Wprowadzenie do optymalizacji Bayesowskiej

Marcin Lisowski

ML Gdańsk, 14 października 2019

Zawartość prezentacji

Motywacja: Założenia

Niech

$$y := f(\boldsymbol{x}), \text{ gdzie}$$

- $\boldsymbol{x} \in \mathcal{X}$ — wektor parametrów eksperymentu; zbiór $\mathcal{X}$ jest ograniczony,
- $y \in f(\mathcal{X})$ — wynik ewaluacji,
- f — nieznana funkcja celu, której ewaluacja jest kosztowna.

Motywacja: Problem optymalizacji

Szukamy $\boldsymbol{x}^*$ globalnie minimalizującego $f(\boldsymbol{x})$:

$$\boldsymbol{x}^* \leftarrow \arg\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$$

Główne problemy

- Przeszukiwanie siatkowe/losowe przestrzeni parametrów jest zbyt kosztowne
- Brak możliwości wykorzystania typowych metod optymalizacji (brak gradientu, jakobianu, hessianu, wypukłości/wklęsłości, ciągłości etc. funkcji celu)
- Funkcja celu może mieć komponent stochastyczny

Motywacja: Przykładowe zagadnienia

- głębokie uczenie
- robotyka,
- sekwencyjne projektowanie,
- konfiguracja algorytmów,
- sieci czujników,
- uczenie przez wzmocnienie,
- planowanie,
- fizyka eksperymentalna,
- modelowanie białek,
- etc.

Optymalizacja Bayesowska: Ask/Tell

Badacz:
„Dotychczas zaobserwowałem $\boldsymbol{x}_{1:n}, y_{1:n}$,
co robić dalej?"

Optymalizator:
„Następnym razem spróbuj $\boldsymbol{x}_{n+1}$"

Optymalizacja Bayesowska: Optymalizator

Model zastępczy $\mathcal{M}$

Pozwala na określenie (gęstości) rozkładu prawdopodobieństwa warunkowego wartości funckji celu w nowym punkcie $\boldsymbol{x}_{n+1}$ na podstawie dotychczasowych obserwacji $\mathcal{D} = \boldsymbol{x}_{1:n}, y_{1:n}$:

$$y_{n+1} | \boldsymbol{x}_{n+1} \sim \mathcal{M}(\mathcal{D}, \boldsymbol{\theta})$$

Funkcja akwizycji α; exploracja kontra exploatacja

Maksimum funkcji akwizycji wskazuje następny wektor wejściowy funkcji celu w oparciu o dotychczasowe obserwacje i model $\mathcal{M}$:

$$\alpha : \mathcal{X} \to \mathbb{R}, \quad \boldsymbol{x}_{n+1} := \arg\max_{\boldsymbol{x} \in \mathcal{X}} \{\alpha\left(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta}\right)\}$$

Artykuł przeglądowy: [SSW[+]16]

Przykłady: Proces Gaussowski

Popularnym modelem zastępczym $\mathcal{M}$ jest *Proces Gaussowski*

Proces Gaussowski jest to kolekcja zmiennych losowych, których dowolny skończony podzbiór ma wspólny rozkład Gaussa.

$$f(\boldsymbol{x}) \sim \mathcal{GP}\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right),$$

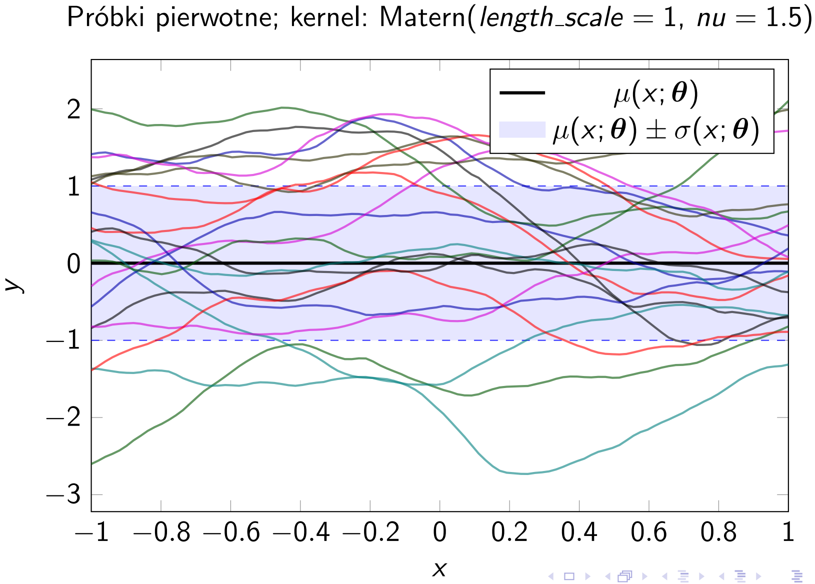
gdzie

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})] \text{ — funkcja wartości średniej,}$$

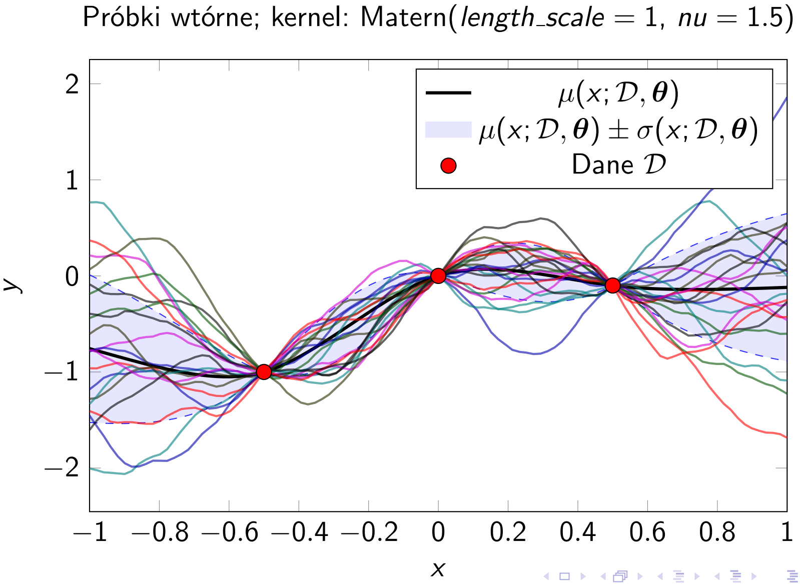
$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))] \text{ — funkcja kowariancji.}$$

Monografia na temat regresji i klasyfikacji przy pomocy procesu Gaussowskiego: [RW05]

Przykłady: Proces Gaussowski



Przykłady: Proces Gaussowski



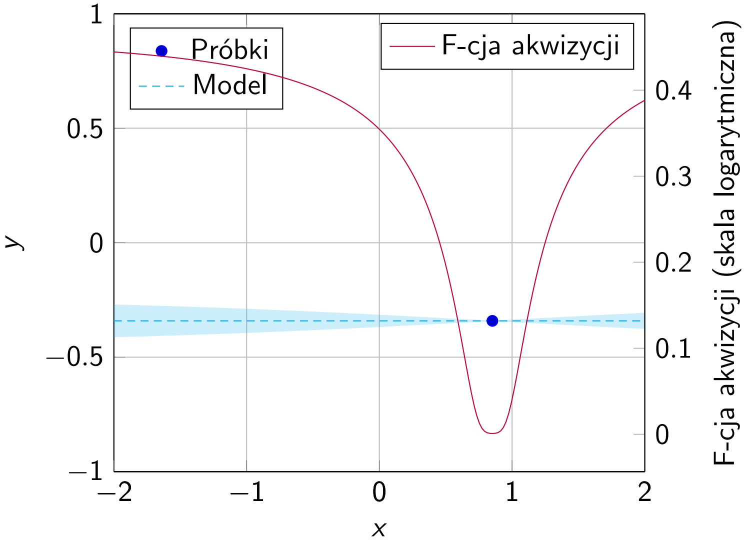
Przykłady: Prawdopodobieństwo poprawy

Przykładem prostej funkcji akwizycji jest *prawdopodobieństwo poprawy* α_{PI} (Probability of Improvement, [Kus64]):

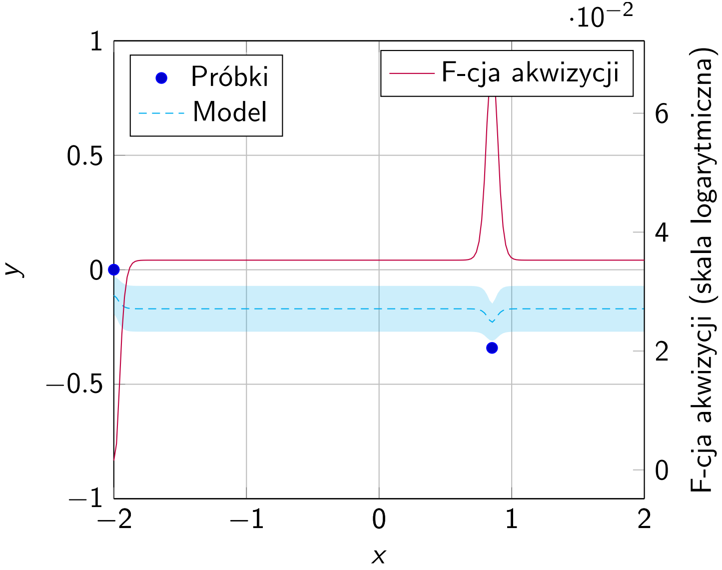
$$\alpha_{PI}(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta}) = \Phi\left(\frac{\min\{y_{1:n}\} - \mu(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta}) - \psi}{\sigma(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta})}\right), \text{ gdzie}$$

$\Phi(x)$ — dystrybuanta rozkładu normalnego,
ψ — współczynnik eksploatacji/eksploracji

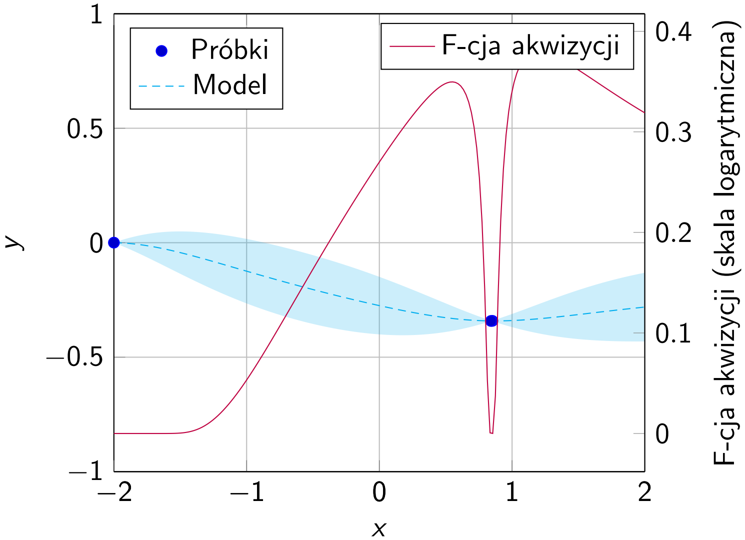
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



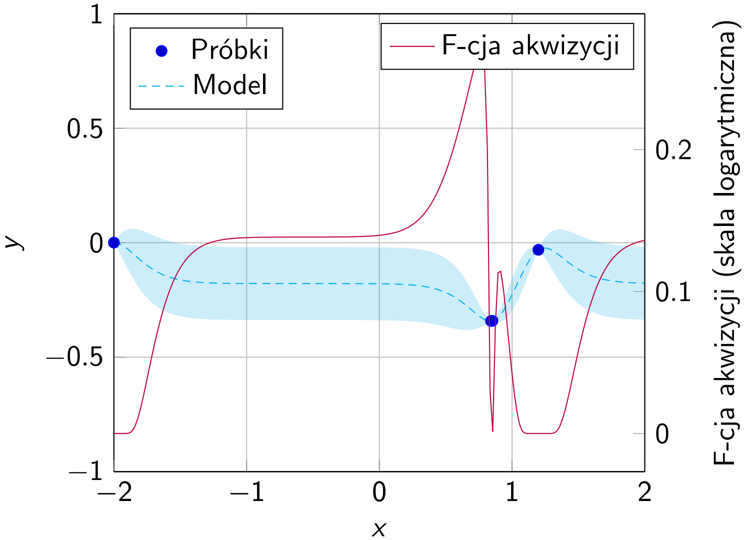
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



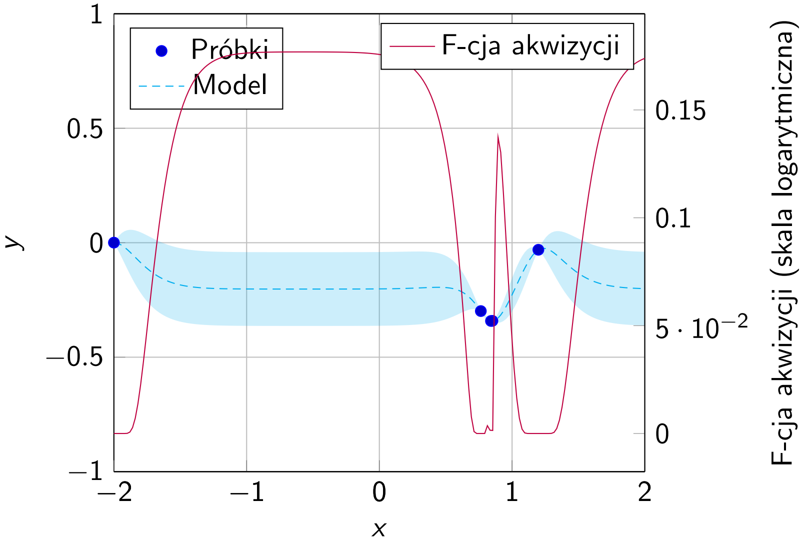
Przykłady: 1. Znana funckja celu $f : [-2, 2] \rightarrow [-1, 1]$



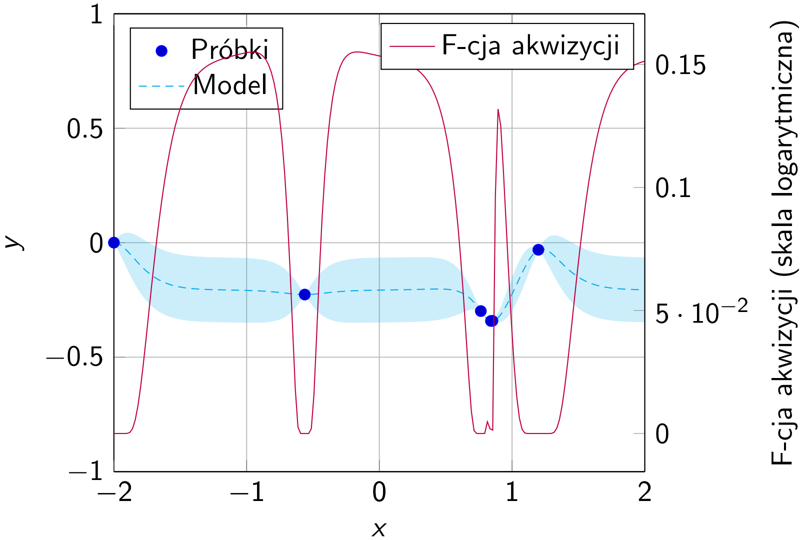
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



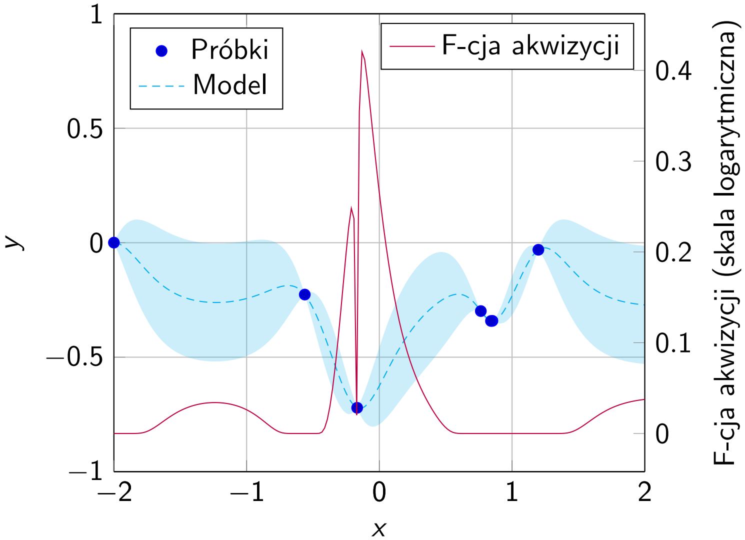
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



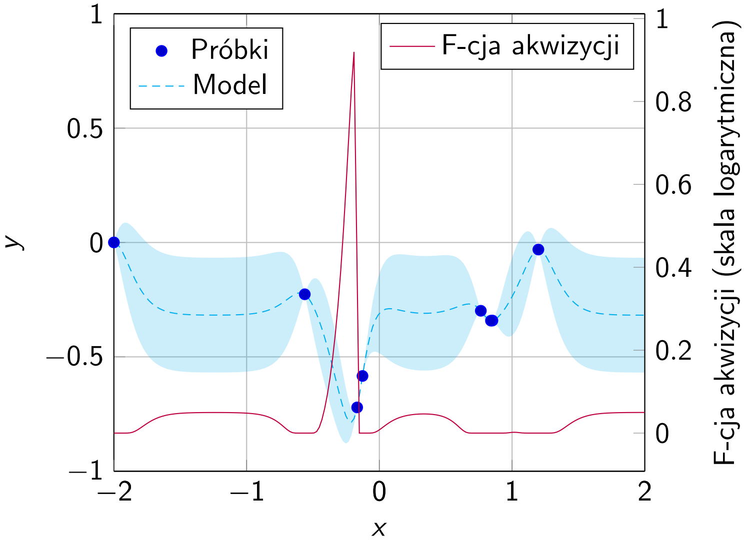
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



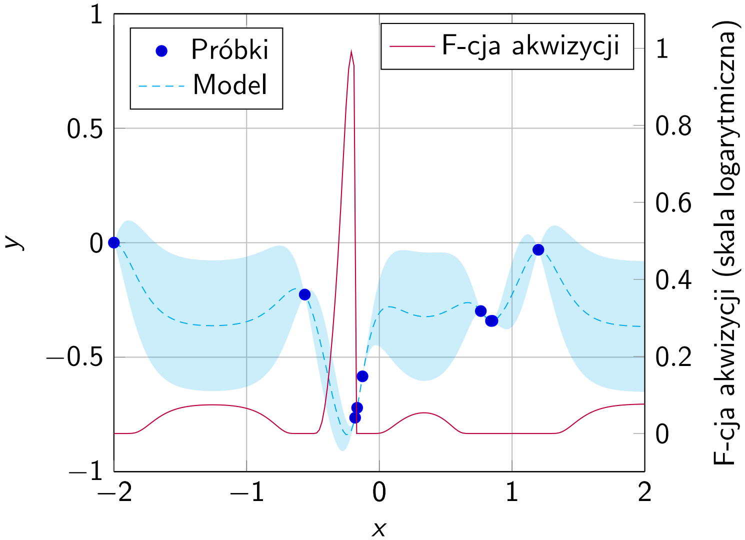
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



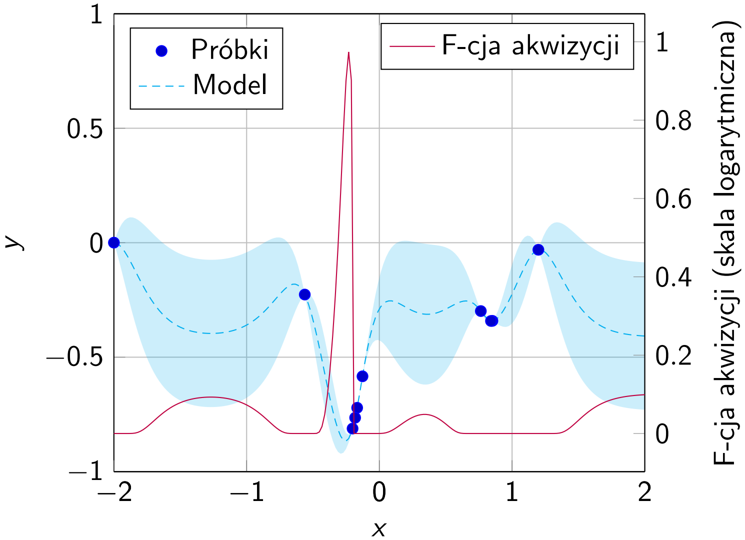
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



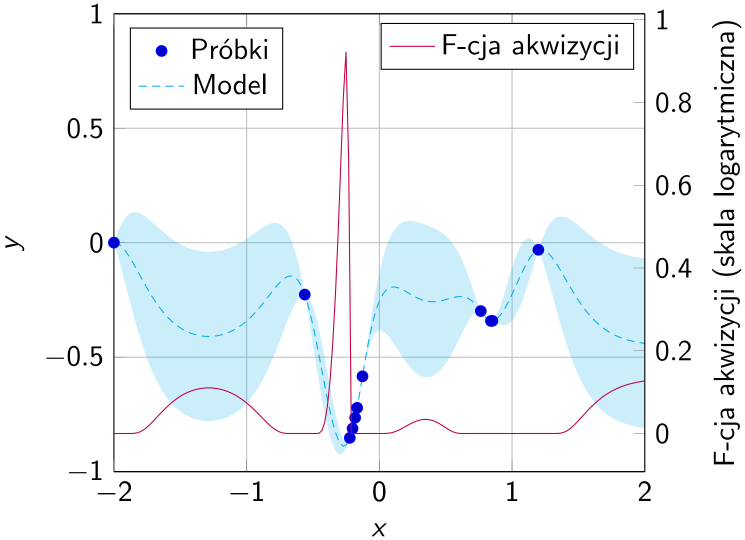
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



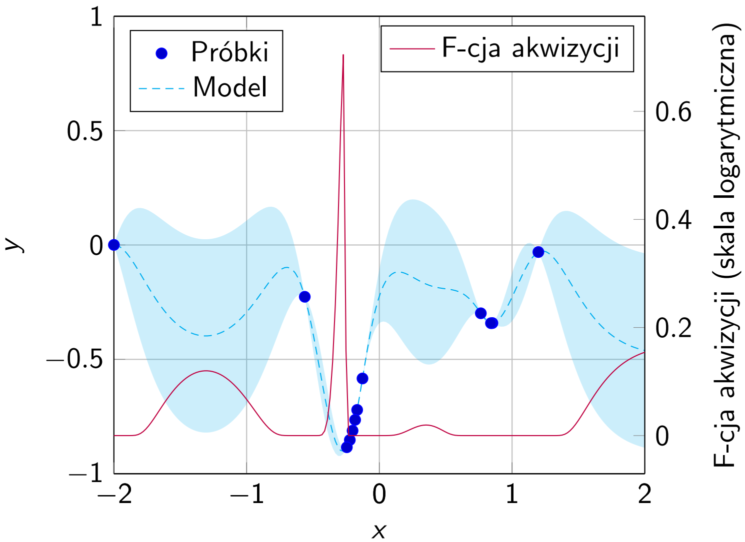
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



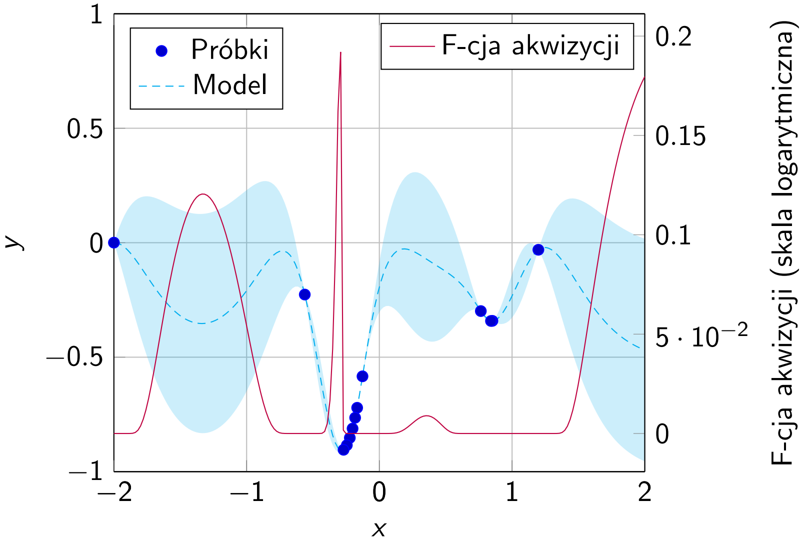
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



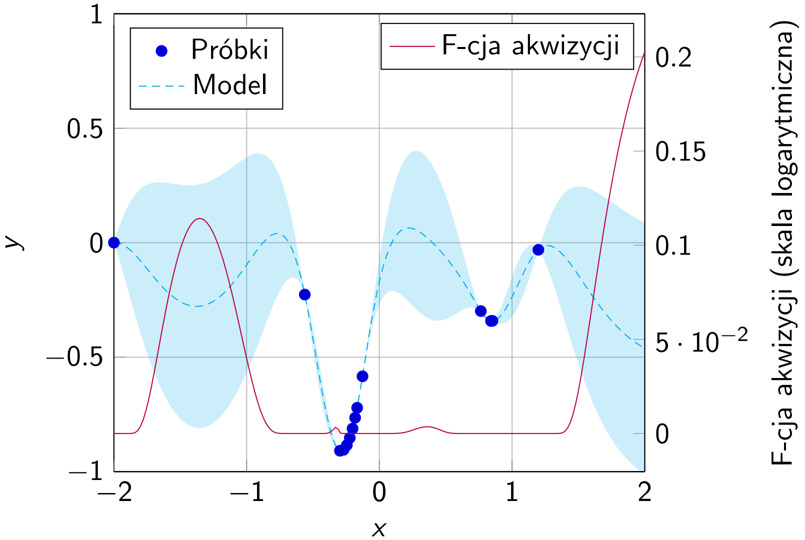
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



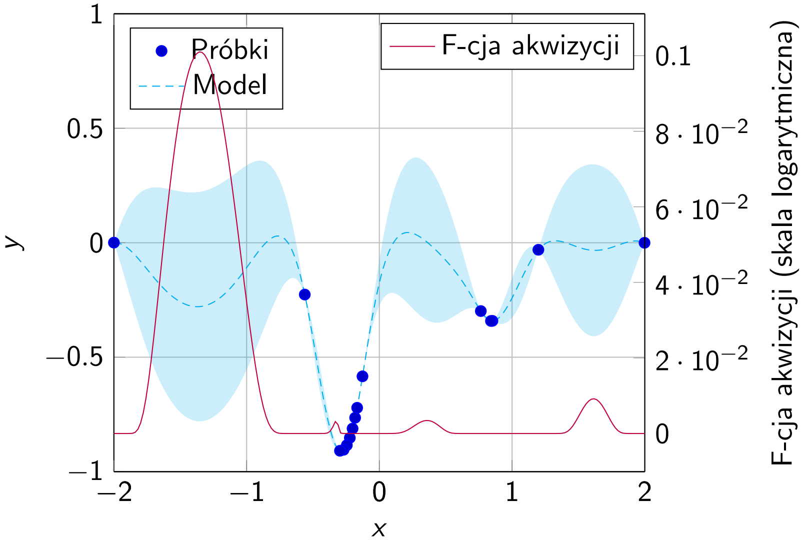
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



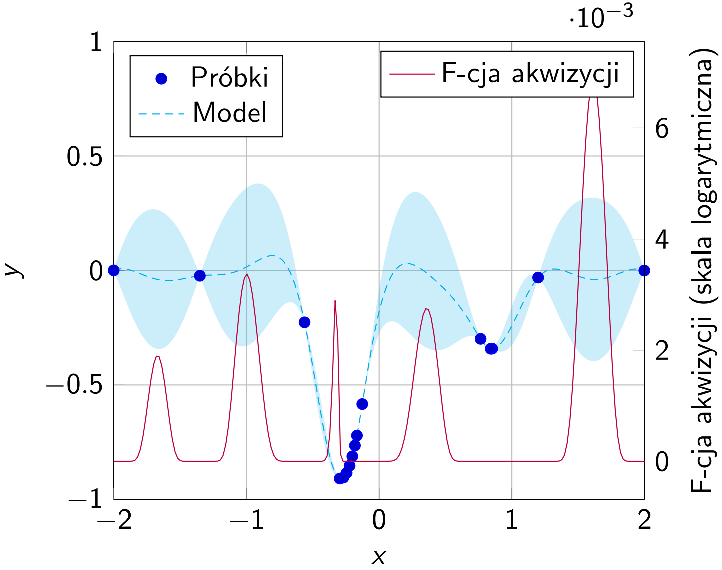
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



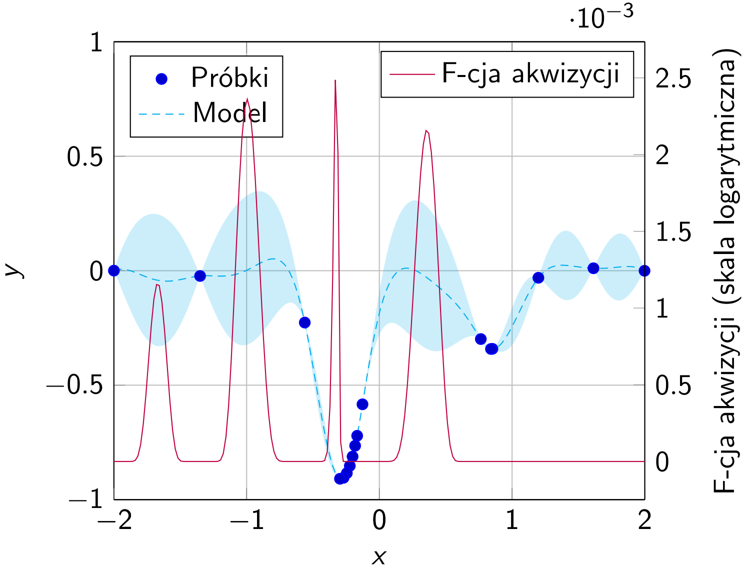
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



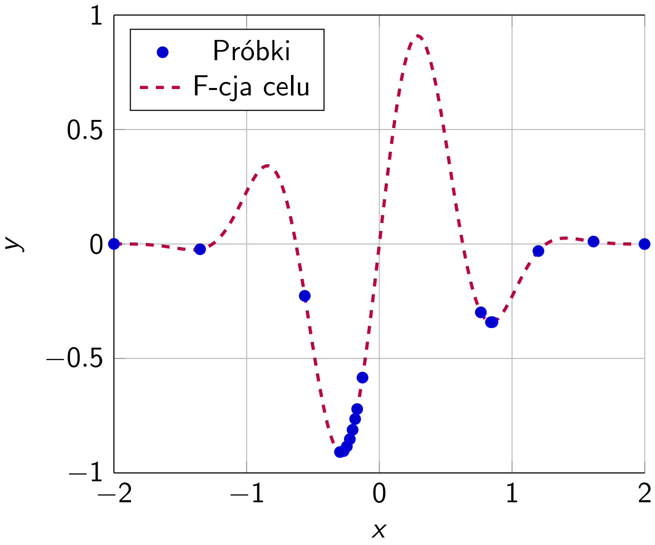
Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$

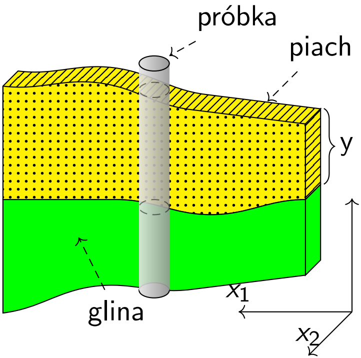


Przykłady: 1. Znana funckja celu $f : [-2, 2] \to [-1, 1]$



Przykłady: 2. Geostatystyka

Odwiert



Warstwy gleby

Przykłady: 2. Geostatystyka — sofrmuowanie problemu

Gdzie znajduje się najgrubsza wieżchnia warstwy piachu?

Sformuowanie problemu:

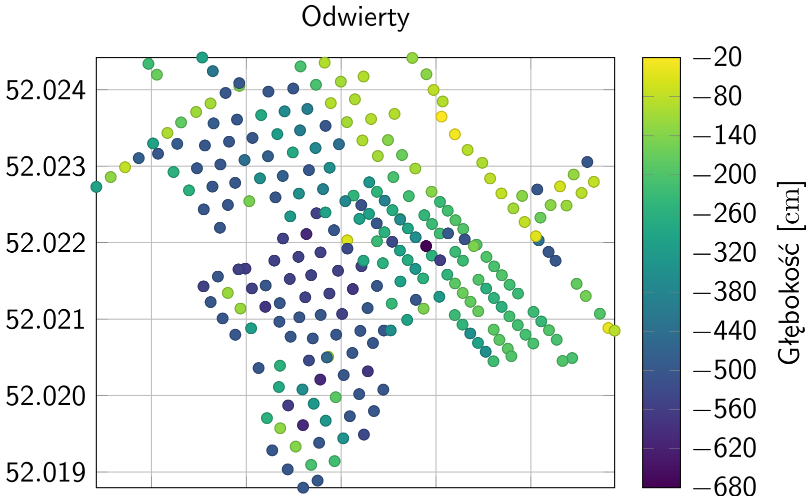
- $\boldsymbol{x}$ — współrzędne geograficzne w obrębie obszaru $\mathcal{X}$,
- y — głębokość/grubość wieżchniej warstwy piachu,
- ewaluacja funkcji celu $f : \mathcal{X} \to \mathbb{R}$ wymaga dokonania odwiertu, jest zatem kosztowna,
- Właściwości f (takie, jak gradient, wypukłość etc.) nie są znane.

Przykłady: 2. Geostatystyka — dane symulacyjne

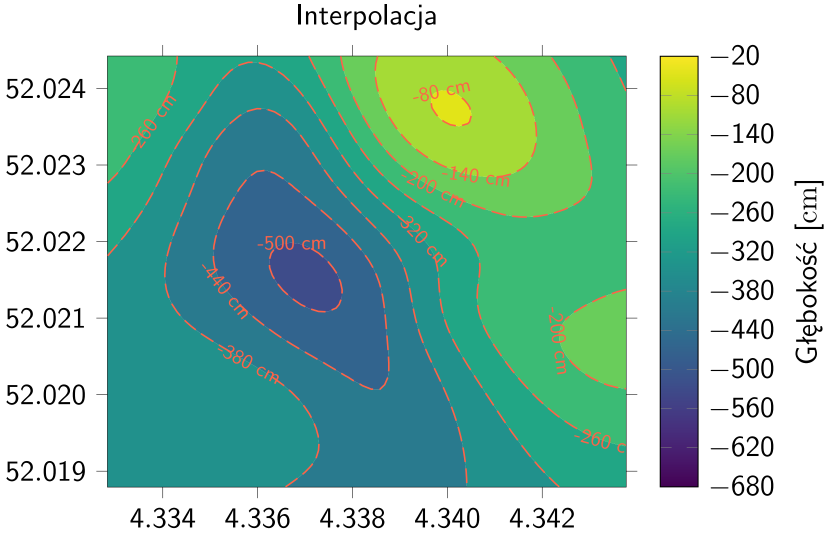
- $\mathcal{X} =$ $[52.0188° \text{ N}, 52.0244° \text{ N}] \times [4.3328° \text{ E}, 4.3438° \text{ E}]$ / WGS84 (okolice miasta Delft w Holandii)
- Interpolacja danych z 518 odwiertów
- Źródło: `https://www.dinoloket.nl`

`https://www.openstreetmap.org`

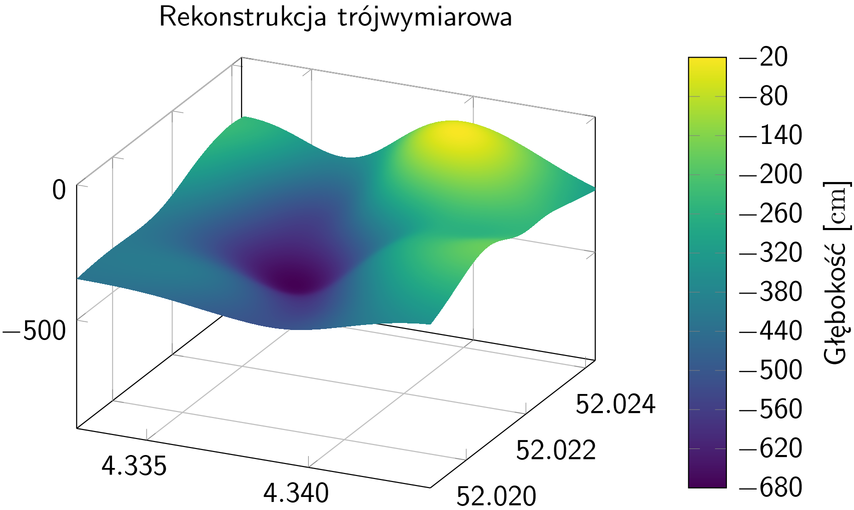
Przykłady: 2. Geostatystyka — dane symulacyjne



Przykłady: 2. Geostatystyka — funkcja celu



Przykłady: 2. Geostatystyka — funkcja celu



Przykłady: 2.1 Geostatystyka — optymalizator

W następującym przykładzie wybrano następującą konfigurację optymalizatora:

Model $\mathcal{M}$

Proces Gaussowski

Funkcja akwizycji α_{EI}

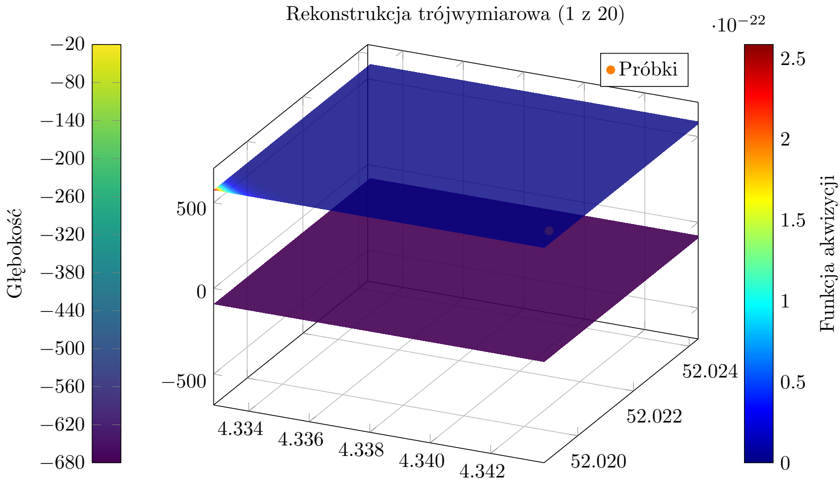
Oczekiwana poprawa (Expected Improvement, EI, [JSW98]);

$$\alpha_{EI}(\boldsymbol{x}; \boldsymbol{\theta}, \mathcal{D}) = \sigma(\boldsymbol{x}; \boldsymbol{\theta}, \mathcal{D})\gamma(\boldsymbol{x})\Phi(\gamma(\boldsymbol{x})) + \mathcal{N}(\gamma(\boldsymbol{x}); 0, 1), \text{ gdzie}$$

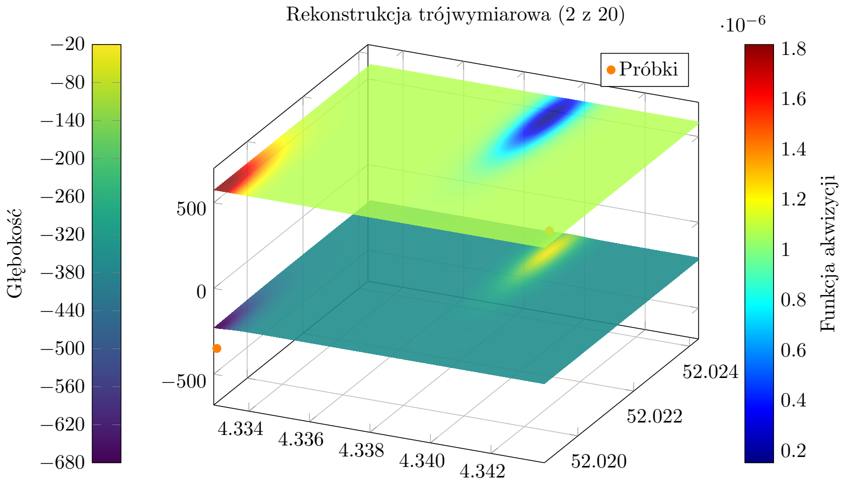
$$\gamma(\boldsymbol{x}) = (\min\{y_{1:n}\} - \mu(\boldsymbol{x}; \boldsymbol{\theta}, \mathcal{D}) + \psi)/\sigma(\boldsymbol{x}; \boldsymbol{\theta}, \mathcal{D}), \text{ oraz}$$

ψ — współczynnik eksploracji/eksploartacji

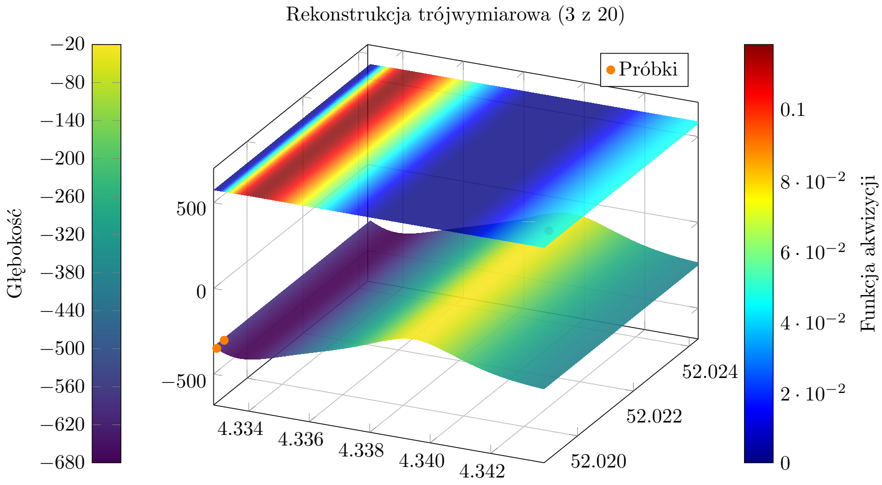
Przykłady: 2.1 Geostatystyka — przebieg



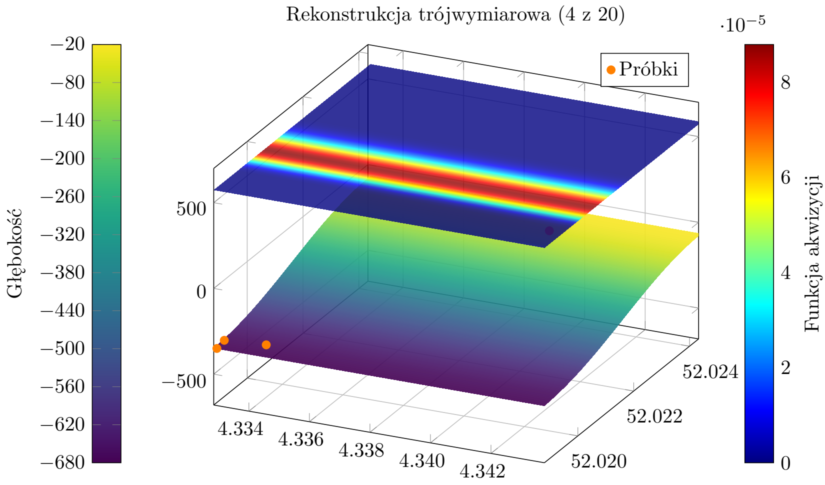
Przykłady: 2.1 Geostatystyka — przebieg



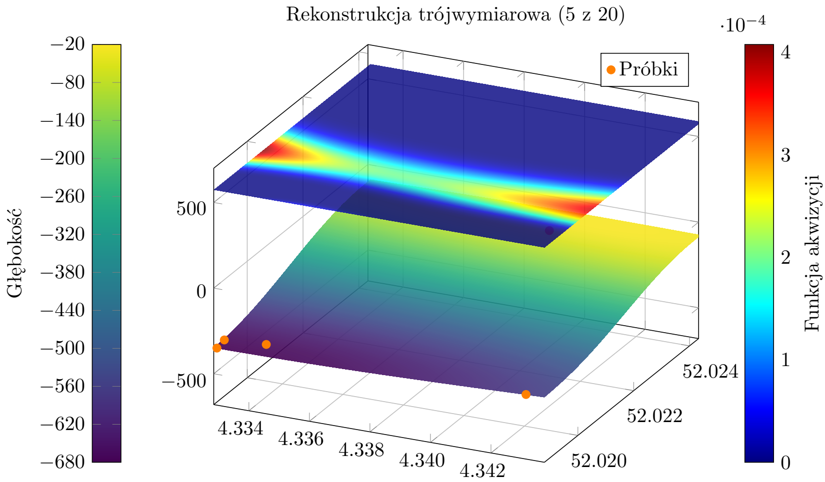
Przykłady: 2.1 Geostatystyka — przebieg



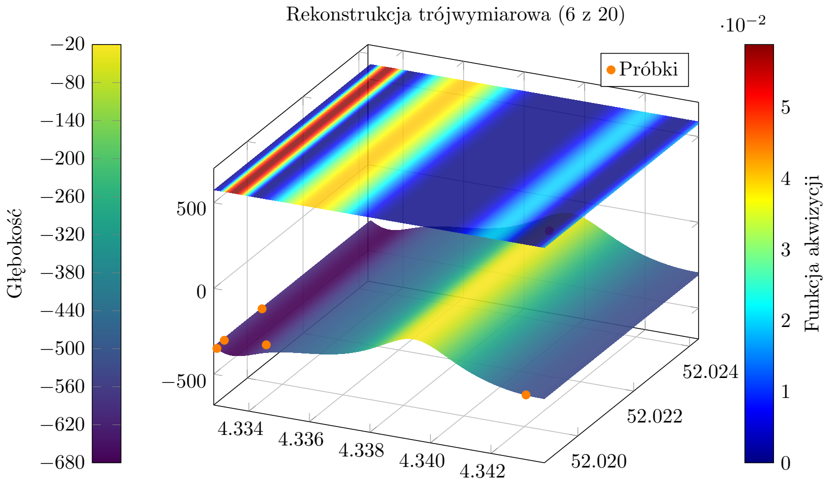
Przykłady: 2.1 Geostatystyka — przebieg



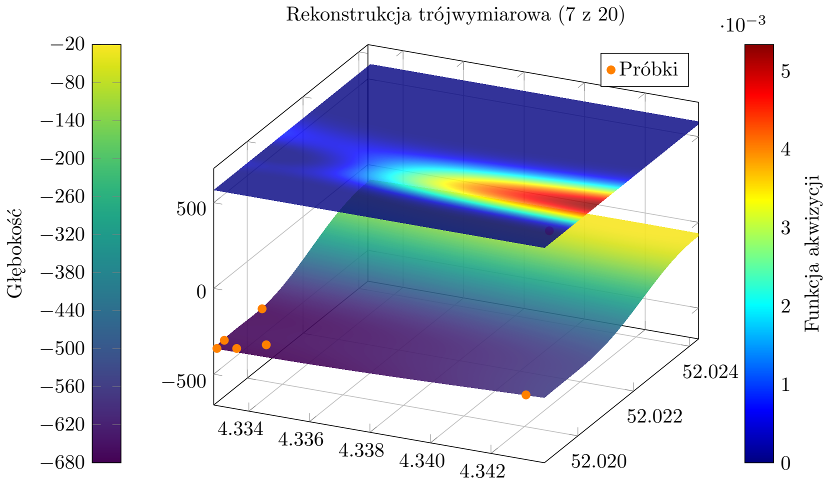
Przykłady: 2.1 Geostatystyka — przebieg



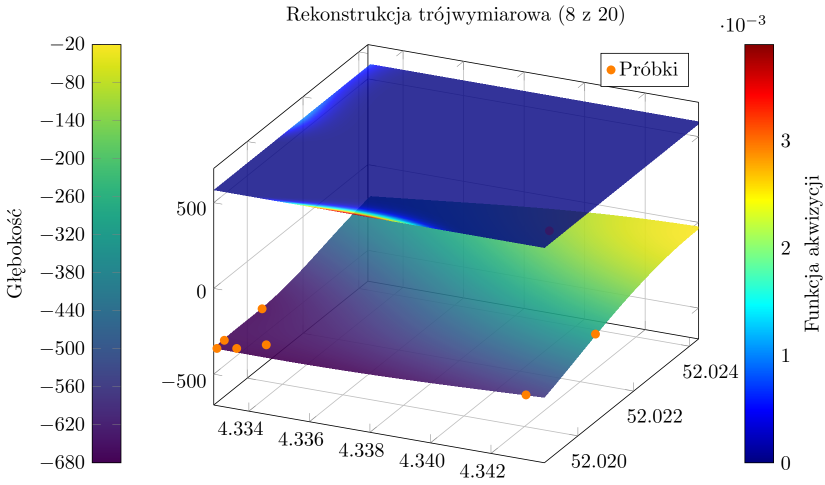
Przykłady: 2.1 Geostatystyka — przebieg



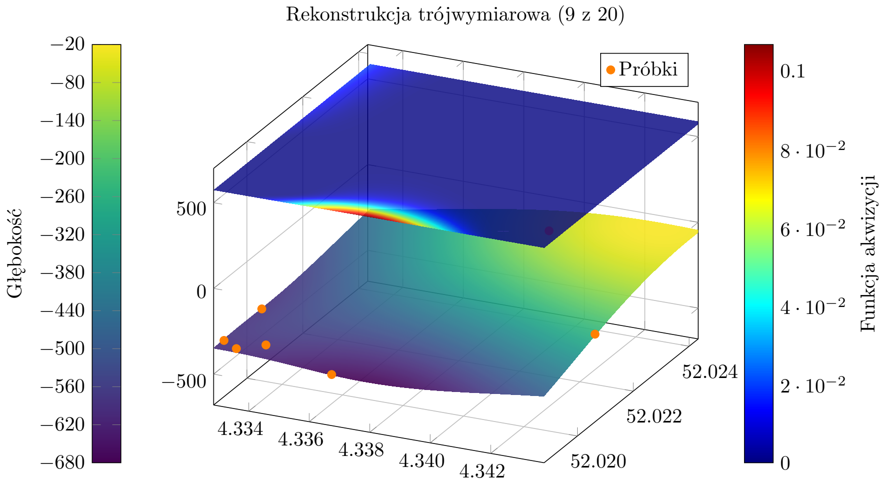
Przykłady: 2.1 Geostatystyka — przebieg



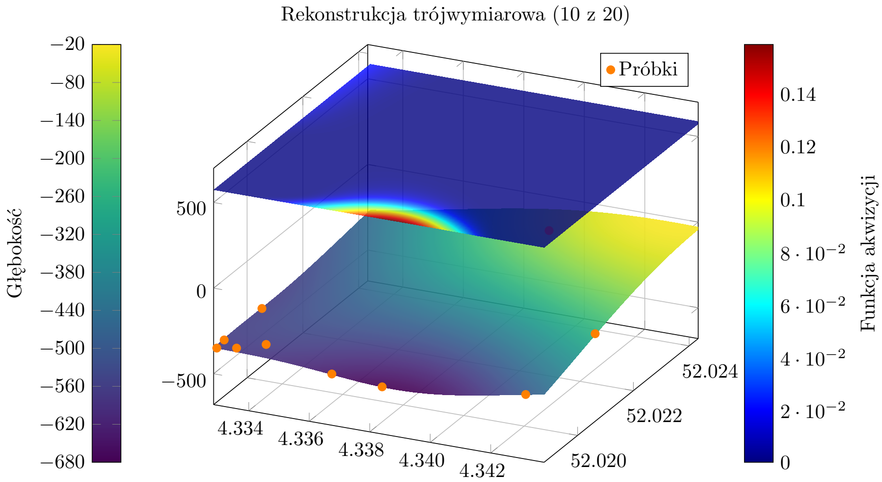
Przykłady: 2.1 Geostatystyka — przebieg



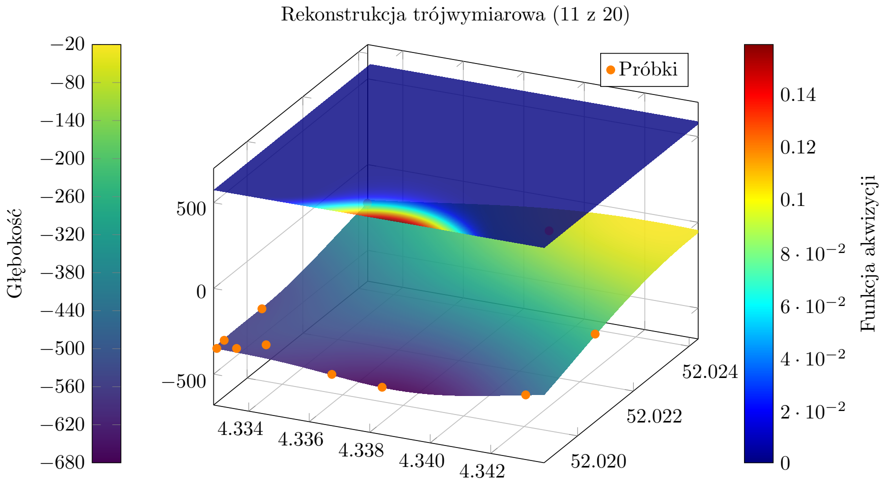
Przykłady: 2.1 Geostatystyka — przebieg



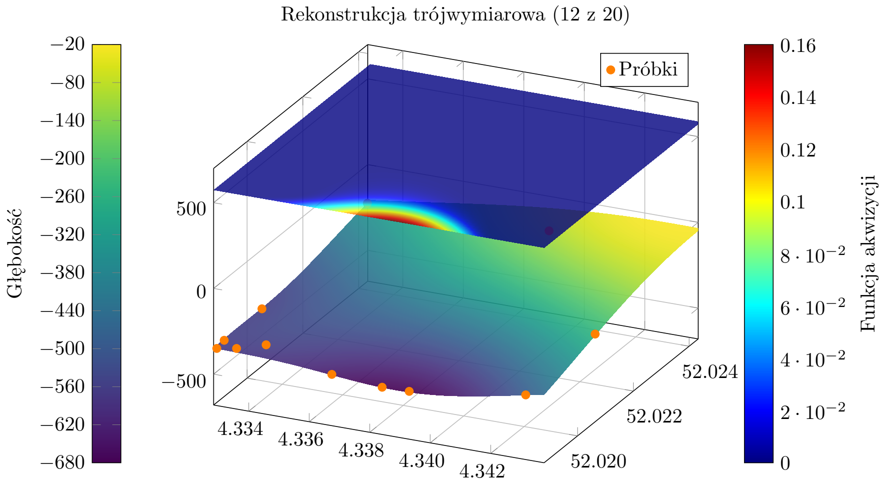
Przykłady: 2.1 Geostatystyka — przebieg



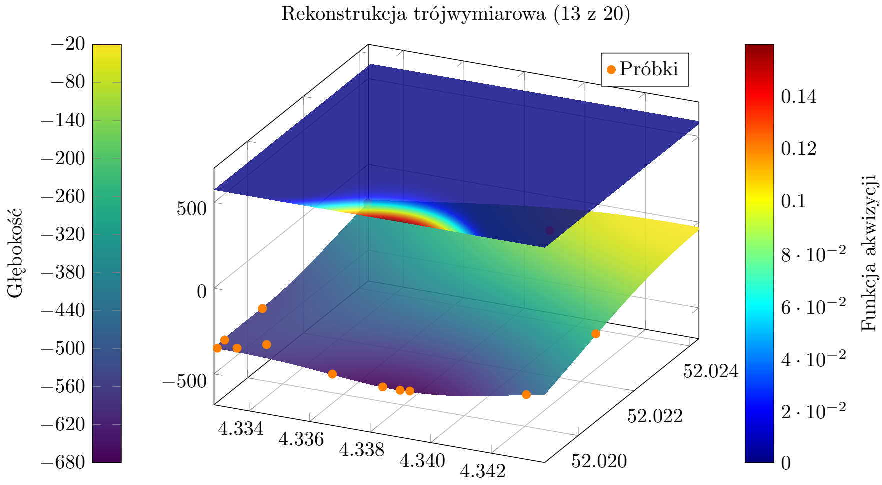
Przykłady: 2.1 Geostatystyka — przebieg



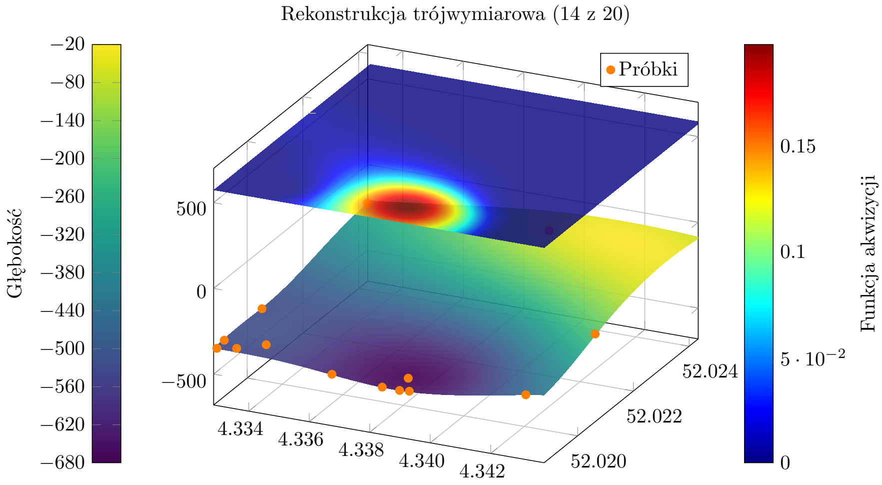
Przykłady: 2.1 Geostatystyka — przebieg



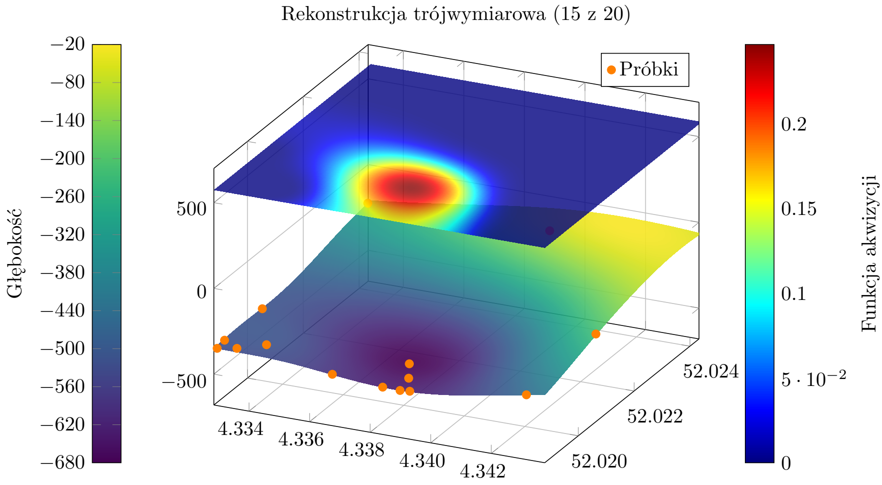
Przykłady: 2.1 Geostatystyka — przebieg



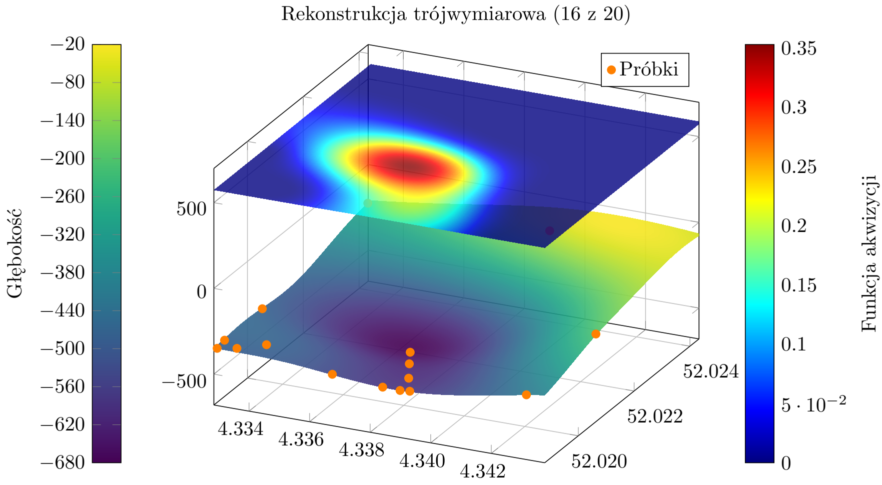
Przykłady: 2.1 Geostatystyka — przebieg



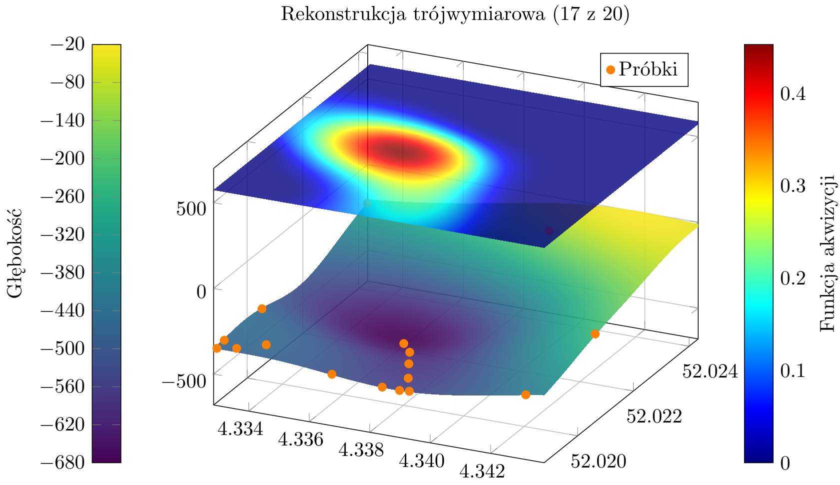
Przykłady: 2.1 Geostatystyka — przebieg



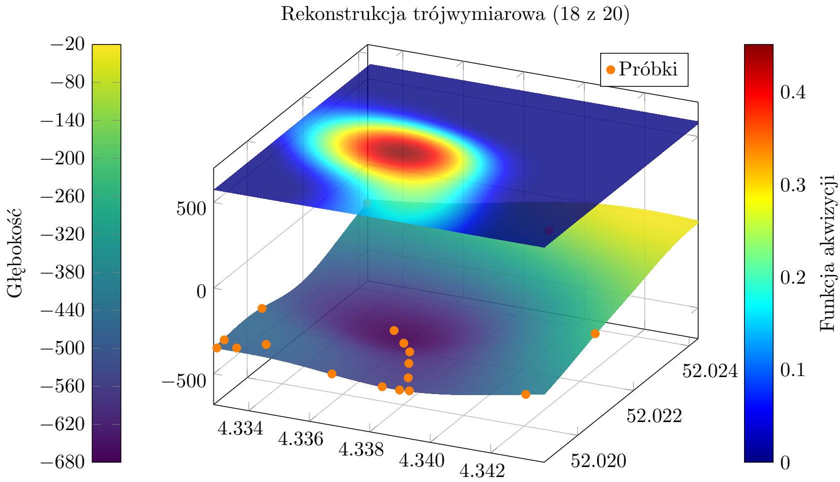
Przykłady: 2.1 Geostatystyka — przebieg



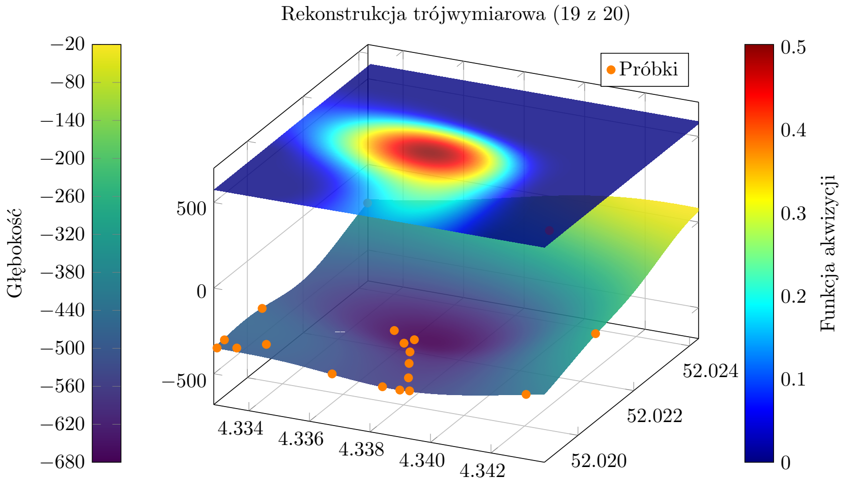
Przykłady: 2.1 Geostatystyka — przebieg



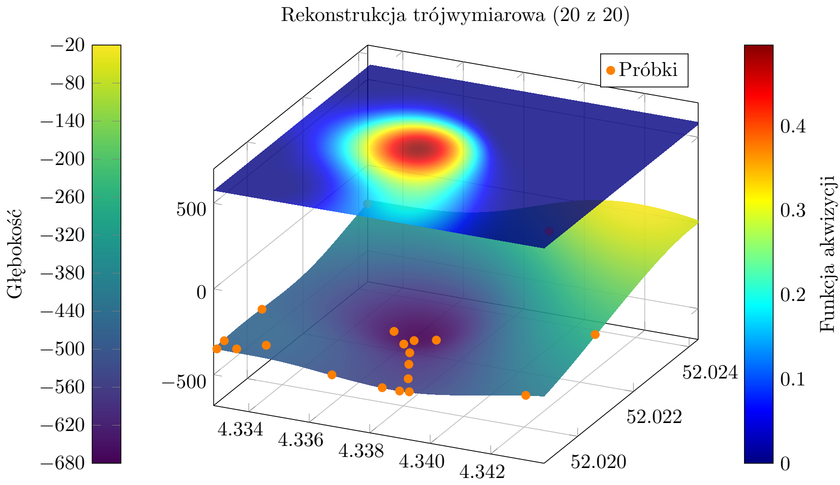
Przykłady: 2.1 Geostatystyka — przebieg



Przykłady: 2.1 Geostatystyka — przebieg



Przykłady: 2.1 Geostatystyka — przebieg



Najlepsza wartość: −515.144 cm, faktyczne minimum ≈ −516.3 cm

Przykłady: 2.2 Geostatystyka — wielokrotna symulacja procesu optymalizacji

Wykonano 100 prób pełnej optymalizacji powyższego wyniku interpolacji:

Modele zastępcze

Wybrano model zastępczy oparty na procesie Gaussowskim z `https://scikit-optimize.github.io`

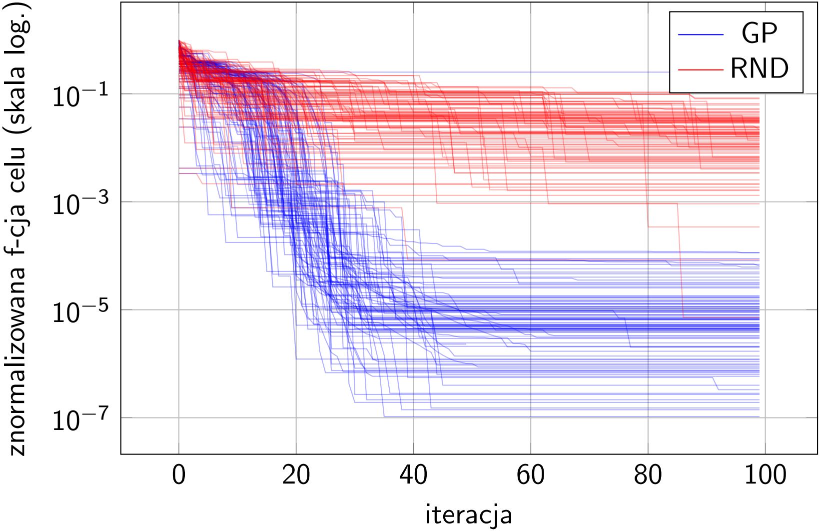
Funkcje akwizycji

Wybrano podstawową funkcję akwizycji z `https://scikit-optimize.github.io`

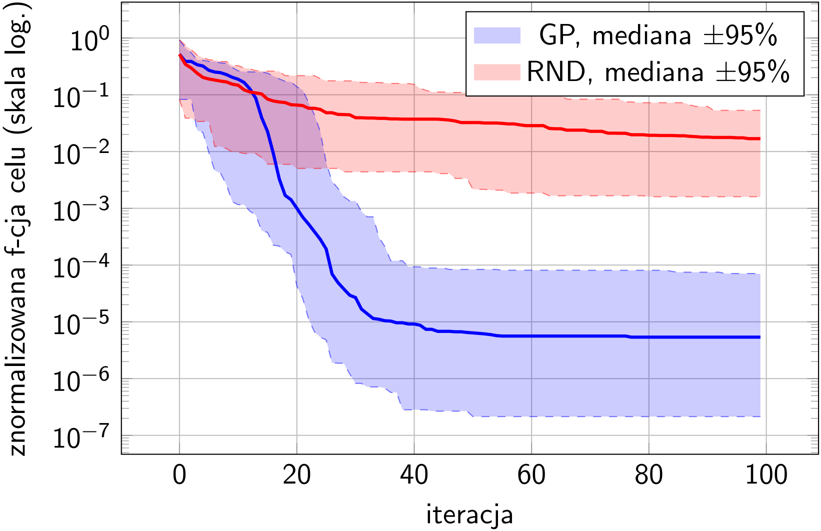
Poziom referencyjny

Dla porównania wykonano również próby przy pomocy zupełnie losowego optymalizatora

Przykłady: 2.2 Geostatystyka – wyniki



Przykłady: 2.2 Geostatystyka – wyniki



Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji

Dolny kres ufności (Lower Confidence Bound, LCB, [SKKS12])

$$\alpha_{LCB}(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta}) = -\mu(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta}) + \beta_t \sigma(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta}), \text{ gdzie}$$

β_t – współczynnik eksploracji/eksploatacji.

Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji

Predykcyjne wyszukiwanie w oparciu o entropię (Predictive Entropy Search, PES, [HLHG14])

$$\alpha_{PES}(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta}) = H\left[p(y|\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{x})\right] - \mathbb{E}_{p(\boldsymbol{x}_*|\mathcal{D},\boldsymbol{\theta})}\left[H\left[p(y|\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{x}, \boldsymbol{x}_*)\right]\right],$$

gdzie:

$$H\left[p(\boldsymbol{x})\right] = -\int_{\mathcal{X}} p(\boldsymbol{x}) \log p(\boldsymbol{x}) d\boldsymbol{x},$$

$\boldsymbol{x}_*$ – argument minimum funkcji celu.

Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji

Rownoległy gradient wiedzy (Parallel Knowledge Gradient, qKG, [WF16])

$$\alpha_{qKG}(\boldsymbol{z}_{1:q}; \mathcal{D}, \boldsymbol{\theta}) = \min_{\boldsymbol{x} \in \mathcal{X}} \mu^{(n)}(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta}) - \mathbb{E}_n \left[\min_{\boldsymbol{x} \in \mathcal{X}} \mu^{(n+q)}(\boldsymbol{x}; \mathcal{D}, \boldsymbol{\theta}) \middle| \boldsymbol{z}_{1:q} \right],$$

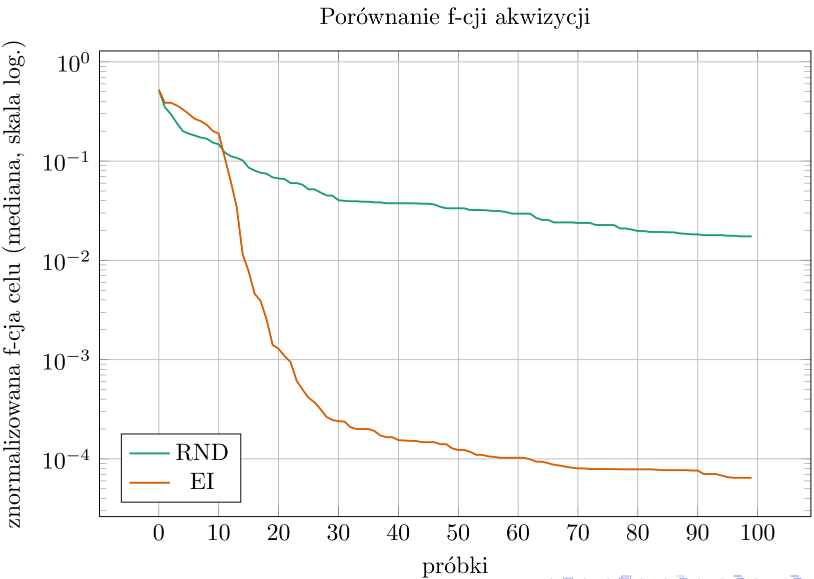
gdzie $\boldsymbol{z}_{1:q}$ to zbiór q kandydatów do próbkowania w aktualnej iteracji.

Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji

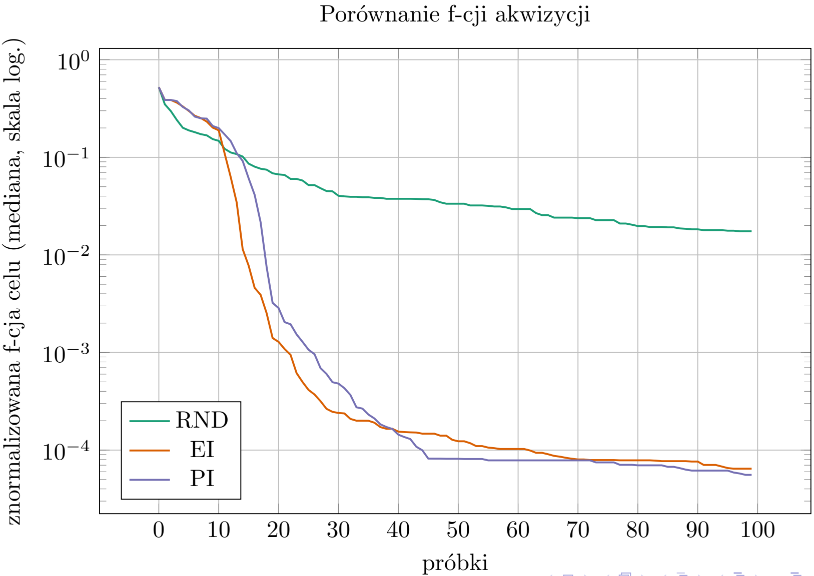
Porównanie:

α	eksp.	próbki	implementacja
PI	100	100	https://scikit-optimize.github.io
EI	100	100	
LCB	100	100	
qKG1	35	52	https://github.com/wujian16/Cornell-MOE
qKG3	40	52	
qKG5	40	52	
PES	45	40	

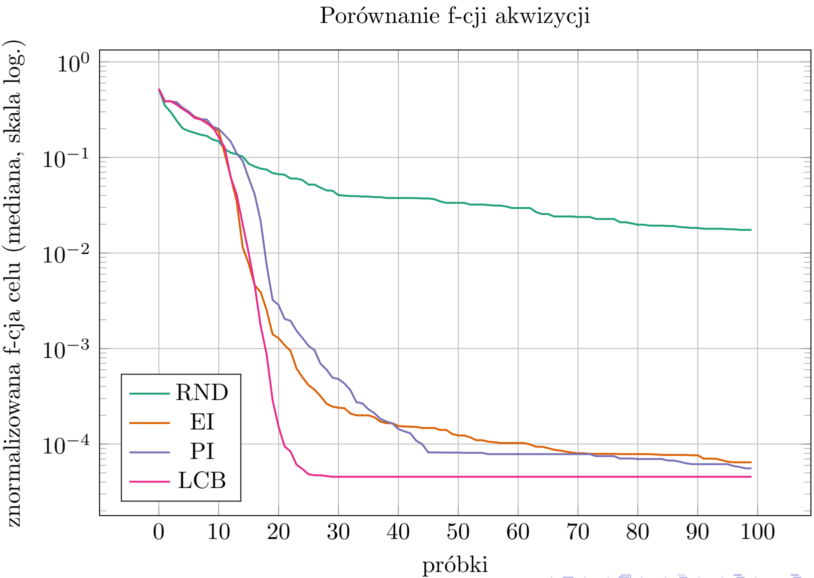
Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji



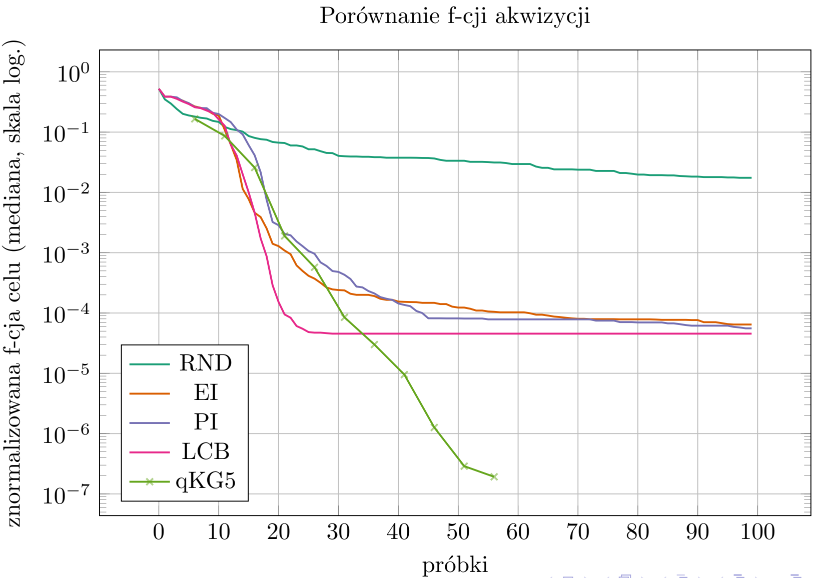
Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji



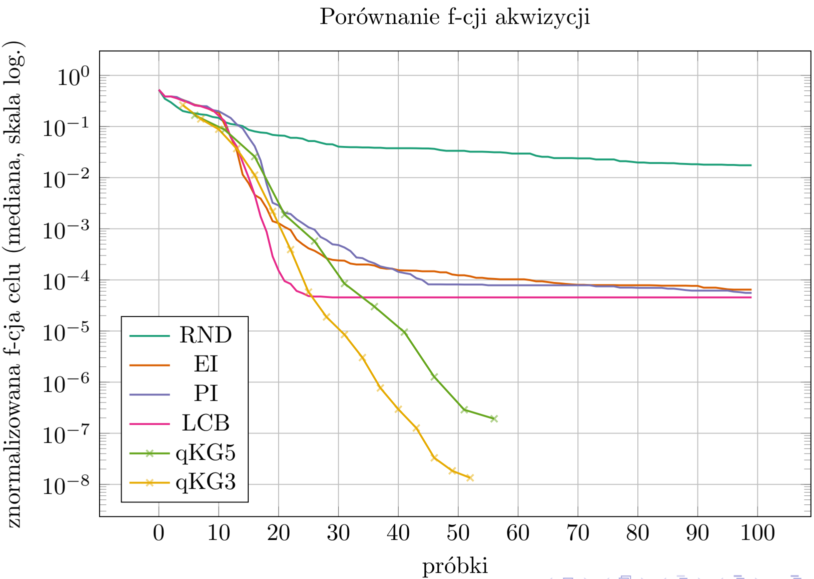
Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji



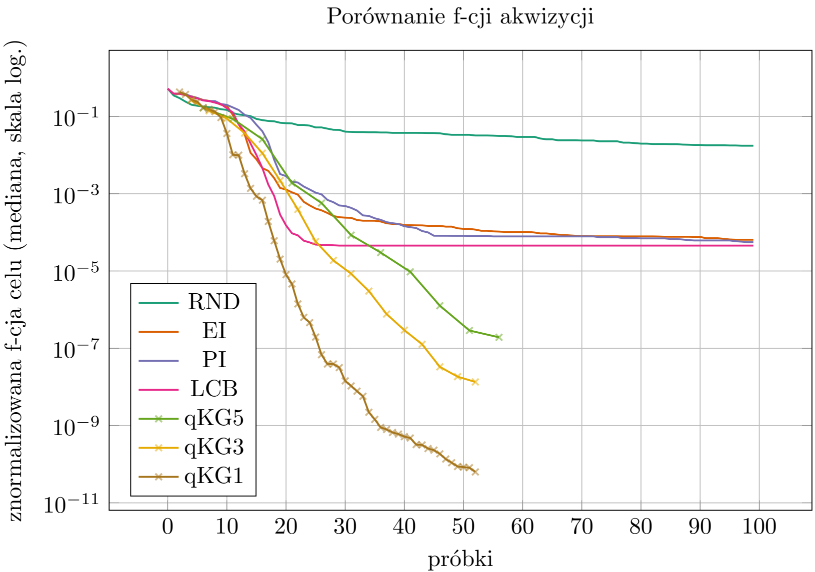
Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji



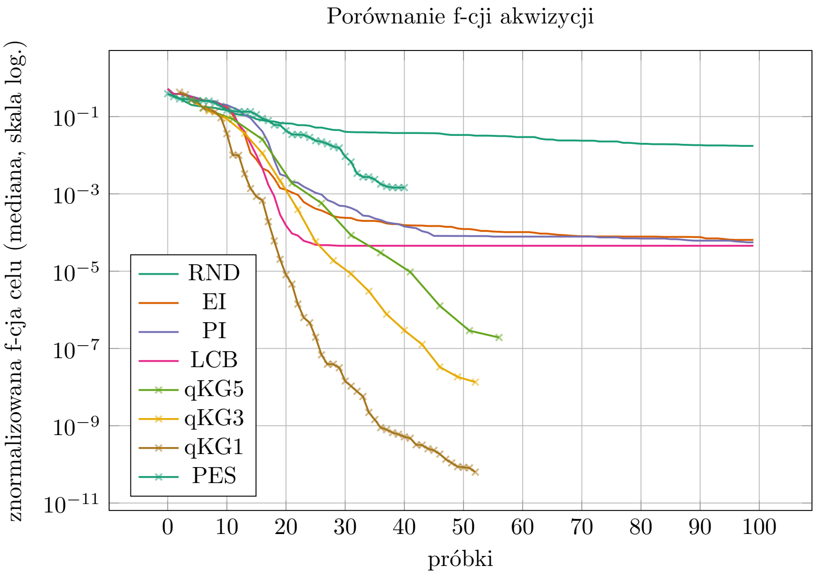
Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji



Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji



Przykłady: 2.3 Geostatystyka – inne funkcje akwizycji



Przykłady: 3. Chód — Opis

Wyszukiwanie optymalnych parametrów chodu robota opisano w [CGS+14].

Krótki opis problemu

- $\boldsymbol{x}$ — 8 ciągłych wymiarów
- y — „zgodność" z pożądaną trajektorią chodu podczas eksperymentu
- Ewaluacja — uruchomienie robota, pomiar trajektorii, porównanie z idealną trajektorią

Przykłady: 3. Chód — Wyniki

Wyszukiwanie optymalnych parametrów chodu robota opisano w [CGS+14].

Wyniki

- Zbieżność po ok 80 ewalucjach
- Trajektoria chodu lepsza, niż po manualnym strojeniu parametrów przez ekspertów
- Skrócenie czasu optymalizacji z kilku dni do kilku godzin
- Film pokazujący chód robota podczas i po zakończeniu optymalizacji: `https://youtu.be/ualnbKfkc3Q`

Oprogramowanie: Optymalizacja Bayesowska

Ax	https://ax.dev
BoTorch	https://github.com/pytorch/botorch
MOE	https://github.com/Yelp/MOE
Cornell-MOE	https://github.com/wujian16/Cornell-MOE
SMAC	http://www.cs.ubc.ca/labs/beta/Projects/SMAC/
SMAC3	https://github.com/automl/SMAC3
GPyOpt	https://sheffieldml.github.io/GPyOpt/
skopt	https://scikit-optimize.github.io/
pyGPGO	https://github.com/hawk31/pyGPGO

Oprogramowanie: Wybrane środowiska probabilistyczne

Pyro https://pyro.ai/

NumPyro https://github.com/pyro-ppl/numpyro

TensorFlow Probability https://www.tensorflow.org/probability

PyMC 3 https://github.com/pymc-devs/pymc3 (defunct?)

PyMC 4 https://github.com/pymc-devs/pymc4 (pre-release)

STAN https://mc-stan.org/

... i wiele innych:
https://en.wikipedia.org/wiki/Probabilistic_programming

Literatura I

[CGS+14] Roberto Calandra, Nakul Gopalan, André Seyfarth, Jan Peters, and Marc Peter Deisenroth, *Bayesian gait optimization for bipedal locomotion*, Learning and Intelligent Optimization (Cham) (Panos M. Pardalos, Mauricio G.C. Resende, Chrysafis Vogiatzis, and Jose L. Walteros, eds.), Springer International Publishing, 2014, pp. 274–290.

[HLHG14] José Miguel Hernández-Lobato, Matthew W. Hoffman, and Zoubin Ghahramani, *Predictive entropy search for efficient global optimization of black-box functions*, 2014.

Literatura II

[JSW98] Donald R. Jones, Matthias Schonlau, and William J. Welch, *Efficient Global Optimization of Expensive Black-Box Functions*, Journal of Global Optimization **13** (1998), no. 4, 455–492.

[Kus64] H. J. Kushner, *A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise*, Journal of Fluids Engineering **86** (1964), no. 1, 97–106.

[RW05] Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian processes for machine learning (adaptive computation and machine learning)*, The MIT Press, 2005.

Literatura III

[SKKS12] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger, *Information-theoretic regret bounds for gaussian process optimization in the bandit setting*, IEEE Transactions on Information Theory **58** (2012), no. 5, 3250–3265.

[SSW+16] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas, *Taking the human out of the loop: A review of bayesian optimization*, Proceedings of the IEEE **104** (2016), no. 1, 148–175.

[WF16] Jian Wu and Peter I. Frazier, *The parallel knowledge gradient method for batch bayesian optimization*, 2016.