九州大学集中講義

深層学習および機械学習の数理

鈴木大慈
東京大学大学院情報理工学系研究科数理情報学専攻
理研AIP

2020年9月2日～4日

リッジ回帰

- カーネル法のアイディア：
  - 機械学習には「内積」が頻繁に現れる.
    → 内積を“工夫”すれば非線形解析ができるはず.

- 例：リッジ回帰（Tikhonov正則化）

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \|\beta\|_2^2$$

変数変換:

- 正則化項のため，$\hat{\beta} \in \mathrm{Ker}(X)^{\perp}$．つまり，$\hat{\beta} \in \mathrm{Im}(X^{\top})$.
- ある $\hat{\alpha} \in \mathbb{R}^n$ が存在して，$\hat{\beta} = X^{\top}\hat{\alpha}$ と書ける.

$$\hat{\alpha} \leftarrow \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \|XX^{\top}\alpha - Y\|_2^2 + \lambda_n \alpha^{\top}(XX^{\top})\alpha.$$

新しい入力xに対しては $y = x^{\top}X^{\top}\hat{\alpha}$ で予測.

- リッジ回帰の変数変換版

$$\hat{\alpha} \leftarrow \underset{\alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{n} \|XX^\top \alpha - Y\|_2^2 + \lambda_n \alpha^\top (XX^\top) \alpha.$$

※ $(XX^\top)_{i,j} = x_i^\top x_j$ はx_iとx_jの内積.

- **カーネル法のアイディア**

  x の間の内積を他の関数で置き換える:

$$x_i^\top x_j \;\rightarrow\; k(x_i, x_j)$$

この $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ をカーネル関数と呼ぶ.

カーネル関数の満たすべき条件

- 対称性：$k(x, x') = k(x', x)$
- 正値性：$\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j k(x_i, x_j) \geq 0, \quad \forall(\{x_i\}_{i=1}^m, \{\alpha_i\}_{i=1}^m, m)$

この条件さえ満たしていればなんでも良い

カーネルリッジ回帰

カーネルリッジ回帰: $K = (k(x_i, x_j))_{i,j=1}^n$ として,

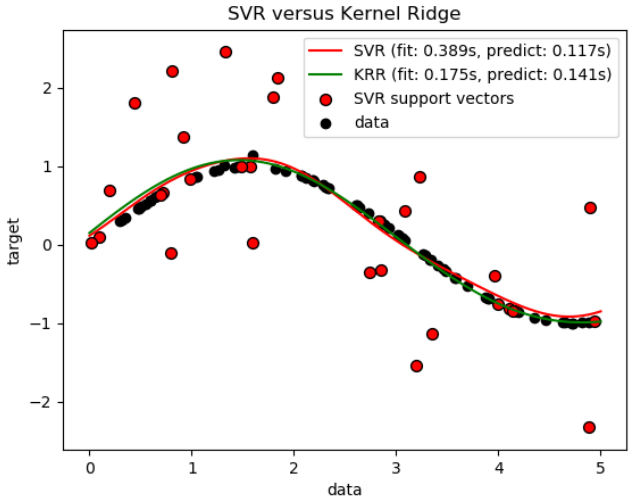
$$\hat{\alpha} \leftarrow \underset{\alpha \in \mathbb{R}^n}{\arg\min} \frac{1}{n} \|K\alpha - Y\|_2^2 + \lambda_n \alpha^\top K \alpha.$$

新しい入力 x に対しては,

$$y = \sum_{i=1}^{n} k(x, x_i)\hat{\alpha}_i$$

で予測.

線形代数で
非線形な回帰を実現.



[https://scikit-learn.org/stable/auto_examples/plot_kernel_ridge_regression.html]

カーネル関数の例

- ガウシアンカーネル

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

- 多項式カーネル

$$k(x, x') = \left(1 + x^\top x'\right)^p$$

- χ^2-カーネル

$$k(x, x') = \exp\left(-\gamma^2 \sum_{j=1}^{d} \frac{(x_j - x'_j)^2}{(x_j + x'_j)}\right)$$

- Matérn-kernel

$$k(x, x') = \int_{\mathbb{R}^d} e^{i\lambda^\top (x - x')} \frac{1}{(1 + \|\lambda\|^2)^{\alpha + d/2}} d\lambda$$

- グラフカーネル，時系列カーネル，...
(ユークリッド空間でなくてもカーネルが定義できればカーネル法が使える)

再生核ヒルベルト空間

カーネル関数 $k(x, x')$ ⇔ 再生核ヒルベルト空間 (RKHS) $\mathcal{H}_k$

定義

集合Ω上の再生核ヒルベルト空間（Reproducing kernel Hilbert space, RKHS）$\mathcal{H}$とは，Ω 上の関数からなるヒルベルト空間であって，任意の$x \in \Omega$ に対し$\phi_x \in \mathcal{H}$が存在し，

$$f(x) = \langle \phi_x, f \rangle_{\mathcal{H}} \quad (f \in \mathcal{H})$$

を満たすものをいう．

- $k(x, y) := \langle \phi_x, \phi_y \rangle_{\mathcal{H}}$は正定値対称カーネル関数
- 逆に正定値対称カーネルが与えられたら対応するRKHSが一意に存在

定理 (Moore-Aronszajnの定理)

$k(x, y)$: 正定値対称カーネル (given)

Ω上の関数からなるヒルベルト空間$\mathcal{H}_k$で以下の条件を満たすものが一意に存在：

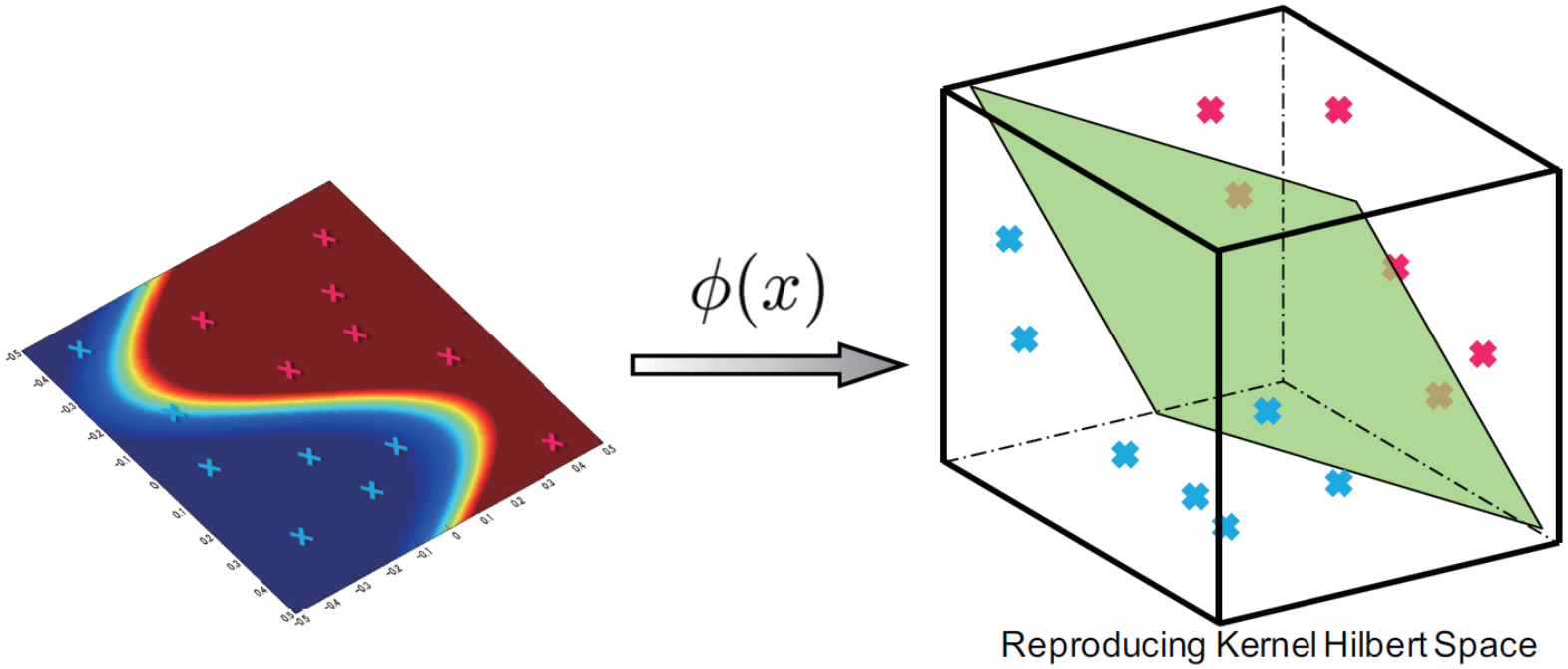
1. $k(x,\cdot) \in \mathcal{H}_k$
2. $f = \sum_{i=1}^{n} \alpha_i k(x_i,\cdot)$なる有限和は$\mathcal{H}_k$内で稠密
3. 再生成：

$$f(x) = \langle k(x,\cdot), f \rangle_{\mathcal{H}_k} \ (\forall x \in \Omega, \forall f \in \mathcal{H}_k).$$

$$f(x) = \left\langle \sum_{i=1}^{n} \alpha_i k(x_i,\cdot), k(x,\cdot) \right\rangle_{\mathcal{H}_k} = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

再生核ヒルベルト空間のイメージ

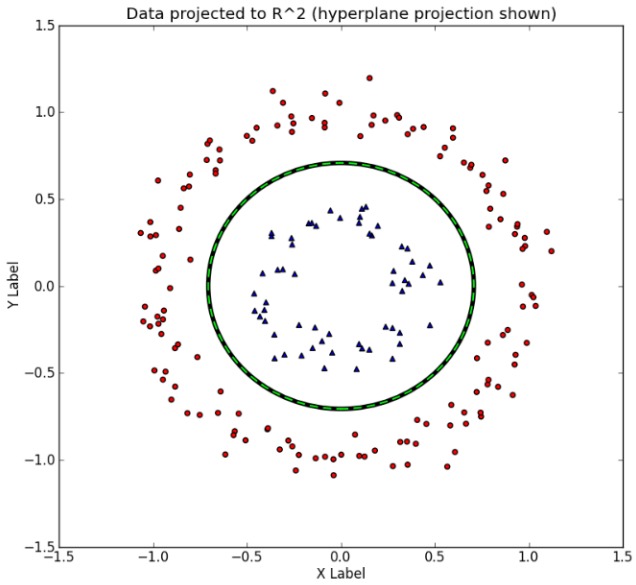
- 高次元(無限次元)特徴空間にϕで写像して推論を行う.
- 再生核ヒルベルト空間では線形な処理が元の空間では非線形処理になる.

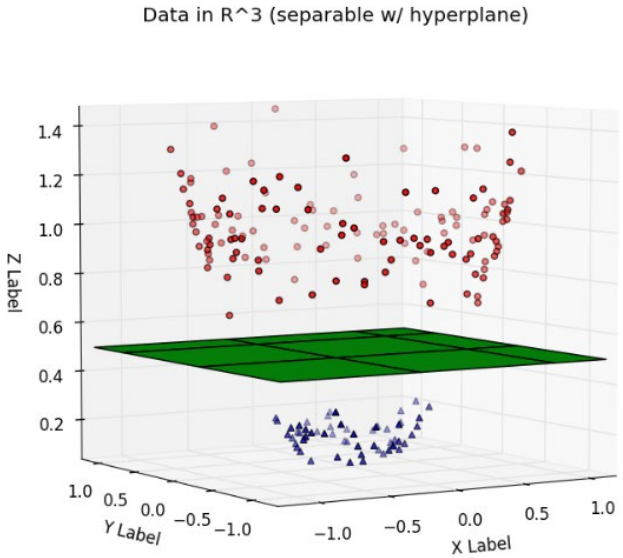


$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

多項式カーネル

非線形写像 ϕ_x





http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html

カーネルリッジ回帰の再定式化

- 再生性: $f \in \mathcal{H}_k$ に対し

$$f(x) = \langle f, \phi(x) \rangle_{\mathcal{H}_k}.$$

- カーネルリッジ回帰

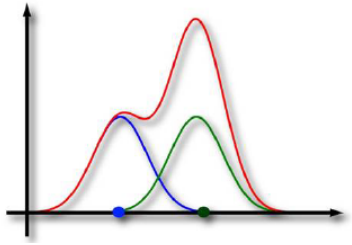
$$\hat{f} \leftarrow \min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + C\|f\|^2_{\mathcal{H}_k}$$

- 表現定理

$$\exists \alpha_i \in \mathbb{R} \quad \text{s.t.} \quad \hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x),$$

$$\Rightarrow \quad \|\hat{f}\|_{\mathcal{H}_k} = \sqrt{\sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)} = \sqrt{\boldsymbol{\alpha}^\top K \boldsymbol{\alpha}}.$$

さきほどのカーネルリッジ回帰の定式化と一致.



一般化

- 回帰だけでなく，判別問題などにも応用可

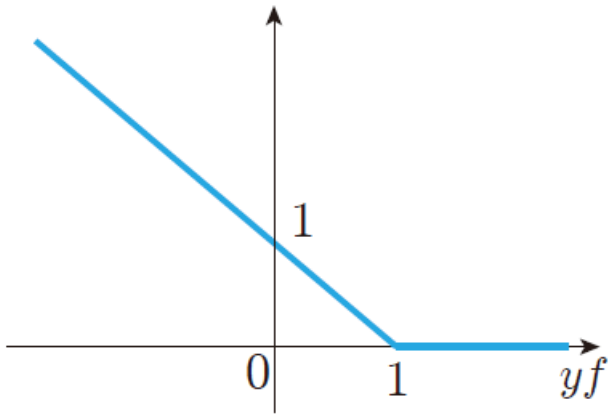
$$\frac{1}{n}\sum_{i=1}^{n}\ell\left(y_i,\sum_{j=1}^{n}k(x_i,x_j)\alpha_j\right)+\frac{C}{2}\alpha^\top K\alpha$$

- 教師なし学習にも応用されている
  - カーネルPCA，カーネルCCA，カーネルICA，...
  - 分布埋め込み

$$P\mapsto\mathbb{E}_{X\sim P}[k(\cdot,X)]\ \text{が単射}$$

サポートベクトルマシン

$y_i \in \{\pm 1\}$: 二値判別
$\ell(y, f) = \max\{-yf + 1, 0\}$: ヒンジ損失
　要番 (ちょうつがい) のような形をしているので「ヒンジ」損失



このとき，カーネル法はサポートベクトルマシン **(SVM)** と呼ばれる．

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \max\left\{1 - y_i\left(\sum_{j=1}^{n} k(x_i, x_j)\alpha_j\right), 0\right\} + \frac{C}{2}\alpha^\top K\alpha.$$

双対問題

$K_{ij} = k(x_i, x_j)$ とする．

SVM の双対問題

$$\min_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \max\left\{1 - y_i\left(\sum_{j=1}^{n} K_{ij}\alpha_j\right), 0\right\} + \frac{C}{2}\alpha^\top K\alpha$$

$$\Leftrightarrow \quad \min_{\beta \in \mathbb{R}^n} \quad -\sum_{i=1}^{n}\beta_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\beta_i\beta_j(y_i y_j K_{ij}) \quad \text{s.t.} \quad 0 \le \beta_i \le \frac{1}{C}.$$

・双対問題は座標ごとに自由に動きやすい→座標降下法が使える．
・主問題の最適解 α^*，双対問題の最適解 β^* に対し最適性条件（KKT 条件）

$$y_i\alpha_i^* = \beta_i^*$$
$$(K\alpha^*)_i y_i \ge 1 \quad \Rightarrow \quad \beta_i^* = \alpha_i^* = 0$$

なる関係が成り立つ．第二式から，

「正しく判別できているサンプル (x_i, y_i) に対応する係数 α_i^* は 0」

が言える．そのほかの $\alpha_i^* \neq 0$ なるサンプル点 x_i をサポートベクトルと呼ぶ．

カーネル関数の表現力

- Universal kernel

  $C_0(\mathbb{R}^d)$ を$\mathbb{R}^d$上の連続関数fで無限遠点で消える関数の集合とする:
  $\forall \epsilon > 0, \{x \mid f(x) \geq \epsilon\}$がコンパクト.

  あるカーネルに対し，そのRKHSが$C_0(\mathbb{R}^d)$内で一様ノルムに関して稠密な時，そのカーネルは「c_0-universal」であるという.
  （$\forall f \in C(X), \forall \epsilon > 0, \exists g \in \mathcal{H}_k$, s.t. $\|f - g\|_\infty < \epsilon$）

$k(x, y) = \psi(x - y)$と，ある有界連続なψを用いて書けるとき（平行移動不変）

kが正定値対称	$\Leftrightarrow$	ある有限非負測度Λが存在して以下のように書ける $\psi(x) = \int_{\mathbb{R}^d} e^{-\mathrm{i}w^\top x} \mathrm{d}\Lambda(w)$ (Bochner)
kがc_0-univeral	$\Leftrightarrow$	Λのサポートが全域: $\mathrm{supp}(\Lambda) = \mathbb{R}^d$.

ガウスカーネル，ラプラスカーネル，Maternカーネルなどはこれを満たす.
多項式カーネルはuniversalではない.

[Sriperumbudur, Fukumizu, Lanckriet: Universality, Characteristic Kernels and RKHS Embedding of Measures, ICML2010]

ガウス過程

ベイズ推定量

パラメータ θ からデータ D_n が出る確率(密度)

データの生成過程を表すパラメータ

$$p(\theta|D_n) = \frac{p(D_n|\theta) \times \pi(\theta)}{\int p(D_n|\theta) \times \pi(\theta)\mathrm{d}\theta}$$

パラメータの事前知識
(事前分布)

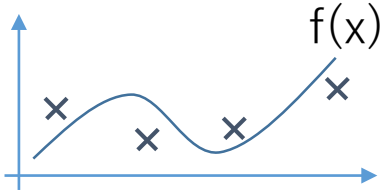
$$\text{posterior} = \frac{(\text{likelihood}) \times (\text{prior})}{\text{normalization constant}}$$

ガウシアンプロセス回帰

$$y = f(x) + \varepsilon$$

正規分布

$$p(f|D_n) = \frac{p(D_n|f)\Pi(f)}{\int p(D_n|f)\Pi(\mathrm{d}f)}$$



$$p(D_n|f) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - f(x_i))^2}{2\sigma^2}\right)$$

Π として**ガウシアンプロセス事前分布**を用いる.

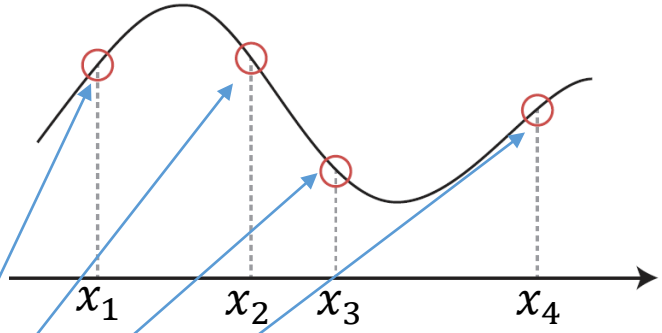
ガウシアンプロセス

- ガウシアンプロセス

関数$x \mapsto f(x)$の確率分布
（無限次元関数空間上の分布：ノンパラメトリック）

$$f \sim \mathrm{GP}$$

任意の有限個の入力に対する関数値$(f(x_1), f(x_2), \ldots, f(x_m))$が多変量正規分布に従う.



$$\begin{bmatrix} f(x_1) \\ f(x_2) \\ f(x_3) \\ f(x_4) \end{bmatrix} \sim N\left(\begin{bmatrix} \mu(x_1) \\ \mu(x_2) \\ \mu(x_3) \\ \mu(x_4) \end{bmatrix}, \begin{bmatrix} k(x_1,x_1) & k(x_1,x_2) & k(x_1,x_3) & k(x_1,x_4) \\ k(x_2,x_1) & k(x_2,x_2) & k(x_2,x_3) & k(x_2,x_4) \\ k(x_3,x_1) & k(x_3,x_2) & k(x_3,x_3) & k(x_3,x_4) \\ k(x_4,x_1) & k(x_4,x_2) & k(x_4,x_3) & k(x_4,x_4) \end{bmatrix} \right)$$

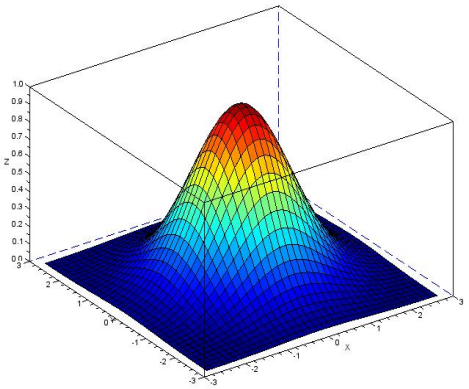
平均　　分散共分散

カーネル関数：$k(x, x') := \mathrm{Cov}_{f \sim \mathrm{GP}}(f(x), f(x'))$

代表的カーネル関数

- ガウスカーネル

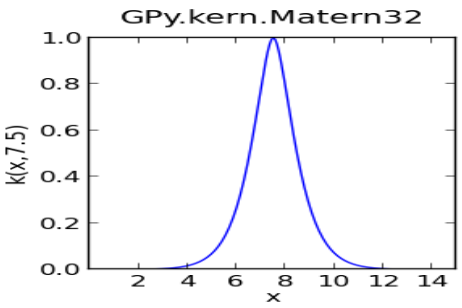
$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$



- Maternカーネル

$$k(x, x') = C_\nu(||x - x'||)$$

ただし，　$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\rho}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{d}{\rho}\right)$



ν は自然数，K_νは第2種変形ベッセル関数
Maternカーネルに対応するGPはν回微分可能な関数を生成

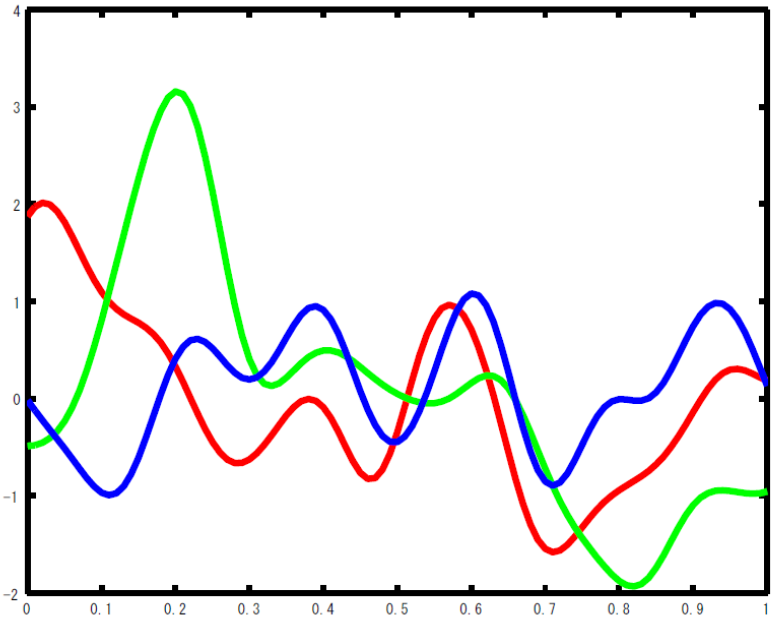
- 多項式カーネル

$$k(x, x') = \left(1 + x^\top x'\right)^p$$

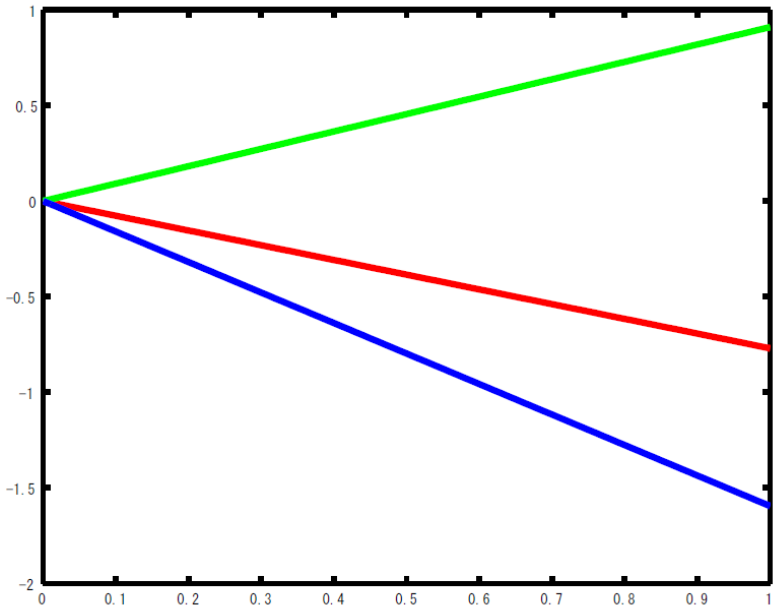
- 線形カーネル

$$k(x, x') = x^\top x'$$

ガウシアンプロセスからのサンプル



(d) Gaussian kernel



(e) Linear kernel

事後分布

データ$\{(x_i, y_i)\}_{i=1}^n, Y = [y_1, \ldots, y_n]^\top$からの事後分布

$$\mathbf{f} = (f(x_1), \ldots, f(x_n))$$ 観測データ点上での関数値

尤度関数（データへの当てはまり）　事前分布

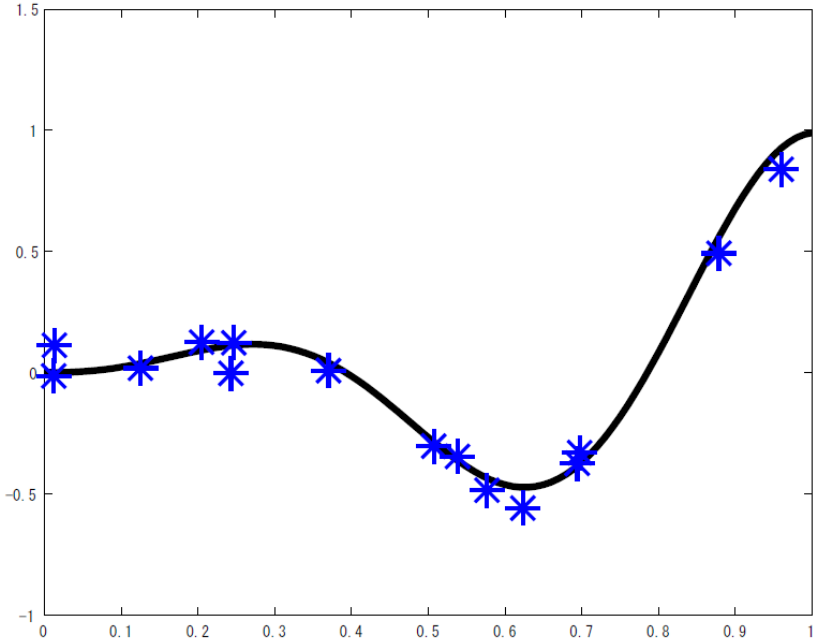
$$p(\mathbf{f}|D_n) = \frac{1}{C} \exp\left(-n\frac{\|\mathbf{f} - Y\|_n^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f}\right)$$
$$= \frac{1}{C} \exp\left(-\frac{1}{2}\left\|\mathbf{f} - (K + \sigma^2 I_n)^{-1} KY\right\|^2_{(K^{-1} + I_n/\sigma^2)}\right)$$

関数値の事後分布

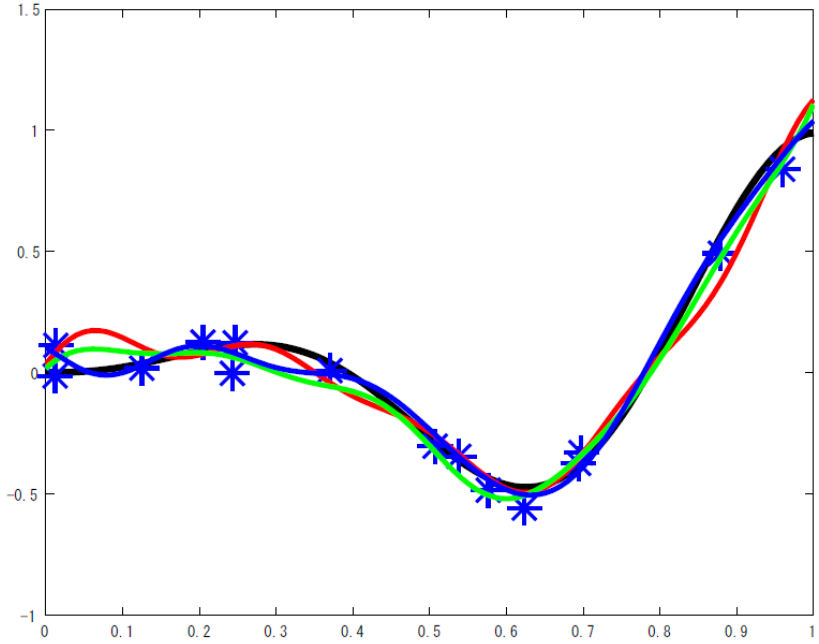
- posterior mean: $\hat{\mathbf{f}} = K(K + \sigma^2 I_n)^{-1} Y.$ 関数値の予測値
  事後分布の平均
- posterior covariance: $K - K(K + \sigma^2 I_n)^{-1} K.$ 関数値の信用区間（自信）
  事後分布の分散

観測データ以外の点における関数値も予測できる.

事後分布からのサンプル

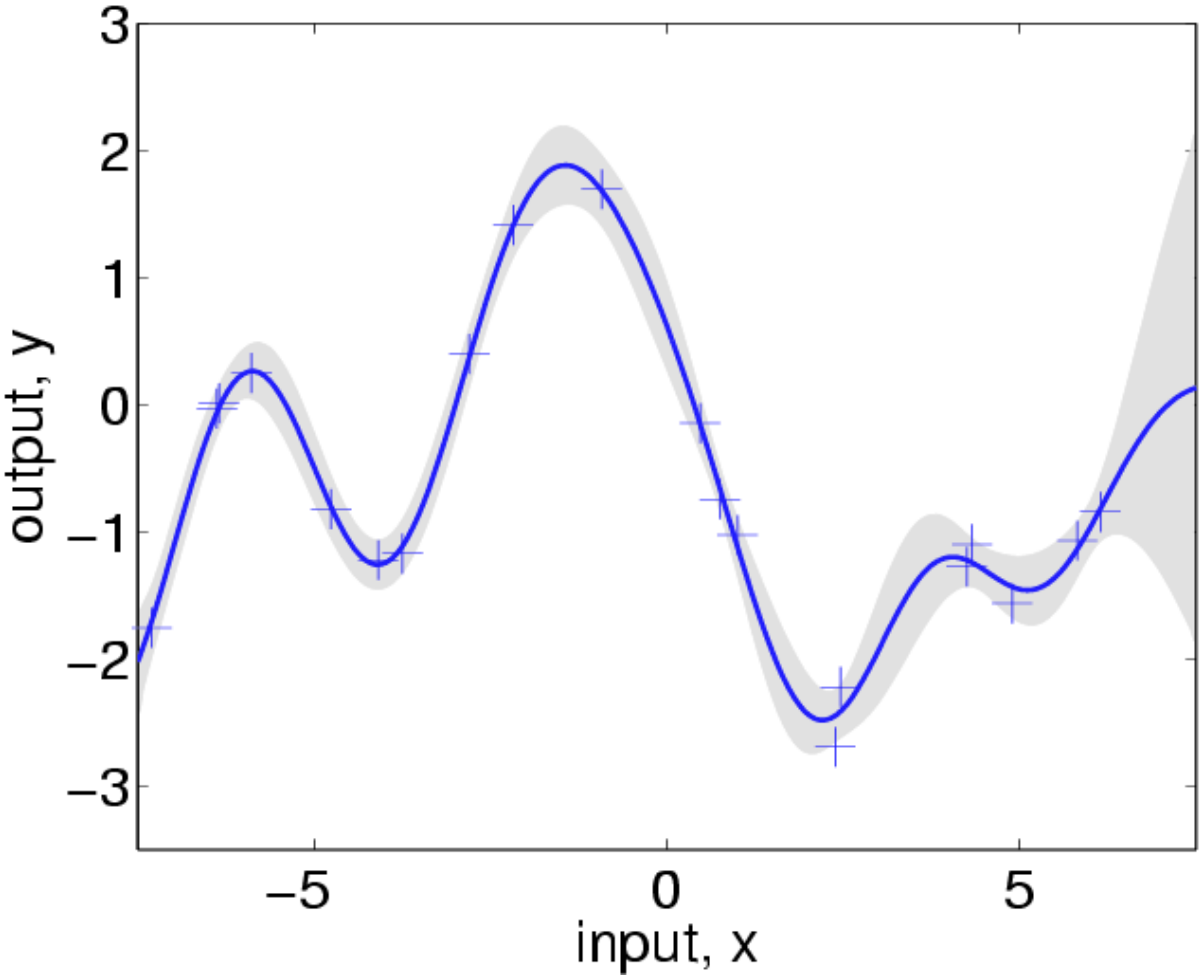


(f) Training Data



(g) Posterior Sample

事後分布の信用区間



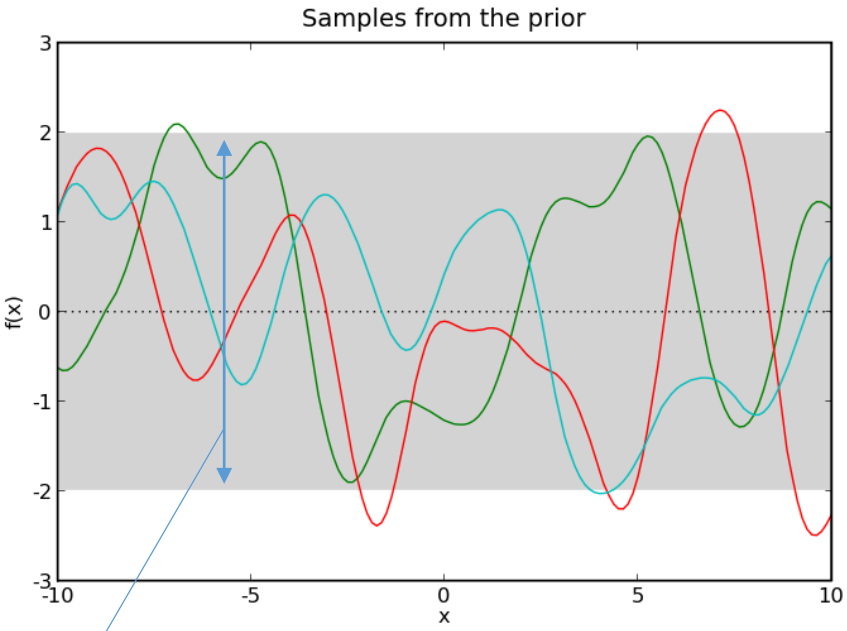
その他の例

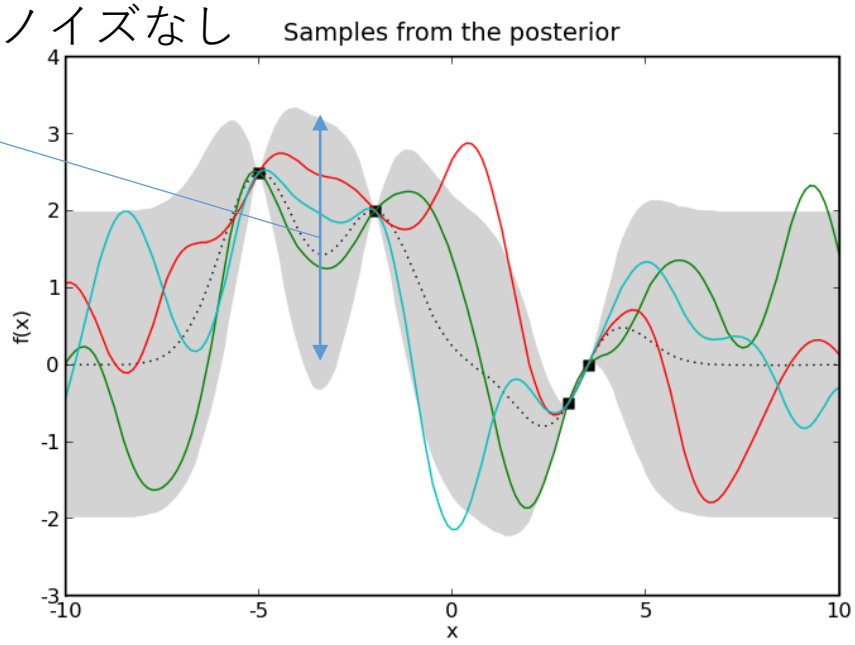
事後分布

ノイズなし

信用区間

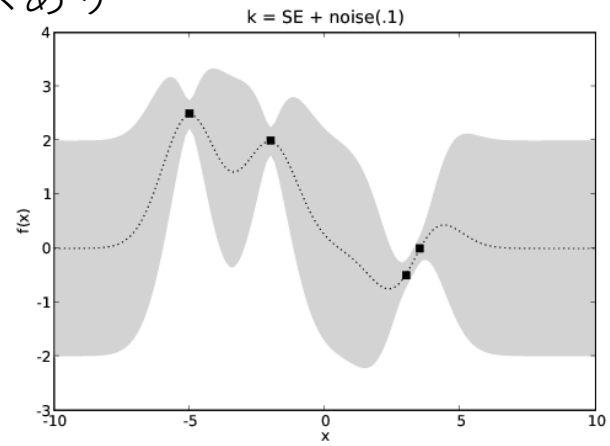
事前分布





信用区間

ノイズあり



MMDと特性的カーネル

MMD

$\phi_k : \mathbb{R}^d \to \mathcal{H}_k$ ：あるRKHS ($\mathcal{H}_k$) への特徴写像

P, P' ：二つの分布

$x \mapsto k(x, \cdot) \in \mathcal{H}_k$ のこと

$$\mathrm{MMD}(P, P') := \left\| \mathbb{E}_P[\phi_k(X)] - \mathbb{E}_{P'}[\phi_k(X)] \right\|_{\mathcal{H}_k}$$

分布を$\mathbb{E}_P[\phi_k(X)]$でRKHS内に埋め込み，そこで距離を測る．

$$\mathrm{MMD}^2 = \mathbb{E}_{X_1, X_2 \sim P,\ X_1', X_2' \sim P'}[k(X_1, X_2) - 2k(X_1, X_1') + k(X_1', X_2')]$$

⇨ データから推定できる

定理

$\mathbb{R}^d \times \mathbb{R}^d$上のカーネル関数が連続かつ特性的 (後述)

⇔　MMDが弱収束位相を距離付けする．

[Simon-Gabriel&Scholkopf: Kernel Distribution Embeddings: Universal Kernels, Characteristic Kernels and Kernel Metrics on Distributions, JMLR2018.]

特性的カーネル

- 特性的カーネル

$$P \mapsto \int k(\cdot, x)\mathrm{d}P(x) \text{ が} \mathbb{R}^d \text{上ボレル確率測度に対して単射}$$

つまり，MMD(P,P')=0がP=P'と同値の時，そのカーネルを特性的と言う．

$M_b(\mathbb{R}^d)$ を$\mathbb{R}^d$上の有限な符号付ボレル測度の集合とする．

カーネル関数kがc_0-universal

$\Leftrightarrow$

$\mu \mapsto \int k(\cdot, x)\mathrm{d}\mu(x) \quad (\mu \in M_b(\mathbb{R}^d))$ が単射．

確率測度に限定したら必ずしも同値ではない．しかし，次が成り立つ．

カーネルkが平行移動不変で$k(x, y) = \phi(x - y)$ $(\phi \in C_0(\mathbb{R}^d))$と書けるとき，

「c_0-universal」 $\Leftrightarrow$ 「特性的」

[Sriperumbudur, Fukumizu, Lanckriet: Universality, Characteristic Kernels and RKHS Embedding of Measures, ICML2010]

応用

- 二標本検定

$$x_1, \dots, x_n \sim P$$
$$y_1, \dots, y_m \sim Q$$

PとQは同じ分布か？

[A. Gretton, et al. A fast, consistent kernel two-sample test. NIPS2009 (2009).]

- 独立性検定：HSIC

$$(x_1, y_1), \dots, (x_n, y_n) \sim P_{XY}$$

XとYは独立か？ ⇒ $\mathrm{MMD}(P_{XY}, P_X P_Y) = 0$の検定

[A. Gretton, et al. Measuring Statistical Dependence with Hilbert-Schmidt Norms. ALT2005 (2005).]

- GANへの応用：MMD-GAN

[C-L Li, et al. MMD GAN: Towards deeper understanding of moment matching network. NeurIPS2017 (2017)].

Generatorの生成する分布と訓練データの分布をMMDで比較して，
MMDを最小化するようにGeneratorを学習．

MMDとその仲間

- 積分型確率測度距離

$$D(P||Q) = \sup_{f \in \mathcal{F}} \mathrm{E}_P[f] - \mathrm{E}_Q[f]$$

- f(x)としてf(x)=xのみを用いれば平均値の差を見ていることになる.
- f(x)として，f(x)=xおよびf(x)=x^2も考えれば二次モーメントの差も考慮できる.
- Fとしてもっと広い関数の集合を考えれば分布の“距離”になる.

- 1-Wasserstein距離

  $\mathcal{F}$ =1-リプシッツ連続な関数の集合

$$\sup_{x,y} \frac{|f(x) - f(y)|}{\|x - y\|} \leq 1$$

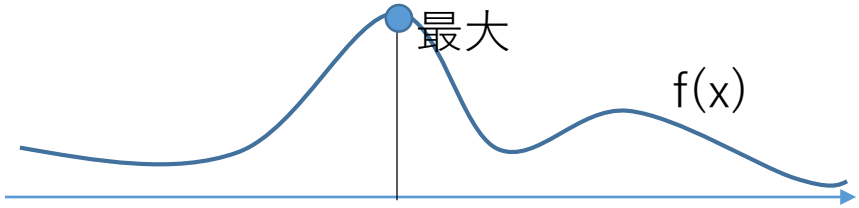
- MMD (Maximum Mean Discrepancy)

  $\mathcal{F}$ =ある再生核ヒルベルト空間の単位球

ベイズ最適化
（参考）

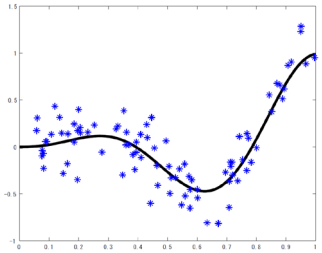
ベイズ最適化の概要

設定：関数f(x)の最大化をしたい.
しかし，関数値を求めるのにコストがかかる.
なるべく少ない回数で最大値に到達したい.



ベイズ推定量

「関数の概形を推定」



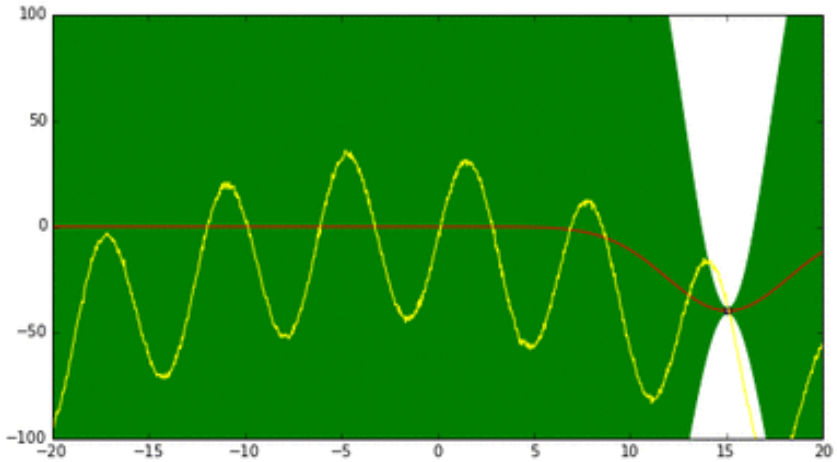
↓

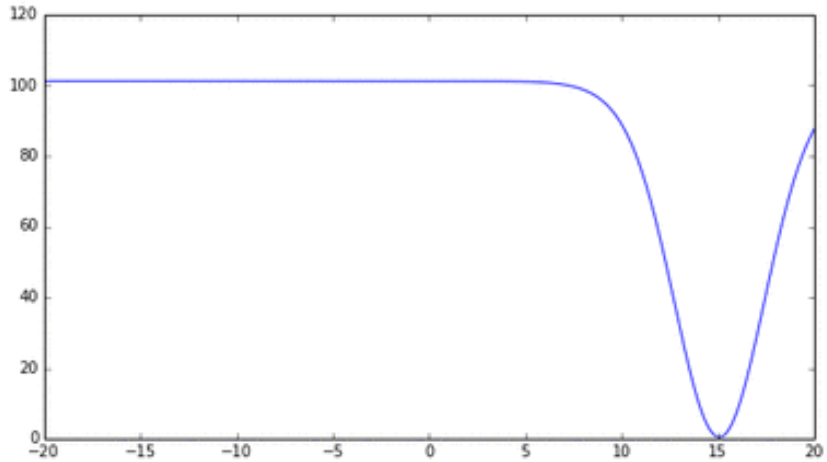
ベイズ最適化

「ベイズ推定量を利用して適切な入力点xを選択」

※ベイズ推定は必ずしもベイズ最適化をする方法ではありません.

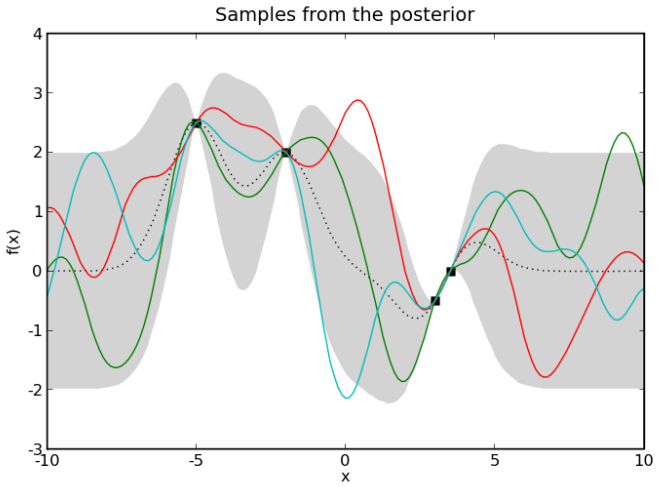
大体の動作


100
50
0
−50
−100
−20
−15
−10
−5
0
5
10
15
20


120
100
80
60
40
20
0
−20
−15
−10
−5
0
5
10
15
20

手順の概要

- 入力点xを適切に選択
  - 一番よさそうな点を選ぶ
  - UCB戦略，Expected Improvement, Thompson sampling
- 関数値をベイズ推定（GP回帰）



- 以上を繰り返す.

基本戦略

- 獲得関数 (acquisition function)

$$x_{t+1} = \mathrm{argmax}_x \, a_t(x)$$

毎回獲得関数を最大にする点を選ぶ

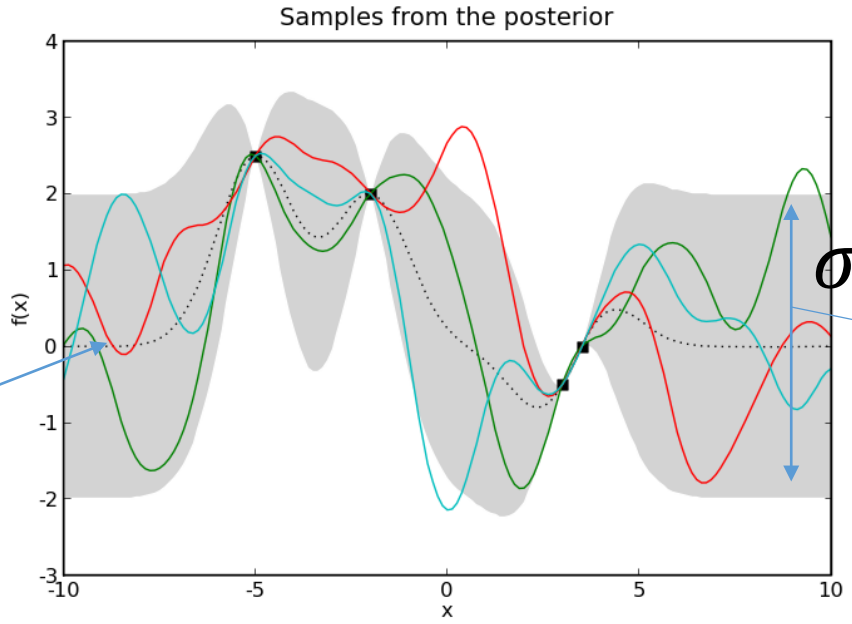
- ベイズ最適化の諸手法はこの獲得関数の違い

UCB戦略

信用区間の最大値を選ぶ

$$a_t(x) = \mu_t(x) + \gamma\, \sigma_t(x)$$

γは上側何％点を用いるかを決定.
γが大きいほどより保守的.



$\sigma_t(x)$

信用区間

$\mu_t(x)$

平均

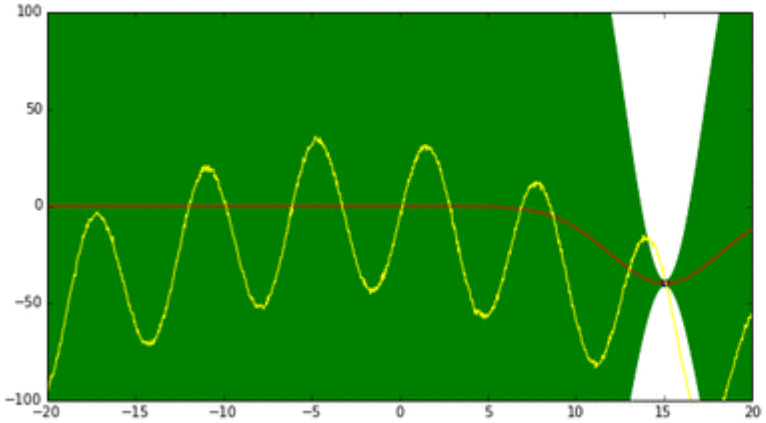
Expected Improvement

- 改善量の期待値が最大である場所を探す

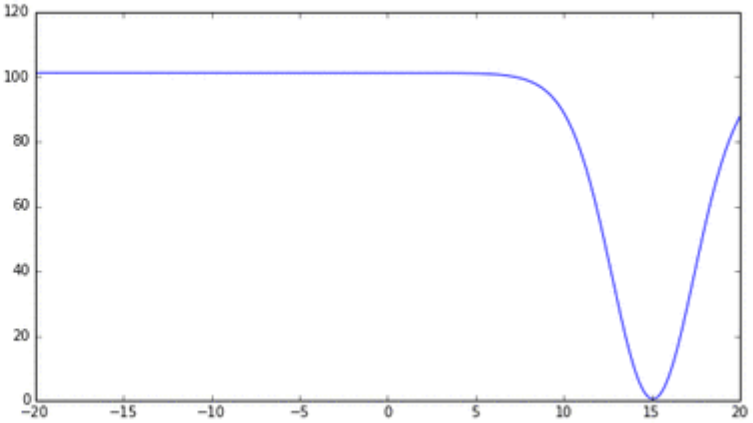
これまでの最大値

$$a_t(x) = \int \max\{0, f(x) - f(x_t^{best})\}\, \mathrm{GP}\left(df \mid \{(x_i, y_i)\}_{i=1}^{t}\right)$$

$$= \left(\mu_t(x) - f\left(x_t^{best}\right)\right) \Phi\left(\frac{\mu_t(x) - f\left(x_t^{best}\right)}{\sigma_t(x)}\right) + \sigma_t(x)\phi\left(\frac{\mu_t(x) - f\left(x_t^{best}\right)}{\sigma_t(x)}\right)$$



事後分布の様子



獲得関数の様子

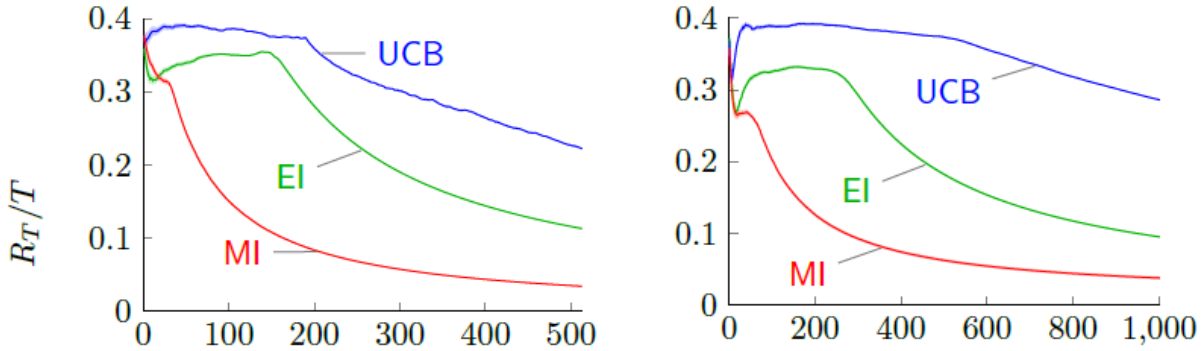
Mutual Information

[Contal, Perchet, Vayatis:ICML20

$$a_t(x) = \mu_t(x) + \gamma \left(\sqrt{\sigma_t^2(x) + \sum_{i=1}^{t-1} \sigma_i^2(x_i)} - \sqrt{\sum_{i=1}^{t-1} \sigma_i^2(x_i)} \right)$$

点xを選ぶことで得られる情報量

理論上UCBよりも効率的（次元への依存性が良い）.



(g) Tsunamis　　(h) Mackey-Glass

実験的にも良い

理論

- 誤差

$$f(x^*) - \max_t f(x_t)$$

ガウシアンカーネルで表現可能な関数 on d次元空間
(非常に滑らか)

- UCB戦略：$\log(t)^{d+1}/t$
- MI戦略：$\log(t)^{(d+1)/2}/t$

(ノイズありの設定)

[Contal, Perchet, Vayatis:ICML2014]

ν回微分可能な関数 on d次元空間

- EI戦略：$(t/\log(t))^{-\frac{\nu}{d}}\left(\log(t)\right)^{1/2}$　　(ノイズなしの設定)

アルゴリズムの途中でランダムに入力点を選ばなければこのレートは達成できない．→セレンディピティ？

[Bull, JMLR, 2011]