehat{K}) \to \frac{L}{2\pi^2} \max_{i \in \overline{1,d}} \left(\frac{h_i}{\lambda_i}\right)^2$.

Доказательство. Оценка сверху

- Для произвольного $\widehat{K}(\boldsymbol{\omega})$ выполнено

$$R^H(L,\boldsymbol{\lambda}) \le \max_{F\in\mathcal{F}(L,\boldsymbol{\lambda})} \Phi(F,\widehat{K}) \le L\left(\frac{1}{2\pi}\right)^2 \times$$

$$\times \max_{\boldsymbol{\omega}} \left\{ \frac{1}{\sum_{i=1}^d \lambda_i^2\omega_i^2} \left[(1-\widehat{K}(\boldsymbol{\omega}))^2 + \sum_{\mathbf{x}\in D_{H-1}\setminus\{\mathbf{0}\}} \widehat{K}^2(\boldsymbol{\omega}+\mathbf{x}) \right] \right\}.$$

- Покажем, что для

$$\widetilde{K}(\boldsymbol{\omega}) = \begin{cases} 1-\|\boldsymbol{\omega}\|, \|\boldsymbol{\omega}\|^2 \le 1, \\ 0, \text{иначе.} \end{cases}$$

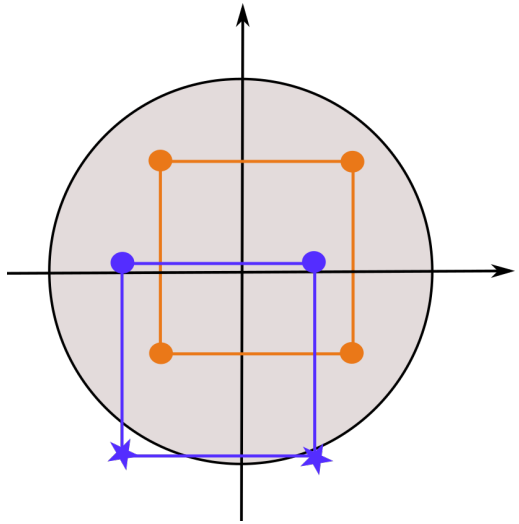
выполнено, что

$$\left[(1-\widetilde{K}(\boldsymbol{\omega}))^2 + \sum_{\mathbf{x}\in\mathbb{Z}^d\setminus\{\mathbf{0}\}} \widetilde{K}^2(\boldsymbol{\omega}+\mathbf{x}) \right] \le 2\|\boldsymbol{\omega}\|^2.$$

Доказательство. Оценка сверху

- Нужно показать, что

$$\sum_{\substack{\mathbf{x}\in\mathbb{Z}^d\setminus\{\mathbf{0}\},\\ \|\boldsymbol{\omega}+\mathbf{x}\|\le 1}} (1-\|\boldsymbol{\omega}+\mathbf{x}\|)^2 \le \|\boldsymbol{\omega}\|^2.$$



- Выполнено, что для $c \ge 0$ и $\omega \ge 0$, таких что $c^2+\omega^2 \le 1$, $c^2+(1-\omega^2) \le 1$:

$$\left(1-\sqrt{c^2+(1-\omega)^2}\right)^2 + \left(1-\sqrt{c^2+\omega^2}\right)^2 \le (1-c)^2.$$

- Используя это утверждение, получаем по индукции по d основной результат.

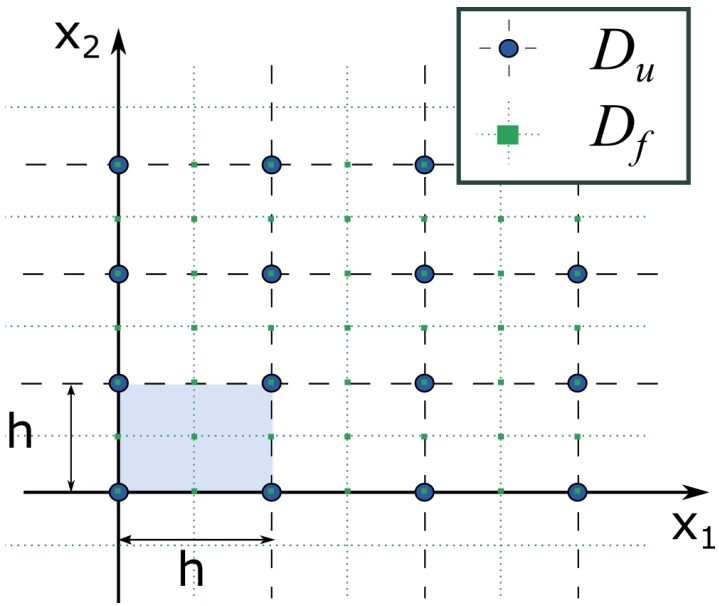
Модель данных разной точности

- Мы моделируем точную функцию как гауссовский процесс [Кеннеди, 2001]

$$u(\mathbf{x}) = \rho f(\mathbf{x}) + g(\mathbf{x}),$$

где ρ — известный коэффициент.

- $f(\mathbf{x})$ и $g(\mathbf{x})$ — два независимых гауссовских процесса со спектральными плотностями $F(\boldsymbol{\omega})$ и $G(\boldsymbol{\omega})$.
- $f(\mathbf{x})$ — модель источника грубых данных.



- $D_u = D_{hI} = D_h$,
- $D_f = D_{\frac{h}{m}I}$, $m \in \mathbb{Z}^+$.

Ошибка интерполяции для $u(\mathbf{x})$ — взвешенная сумма ошибок интерполяции для $f(\mathbf{x})$ и $g(\mathbf{x})$

Теорема

Ошибка интерполяции

$$\sigma^2_{h,m}(\widetilde{u}, F, G, \rho) = \frac{1}{\mu(\Omega_{hI})} \int_{\Omega_{hI}} \mathbb{E}\left[\widetilde{u}(\mathbf{x}) - u(\mathbf{x})\right]^2 d\mathbf{x}.$$

Минимум ошибки интерполяции имеет вид:

$$\sigma^2_{h,m}(\widetilde{u}, F, G, \rho) = \sigma^2_{hI}(\widetilde{g}, G) + \rho^2 \sigma^2_{\frac{h}{m}I}(\widetilde{f}, F),$$

где $\widetilde{f}$, $\widetilde{g}$ *минимизируют* $\sigma^2_{\frac{h}{m}I}(\widetilde{f}, F)$ *и* $\sigma^2_{hI}(\widetilde{g}, G)$ *соответственно.*

Минимаксная ошибка интерполяции для $u(\mathbf{x})$

Теорема

Будем считать, что спектральная плотность гауссовских процессов $f(\mathbf{x})$ и $g(\mathbf{x})$ неизвестна, но выполнено:

$$\mathbb{E}\sum_{i=1}^{d}\left[\frac{\partial f(\mathbf{x})}{\partial x_i}\right]^2 \leq L_f,\ \mathbb{E}\sum_{i=1}^{d}\left[\frac{\partial g(\mathbf{x})}{\partial x_i}\right]^2 \leq L_g.$$

Тогда минимаксная ошибка интерполяции

$$R^{h,m}(L_f, L_g) = \inf_{f,g}\sup_{F\in\mathcal{F}(L_f),G\in\mathcal{F}(L_g)}\sigma^2_{h,m}(\widetilde{u}, F, G, \rho)\,.$$

для $u(\mathbf{x}) = \rho f(\mathbf{x}) + g(\mathbf{x})$ и наблюдений $u(\mathbf{x})$ в D_h и $f(\mathbf{x})$ в $D_{\frac{h}{m}}$:

$$R^{h,m}(L_f, L_g) = \rho^2\frac{L_f}{2}\left(\frac{h}{m\pi}\right)^2 + \frac{L_g}{2}\left(\frac{h}{\pi}\right)^2.$$

Задача выбора оптимального соотношения между размерами выборок

Мы решаем задачу в следующей постановке:

- стоимость вычисления грубой функции 1, а точной функции – в c раз больше;
- стоимость вычисления значения функции во всех точках выборки пропорциональна $1/h^d$ — количеству точек в гиперкубе единичного объема;
- общий бюджет равен Λ.

В таких предположениях

$$\Lambda = c\frac{1}{h^d} + \delta\frac{1}{h^d},$$

где $\delta = m^d$ — отношение между размерами выборок.

Оптимальное соотношение между размерами выборок имеет явный вид

Теорема

Минимум минимаксной ошибки интерполяции по δ для бюджета $\Lambda = c\frac{1}{h^d} + \delta\frac{1}{h^d}$ имеет вид

$$R^{h,\delta^{\frac{1}{d}}}(L_f, L_g) = \rho^2 \frac{L_f}{2}\left(\frac{c+\delta^*}{\pi\Lambda\delta^*}\right)^{\frac{2}{d}} + \frac{L_g}{2}\left(\frac{c+\delta^*}{\pi\Lambda}\right)^{\frac{2}{d}},$$

причем оптимальное соотношение между размерами выборок

$$\delta^* = \left(\frac{L_f}{L_g} c\rho^2\right)^{\frac{d}{d+2}}.$$

Отношение минимаксных ошибок интерполяции $\frac{R_2}{R_1}$ можно выписать явно

- $R_2 = R^{h,\mu^*}(L_f, L_g, \rho)$ — ошибка интерполяции при использовании данных разной точности;
- $R_1 = R^{h}(L_f, L_g, \rho)$ — ошибка интерполяции при использовании данных одной точности.
- Тогда

$$\frac{R_2}{R_1} = \frac{\left(1 + \left(\frac{L_f^d \rho^{2d}}{L_g^d c^2}\right)^{\frac{1}{d+2}}\right)^{\frac{d+2}{d}}}{1 + \rho^2 \frac{L_f}{L_g}}.$$

При $r \to 0$ становится невыгодным использовать данные разной точности, при $r \to 1$ $\frac{R_2}{R_1} \approx \frac{1}{c^{\frac{2}{d}}}$

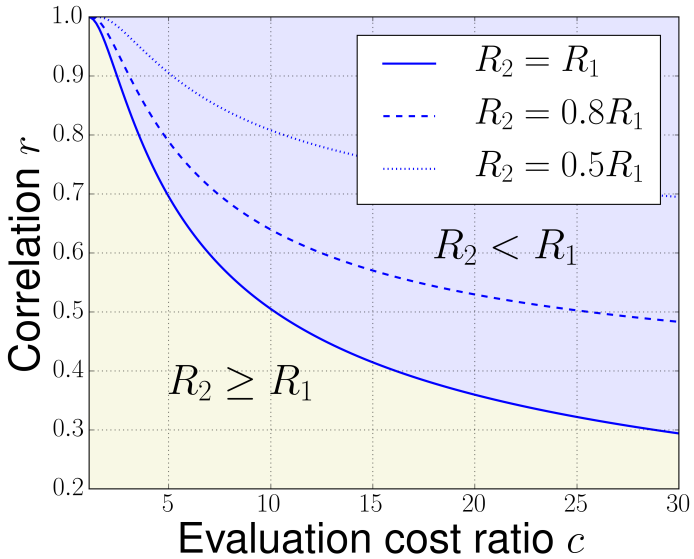
- Пусть $\mathbb{E}f^2(\mathbf{x}) = V_f$, $\mathbb{E}g^2(\mathbf{x}) = V_g$.
- Тогда коэффициент корреляции

$$\operatorname{corr}(u(\mathbf{x}), f(\mathbf{x})) = r = \frac{1}{\sqrt{1 + \frac{V_g}{V_f}\frac{1}{\rho^2}}}.$$

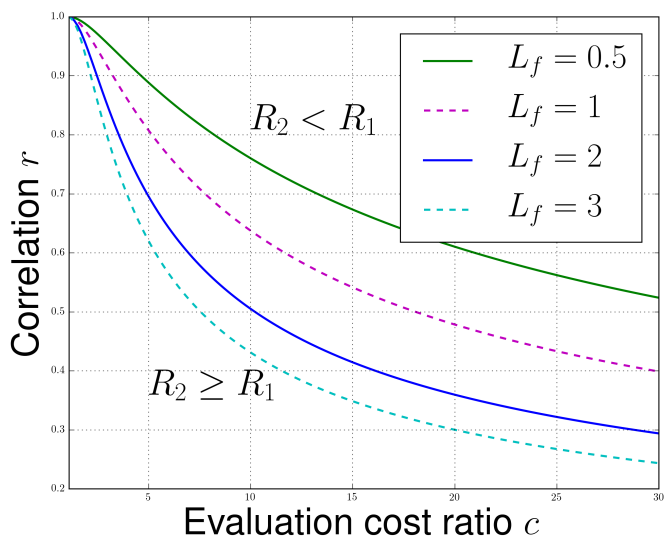
$$r \to 0 : \frac{R_2}{R_1} \approx 1 + \frac{d+2}{d}\left(\frac{L_f V_f}{L_g V_g}\right)^{\frac{d}{d+2}} \frac{r^{\frac{2d}{d+2}}}{c^{\frac{2}{d+2}}},$$

$$r \to 1 : \frac{R_2}{R_1} \approx \frac{1}{c^{\frac{2}{d}}} + \frac{2+d}{d}\left(\frac{L_g V_f}{L_f V_g}\right)^{\frac{d}{d+2}} \frac{(1-r^2)^{\frac{d}{d+2}}}{c^{\frac{4}{d(d+2)}}}.$$

Можно выделить области, в которых использование данных разной точности имеет смысл



(a) Кривые $R_2 = kR_1$ для $L_f = 2$



(b) Кривые $R_2 = R_1$ для разных L_f

Figure: Область, в которой R_2 меньше R_1 для $d = 1$, $L_g = 1$

Алгоритм выбора плана эксперимента для данных разной точности

Вход: коэффициент корреляции r между функциями разной точности, бюджет Λ, относительная стоимость вычислений точной функции c.

1. Оценить $\rho^2 = \frac{1}{1-\frac{1}{r^2}}$.
2. Оценить $\delta^* = \left(c\rho^2\right)^{\frac{d}{d+2}}$.
3. Получить выборки точных данных размера $n_u = \frac{\Lambda}{c+\delta^*}$ и грубых данных размера $n_f = \frac{\Lambda\delta^*}{c+\delta^*}$, такие что $D_u \subseteq D_f$.

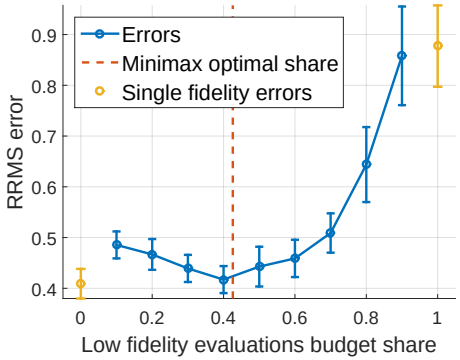
Тестирование

- Для модели $\widetilde{u}(\mathbf{x})$ и тестовой выборки $S_* = \{\mathbf{x}_i^*, u_i^* = u(\mathbf{x}_i^*)\}_{i=1}^{n_t}$ ошибка RRMS есть:

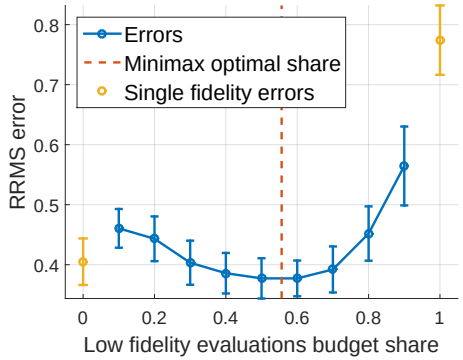
$$\mathrm{RRMS} = \sqrt{\frac{\sum_{i=1}^{n_t}(u_i^* - \widetilde{u}(\mathbf{x}_i^*))^2}{\sum_{i=1}^{n_t}(u_i^* - \overline{u})^2}},$$

  где $\overline{u} = \frac{1}{n_t}\sum_{i=1}^{n_t} u_i^*$.
- Для оценки ошибки используется скользящий контроль.
- Искусственные данные — реализации гауссовских процессов с заданной ковариационной функцией.
- Реальные данные взяты из аэрокосмической отрасли, астрономии и метаобучения.

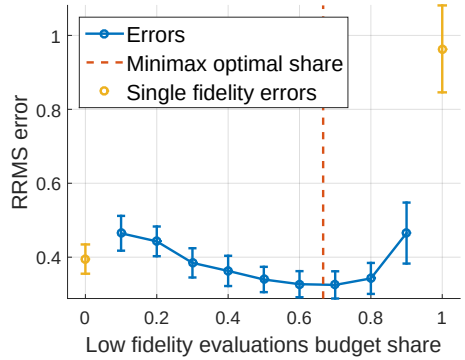
Теоретические оптимальное значение близко к оптимальному на искусственных данных



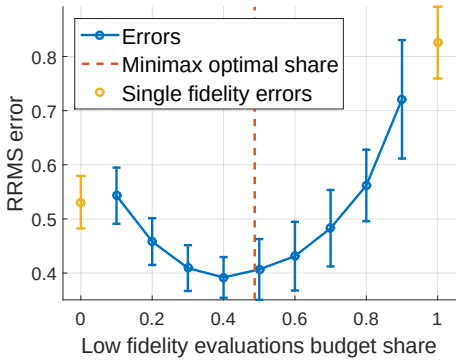
(a) $c = 5, r = 0.8$



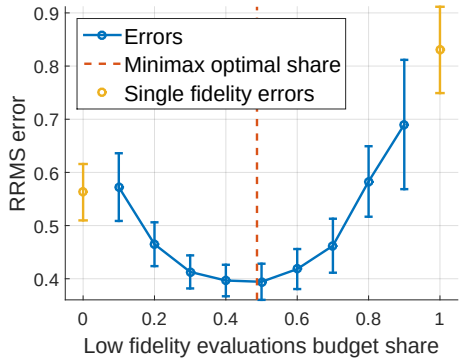
(b) $c = 5, r = 0.9$



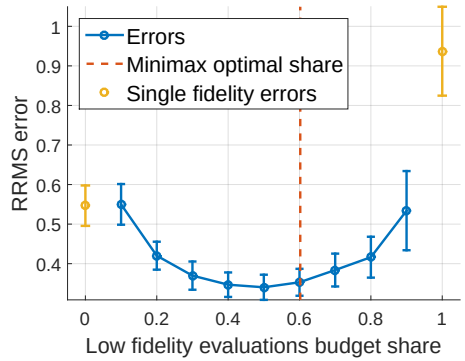
(c) $c = 5, r = 0.95$



(d) $c = 10, r = 0.8$



(e) $c = 10, r = 0.9$



(f) $c = 10, r = 0.95$

Figure: Зависимость ошибки RRMS от доли бюджета, использованного для источника данных низкой точности, $d = 3$

Мы сравниваем наш подход с набором эвристик

- High — используем только данные высокой точности;
- EqSize — $n_u = n_f$;
- EqBudget — $cn_u = n_f$;.
- MinMinimax — подход, предложенный в работе.
- Low — используем только данные низкой точности.

Результаты работы на искусственных данных

δ^*	d	High	EqSize	EqBudget	MinMinimax	Low
$2c$	1	0.0575	0.0936	**0.0196**	**0.0222**	1.8481
	2	0.2551	0.3202	**0.1289**	**0.1440**	1.2279
	3	0.4830	0.5503	**0.3010**	**0.3073**	0.9639
$\frac{1}{2}c$	1	0.0563	0.0801	**0.0469**	**0.0402**	1.5205
	2	**0.2574**	0.3437	0.2795	**0.2452**	0.8977
	3	**0.4691**	0.5392	0.4784	**0.4499**	0.7727
$\frac{1}{3}c$	1	**0.0484**	0.0797	0.0653	**0.0429**	1.2927
	2	**0.2631**	0.3300	0.3707	**0.2932**	0.9894
	3	**0.4978**	0.5614	0.5846	**0.5192**	0.9183

Table: Относительные среднеквадратичные ошибки, усреднение по 50 запускам

Результаты работы на реальных данных

Задача	High	EqSize	EqBudget	MinMinimax	Low
cPress12	0.5599	0.6019	0.3580	**0.2779**	**0.2843**
cPress13	0.5596	0.5759	**0.3861**	**0.3481**	0.5435
Euler	**0.7674**	0.8925	0.8462	**0.7420**	0.9139
Airfoil	0.5462	0.5946	0.5390	**0.5221**	**0.4852**
Disk1	0.2999	0.3400	**0.1922**	**0.1922**	**0.1638**
Disk2	0.4460	0.4570	**0.2998**	**0.2998**	**0.2723**
SVM	**0.1487**	**0.1492**	0.1849	0.1642	0.6081
Supernova	**0.0395**	0.0484	**0.0180**	0.0575	0.0575

Table: Ошибки RRMS, усреднение по 20 запускам, $c = 5$, бюджет 300

Другие задачи

- Неправильная спецификация модели;
- Как влияет на ответ шум в данных;
- Улучшение алгоритма выбора отношения размеров выборок;
- Создание репрезентативной базы данных разной точности.

Выводы

- Минимаксная ошибка интерполяции для регрессии на основе гауссовских процессов порядка h^2.
- Можно явно получить отношение размеров выборок разной точности, доставляющее минимум минимаксной ошибки интерполяции.
- Такое отношение может быть использовано для выбора плана эксперимента в реальных задачах.

Спасибо за внимание!