O wielorękich bandytach, nieuczciwych kasynach i sprytnych statystykach

Plan prezentacji

- Wieloręki bandyta
- Eksploracja vs Eksploatacja
- Motywacje
- Rozwinięcie problemu
- Algorytmy
- Optymalizacja Bayesowska

Wieloręki bandyta

Model

Wieloręki bandyta

Model

- Bandyta ma K ramion
- Każde ramię i płaci 1 PLN z prawdopodobieństwem p_i
- Nie znamy $\{p_i\}$ ale wiemy, że są stałe w czasie
- W każdym kroku t wybieramy ramię a_t którym gramy
- Na podstawie naszego wyboru otrzymujemy wygraną:

  $$r_t \sim \text{Bernouli}(p_{a(t)})$$

- W jaki sposób grać, aby zmaksymalizować wygraną?
- Możemy rozważać koszt gry c
- Szukamy strategi π która minimalizuje stratę:

  $$R_t = p^* t - E[\ \text{sum}_t\ r_{\pi(t)}\ ]$$

Wieloręki bandyta

Model

- Mamy skończony budżet i w ramach tego budżetu chcemy osiągnąć największe zyski (ROI). Jednocześnie optymalizujemy i robimy użytek z naszej wiedzy.
- Pokrewne problemy:
  - Optymalizacja funkcji przy założonym budżecie – użytek z naszej wiedzy będziemy robić później (optymalizacja Bayesowska – będzie w dalszej części).
  - Minimalizacja wariancji estymatora przy założonym budżecie (optymalne projektowanie eksperymentu, aktywne uczenie http://burrsettles.com/pub/settles.activelearning.pdf)

Wieloręki bandyta

Podejście „naiwne” (zachłanne)

- Gra trwa T rund (liczbę rund znamy na początku)
- Na początku gramy każdym ramieniem N razy (faza eksploracji K N < T)
- Pozostały czas gramy ramieniem które wypadło najlepiej w fazie eksploracji (eksploatacja)

Wieloręki bandyta

Podejście „naiwne” (zachłanne)

- $N = T^{2/3}(\log T)^{1/3}$
- $R_T \leq O(T^{2/3} (\log T)^{1/3}) \leftarrow$ na koniec gry
- Czy możemy grać lepiej (strategie adaptatywne)?
- O ile lepiej możemy grać (ograniczenie dolne na R_T)?

Wieloręki bandyta

Podejście „naiwne”- dlaczego to działa

- Dla każdego ramienia zachodzi $|S_n - N p| \leq \varepsilon$ i wybieramy złe ramię => dużo nie tracimy:

$$R_T \leq N + O(\varepsilon\ (T-KN))$$

- Jakie jest prawdopodobieństwo, że tracimy dużo, tj.:

$$|S_n - N p| > \varepsilon?$$

- Ograniczenie Hoeffdinga:

$$P(|S_n - N p| \leq \varepsilon) \geq 1 - \exp(-2\ \varepsilon^2 N)$$

- $E[R_T] = E[R_T|ok]p(ok) + E[R_T|bad]p(bad)$
- $\varepsilon=(2 \log T/N)^{1/2}$, $N = T^{2/3}(\log T)^{1/3}$ => $R_T \leq O(T^{2/3} (\log T)^{1/3})$

Wieloręki bandyta

Algorytm ε-zachłanny

- Z prawdopodobieństwem (1 - ε), na podstawie dotychczasowych obserwacji wybierz ramię z największym prawdopodobieństwem wygranej.
- Z prawdopodobieństwem ε wybierz losowe ramię.
- $\varepsilon_t \sim t^{-1/3}$
- $R_t \leq O(t^{2/3} (\log t)^{1/3}) \leftarrow$ w każdym momencie gry
- Czy można lepiej

- Co jeśli kasyno oszukuje? Algorytm może przynieść duże straty: $R_T = o(T) \leftarrow$ o tym jeszcze będzie później
- Eksplorujemy przestrzeń parametrów w sposób nieefektywny.

Wieloręki bandyta

Górny przedział ufności (UCB1)

- Wybierz każde ramię raz
- W każdej rundzie *t*:
  - Oblicz średnią wartość ramienia: $w_t(a)$
  - Oblicz przedział ufności dla ramienia: $r_t(a) = (2 \log T/n_t(a))^{1/2}$
  - Graj optymistycznie, tj. wybierz ramię: $\arg\max_a w_t(a)+r_t(a)$
  - Lepsze szacowanie przedziału ufności – UCB-tuned

Wieloręki bandyta

Górny przedział ufności (UCB1)

- $R_t \leq O((Kt\ (\log T))^{1/2})$
- Jeśli coś wiemy o p_i: $R_t \leq O(\log T)[\sum_a 1/\Delta(a)]$

Wieloręki bandyta

Jak dobry może być algorytm

- Rodzina wrednych Bandytów:
  $I_i=\{p_0=1/2, p_1=1/2, \ldots, p_i=(1+\varepsilon)/2, \ldots, p_K=1/2\}$
- Trudno powiedzieć z którym bandytą mamy do czynienia:
  $|p(A)-q(A)| \leq \varepsilon\, T^{1/2}$
- Algorytm: A_i <=> i
- $R_T \geq \Omega((K\,T)^{1/2})$
- Nawet jeśli coś wiemy o p_i: $R_t \geq \Omega(\log t)$

Wieloręki bandyta

Eksploracja vs eksploatacja

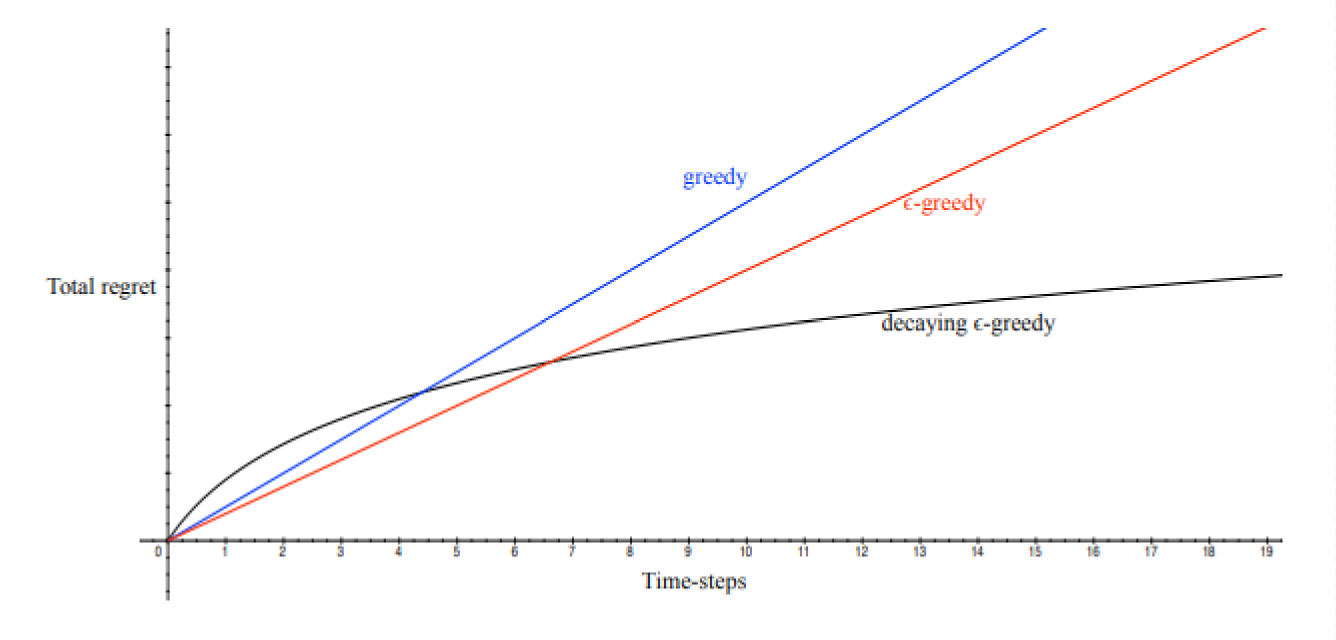
- Eksploatacja – wykorzystaj dotychczasowe dane do podjęcia najlepszej decyzji
- Eksploracja – zgromadź więcej danych

- Najlepsza strategia długoterminowa nie musi być lokalnie optymalna.
- Od czasu do czasu warto wyjść poza „strefę komfortu” oraz inne życiowe prawdy.

Wieloręki bandyta

Eksploracja vs eksploatacja

- Jeśli bez końca eksplorujemy to nasze straty będą liniowe w czasie.



Motywacje

- Marketing: optymalizacja cen, optymalizacja asortymenty, optymalizacja komunikatów/reklam, optymalizacja układu strony (Amazon, Yahoo!)
- Koszty biznesowe nieoptymalnego działania
- Optymalizacja portfolio (np. przydział środków na projekty R&D)
- Optymalizacja terapii, testy kliniczne

Warianty

- Przestrzeń parametrów jest dyskretna czy ciągła?
- Czy jest jakaś „korelacja przestrzenna” między bandytami/ramionami? Bliskie *i* implikuje bliskie p_i?
- Czy mamy jakąś dodatkową wiedzę o rozkładzie p_i?
- Czy p_i zmienia się w czasie (stacjonarność vs niestacjonarność)?
- Czy p_i zależy od poprzednich ewaluacji (markowość)?
- Czy jest dodatkowy „kontekst” od którego zależy wygrana (np. informacje o użytkowniku do którego adresujemy reklamę, parametry reklamy)?
- Bandyta przeciwnik (adversarial bandit) – nieuczciwe kasyno zna historię naszych ruchów oraz algorytm i na tej podstawie ustala p_i w następnej rundzie.
- Opóźnienia informacji o wygranej (np. poznajemy wygraną 10 rund później, CTR).
- Możliwość przeprowadzenia kilku gier równolegle
- Skończony/otwarty horyzont czasowy.
- Inne modele straty (np. probably approximately correct).

Wieloręki bandyta

Algorytm Bayesowski

- Czy możemy zrobić użytek z naszej wiedzy o rozkładzie wygranych $\{p_i\}$, korelacjach między bandytami, znajomości kontekstu?
- Strata Bayesowska/strukturalna

  $BR_t = E_\Theta[p^* t - E[\ sum_t\ r_{\pi(t)}\ ]]$

  Θ – parametry od których zależy wygrana

Wieloręki bandyta

Algorytm Bayesowski

- Wyznacz rozkład aposteriori prawdopodobieństwa wygranej p_i, przy założeniu dotychczasowych obserwacji (do chwili t).
- Policz wartość oczekiwaną zysku dla każdego i:

  $$r^*_i = E[\Sigma_{t'>t}\, r_i | p_i\,]$$

- Optymalizuj względem i

Wieloręki bandyta

Algorytm Bayesowski

- Elastyczność, można uwzględnić:
  - wiedzę aprioi
  - zmienne ciągłe i dyskretne
  - strukturę (korelacje, warunek Lipschitza)
  - zmienność w czasie
  - kontekst
  - szum/brak szumu (optymalizacja funkcji deterministycznych: rozkład aprioi → klasa funkcji)
- Duża złożoność obliczeniowa (aproksymacja przy użyciu MCTS), dla niektórych przypadków łatwiej

Wieloręki bandyta

Algorytm Bayesowski

- Heurystyka: „bądź optymistą – wybieraj najbardziej obiecujące ramię”
- Funkcja użyteczności (np. UCB, p. poprawy)

Wieloręki bandyta

Próbkowanie Thompsona

- Wylosuj p_i z rozkładu a posteriori, przy założeniu dotychczasowych obserwacji (do chwili t).
- Wyznacz i optymalne dla wylosowanego p_i

- Algorytm łatwy do implementacji i tani obliczeniowo
- Równoległe eksperymenty – losujemy wiele p_i
- Wyniki eksperymentów potwierdzają skuteczność w porównaniu z innymi algorytmami
- $R_T = O((T K \log T)^{1/2})$
  - r_t - rozkład Bernoullego, rozkład jednorodny a priori
  - r_t - rozkład Gaussa, rozkład jednorodny a priori

Wieloręki bandyta

Hedge, EXP3, EXP4, EXP...

- Wybierz ramię *i* z prawdopodobieństwem:

  $p_{i,t}=(1-\gamma)\, w_{i,t}/\Sigma_j\, w_{j,t}+\gamma/K$
- Aktualizacja wag:

  $w_{i,t+1} = w_{i,t} exp(\gamma'\, r_i / K)$

  $w_{i,t+1} = w_{i,t} exp(\gamma'\, r_i / p_{i,t}\, K)$
- Gra z przeciwnikiem (adversial bandit):
  - Na początku rundy przeciwnik ustala wypłaty r_i, przeciwnik może znać historię ruchów.
  - Gracz wybiera ramię i otrzymuje wypłatę; gracz nie zna r_i
  - R_t = r*t - E[$sum_t\, r_{\pi(t)}$]
- $R_T \leq$ O(sqrt(T K log(K))); jeśli istnieje optymalna strategia, jesteśmy w stanie się jej nauczyć

Wieloręki bandyta

Niestacjonarność

- Kasyno co jakiś czas manipuluje przy bandytach
  - Zmiana skuteczności leków (np. nabywanie odporności przez bakterie)
  - Zmiana zachowań użytkowników serwisów internetowych
- Zapominanie starych wyników (UCB, EXP):
  - Wykładnicze dyskontowanie
  - Okno
  - Restart
- Wykrywanie zmian strukturalnych (EXP.R)

Wieloręki bandyta

Kontekst

- Kontekst bandyty i kontekst gracza:
  - Podobni bandyci dają podobne wygrane
  - Podobni gracze mają podobne szczęści
- Posiadamy dodatkowe informacje które możemy wykorzystać:
  - Dane pacjenta, historia choroby, dawka leku
  - Historia przeglądania, kategoryzacja produktów
  - Parametry układu treści na stronie, kategoryzacja tematyczna treści

Wieloręki bandyta

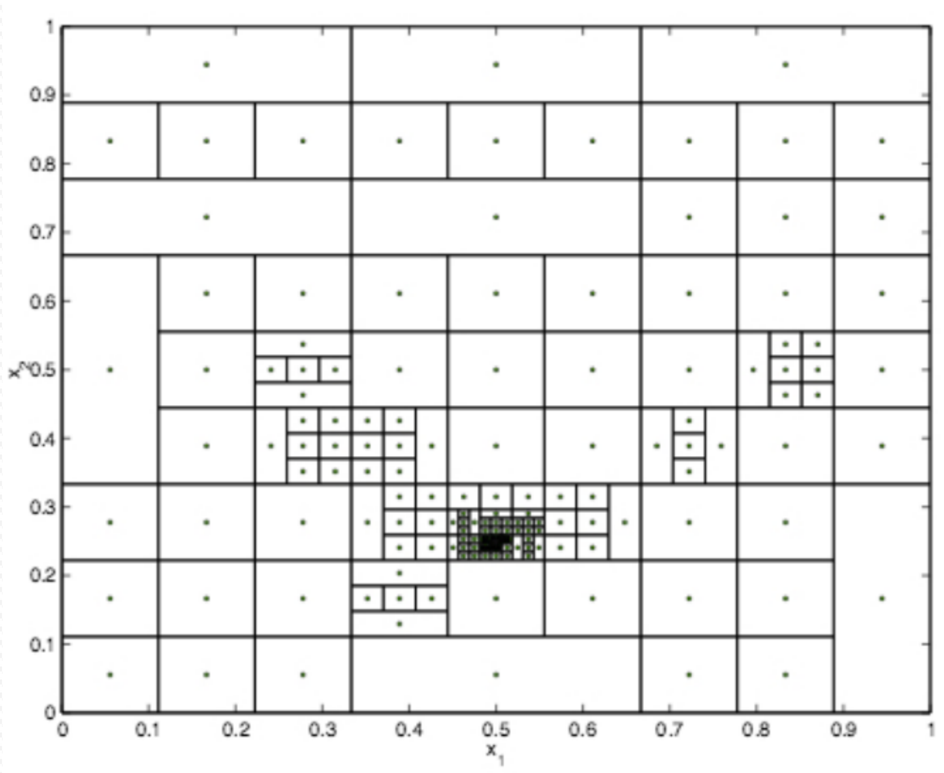
Kontekst

- Dla skończonej liczby kontekstów redukcja do KC niezależnych ramion; nie korzystamy z informacji o podobieństwie.
- Korelacje, ograniczenie/ciągłość Lipschitza: $|r_x - r_y| \leq L|x-y|$
  - Dyskretyzacja: $R_T \leq R_T(S) + DE(S)$
  - Adaptatywna dyskretyzacja – rozmiar siatki zależy od przedziału ufności
  - $R_T \leq O(T^{d+1/d+2} (\log T)^{1/d+2})$
- Podział przestrzeni przy użyciu drzewa/podejście hierarchiczne (BAST, HOO)

Wieloręki bandyta

Kontekst

- Adaptatywna dyskretyzacja: DIviding RECTangles - DIRECT, Perttunen at al. 1993



Wieloręki bandyta

Kontekst

- Model liniowy – LinUCB (niezależna parametryzacja)

  $\arg\max_a w_t(a)+r_t(a)$

  $w_t(a) = x_{t,a}\ \theta_{t,a},\ r_t(a) = \alpha\ \mathrm{sqrt}(x_{t,a}\ A_{t,a}^{-1}\ x_{t,a})$

  $\theta_{t,a} = A_{t,a}^{-1} b_{t,a}$

  $A_{t,a} = A_{t-1,a} + x_{t,a} x_{t,a}$

  $b_{t,a} = b_{t-1,a} + x_{t,a} r_{t,a}$

Optymalizacja Bayesowka

Wprowadzenie

- Optymalizacja funkcji f(x) której ewaluacja jest kosztowna:
  - długi czas obliczania funkcji – optymalizacja meta parametrów w uczeniu maszynowym
  - czynnik ludzki (np. uczenie preferencji, modelowanie materiałów 3D https://arxiv.org/pdf/1012.2599.pdf)
  - adaptatywny eksperyment, testy A/B
  - Przeszukiwanie drzew metodami monte carlo (MCTS)
  - geologia, ekologia
- Optymalizacja f(x) przy założonym budżecie – minimalizacja liczby wywołań f(x)

Optymalizacja Bayesowka

Założenia

- Założenia odnośnie „gładkości” optymalizowanej funkcji
- Rozkład prawdopodobieństwa służy do modelowania niewiedzy

Optymalizacja Bayesowka

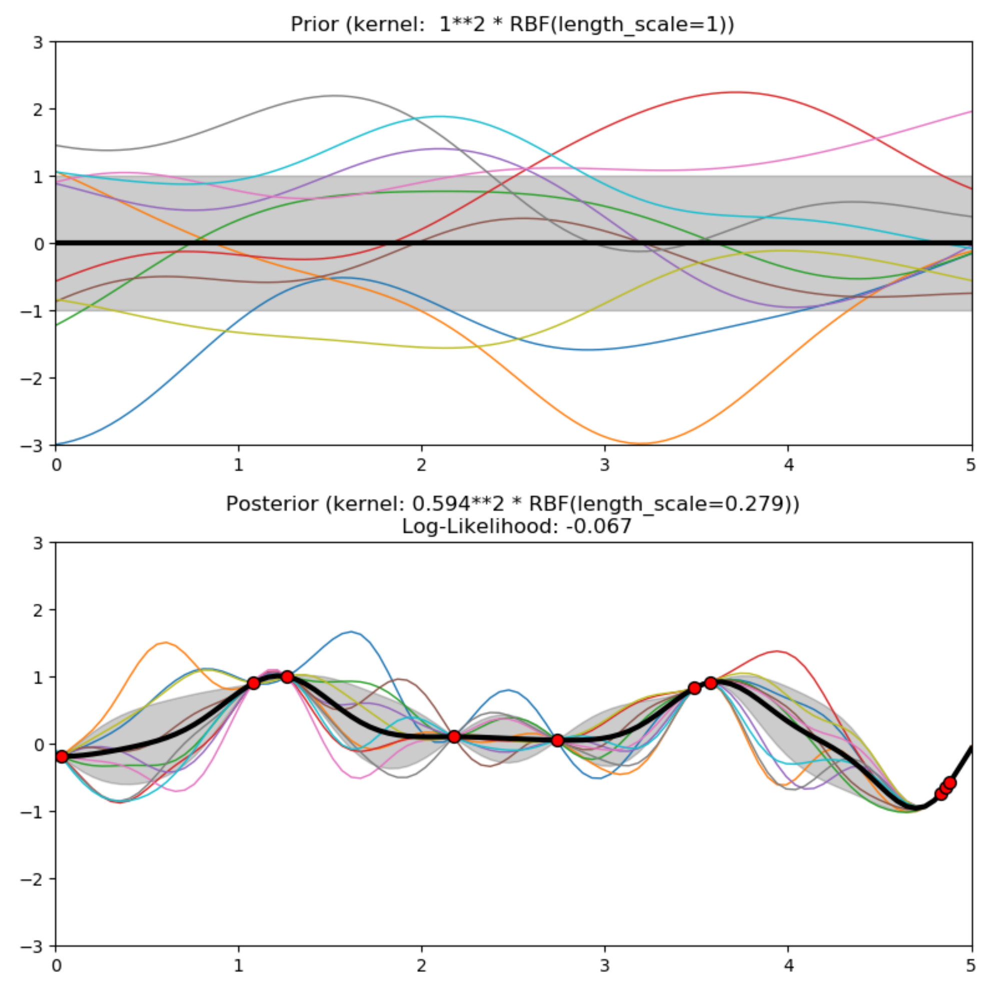
Rozkład apriori funkcji

- Założenia odnośnie „gładkości” optymalizowanej funkcji f – ciągłość w sensie Lipschitza:

  $|f(x) - f(y)| \leq L|x-y|$
- Rozkład apriori powinien umożliwiać łatwą aktualizację wraz z nadchodzącymi danymi oraz łatwe obliczanie „odchylenia” od wartości „średniej”
- Lasy losowe (SMAC)
- Proces Gaussowski f ~ GP(m, k), gdzie:
  - m – wartości średnie
  - k – funkcja kowariancji, np. RBF ~ exp(- $|x-x'|^2/2\theta^2$)

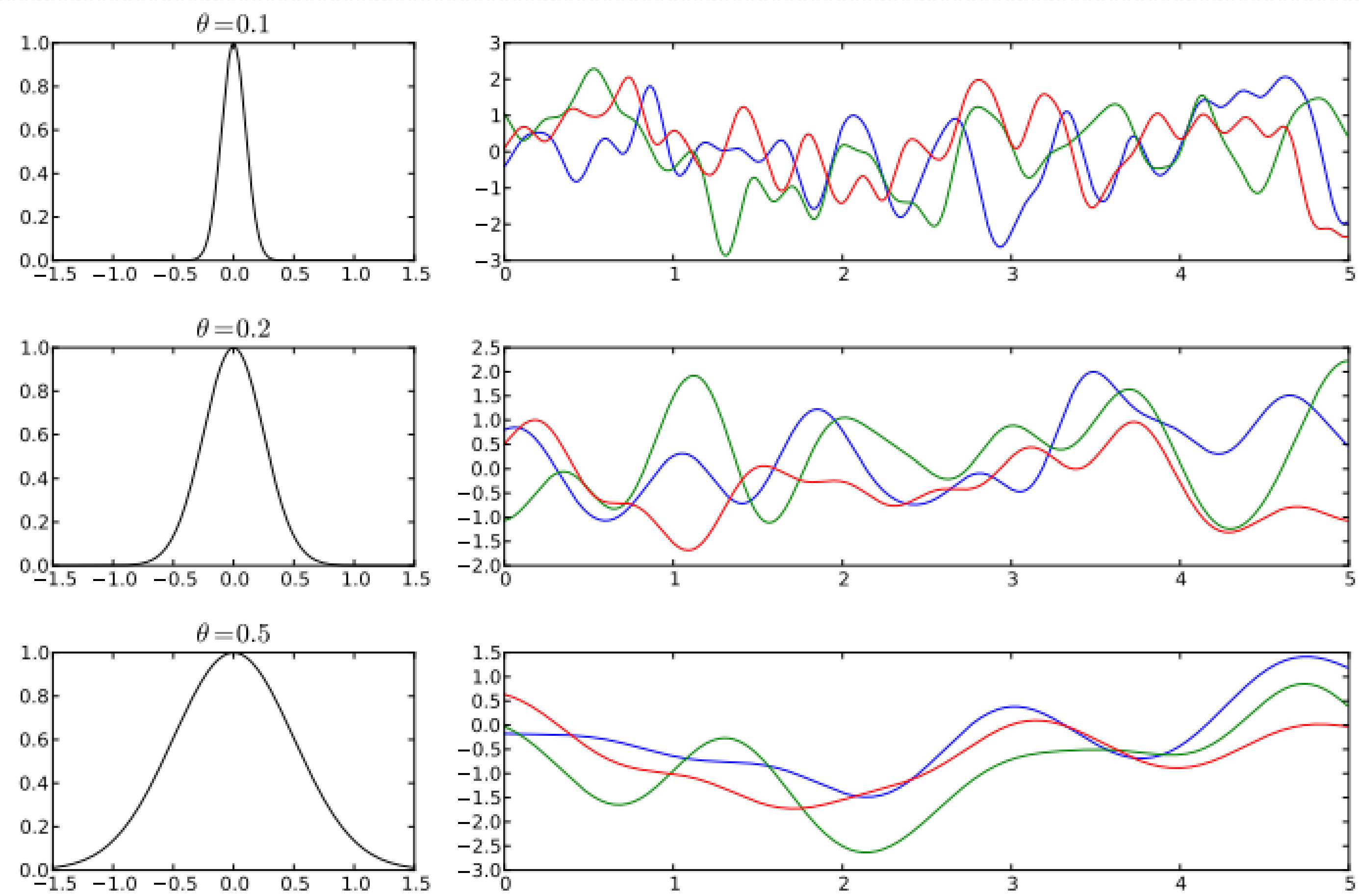
Optymalizacja Bayesowka

Rozkład apriori funkcji



(scikit-learn)

Optymalizacja Bayesowka
Rozkład apriori funkcji



(E. Brochu, V. M. Cora, N. de Freitas, 2010)

Optymalizacja Bayesowka

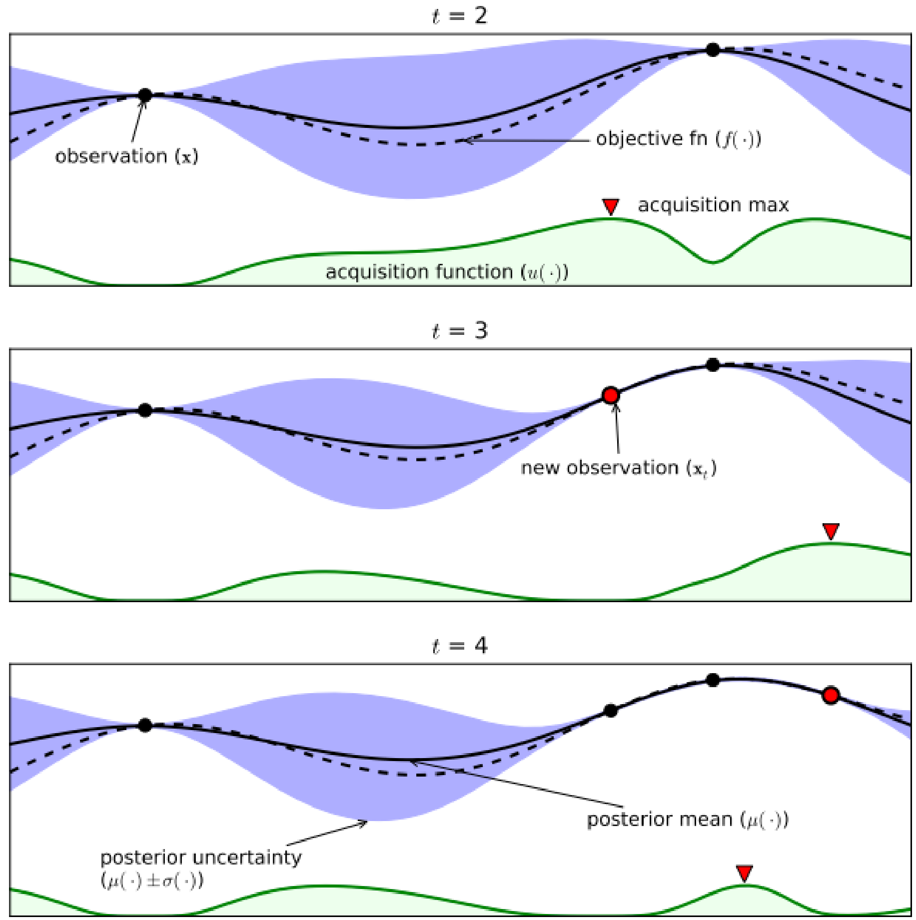
Algorytm

- Dla danego kroku t mamy zbiór danych:
  $D_t = \{(x_1, y_1), \dots, (x_t, y_t)\}$
- Optymalizujemy funkcję użyteczności dla rozkładu aposteriori:
  $x_{t+1} = \mathrm{argmax}_x\ u(f|D_t)$
- Obliczamy $y_{t+1} = f(x_{t+1})$
- $D_{t+1} = D_t + \{(x_{t+1}, y_{t+1})\}$

- Funkcja użyteczności: np. UCB = $\mu(x) + \mathrm{std}(x)$
- Rozkład aposteriori dla x_{t+1}
  $\mu(x_{t+1}) = k^T K^{-1} y$
  $\sigma^2(x_{t+1}) = k(x_{t+1}, x_{t+1}) - k^T K^{-1} k$

Optymalizacja Bayesowka

Algorytm

- 

(E. Brochu, V. M. Cora, N. de Freitas, 2010)

Optymalizacja Bayesowka

Pakiety w Pythonie

- Ogólnego przeznaczenia:
  - bayesian-optimization
    https://github.com/fmfn/BayesianOptimization
  - pyGPGO
    http://pygpgo.readthedocs.io/en/latest
  - hyperopt
    http://hyperopt.github.io/hyperopt/
- Automatyczne uczenie maszynowe – optymalizacja pipeline (preprocessing, wybór modelu i optymalizacja metaparametrów)
  - autosklearn
    https://github.com/automl/auto-sklearn

Literatura

- *Introduction to Multi-Armed Bandits*, A. Slivkins
- *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*, S. Bubeck, N. Cesa-Bianchi
- *Gambling in a rigged casino: The adversarial multi-armed bandit problem*, P. Auer, N. Cesa-Bianchi, Y. Freund, R. E. Schapire
- *An Empirical Evaluation of Thompson Sampling*, O. Chapelle, L. Li
- *A Tutorial on Thompson Sampling*, D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen

Literatura

- *A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit*, G. Burtini , J. Loeppky, R. Lawrence
- *A Survey on Contextual Multi-armed Bandits,*Li Zhou
- *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*, E. Brochu, V. M. Cora, N. de Freitas
- *Taking the Human Out of the Loop: A Review of Bayesian Optimization,* B. Shahriari, K. Swersky, Z. Wang, et. al.
- *Multi-Armed Bandit Algorithms and Empirical Evaluation*, J. Vermorel, M. Mohri
- *Algorithms for the multi-armed bandit problem*, V. Kuleshov, D. Precup
- https://dataorigami.net/blogs/napkin-folding/79031811-multi-armed-bandits