カーネル法の数理

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
数理第六研究室

Outline

線形回帰

デザイン行列 $X = (X_{ij}) \in \mathbb{R}^{n \times p}$.　　$Y = [y_1, \dots, y_n]^\top \in \mathbb{R}^n$.
真のベクトル $\beta^* \in \mathbb{R}^p$:

$$\text{モデル}: \quad Y = X\beta^* + \xi.$$

リッジ回帰（Tsykonov 正則化）

$$\hat{\beta} \leftarrow \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \frac{1}{n}\|X\beta - Y\|_2^2 + \lambda_n \|\beta\|_2^2.$$

線形回帰

デザイン行列 $X = (X_{ij}) \in \mathbb{R}^{n \times p}$.　　$Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$.
真のベクトル $\beta^* \in \mathbb{R}^p$:

$$\text{モデル}: \quad Y = X\beta^* + \xi.$$

リッジ回帰（Tsykonov 正則化）

$$\hat{\beta} \leftarrow \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \frac{1}{n}\|X\beta - Y\|_2^2 + \lambda_n \|\beta\|_2^2.$$

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．

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

※ $(XX^\top)_{ij} = x_i^\top x_j$ より，観測値 x_i と x_j の<u>**内積**</u>さえ計算できればよい．

リッジ回帰のカーネル化

リッジ回帰（変数変換版）

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

※ $(XX^\top)_{ij} = x_i^\top x_j$ はサンプル x_i と x_j の**内積**.

リッジ回帰のカーネル化

リッジ回帰（変数変換版）

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

※ $(XX^\top)_{ij} = x_i^\top x_j$ はサンプル x_i と x_j の**内積**.

- **カーネル法のアイディア**
  x の間の内積を他の非線形な関数で置き換える:

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

この $k : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ を**カーネル関数**と呼ぶ.

Definition (正定値カーネル)

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

逆にこの性質を満たす関数なら何でもカーネル法で用いて良い.

カーネルリッジ回帰

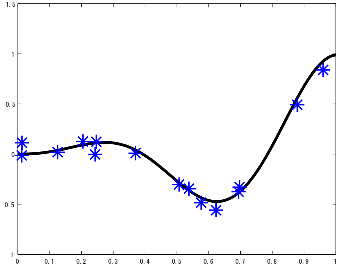
カーネルリッジ回帰: $K = (k(x_i, x_j))_{i,j=1}^{n}$ として,

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

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.



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,

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

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.

カーネル関数 $\Leftrightarrow$ **再生核ヒルベルト空間 (RKHS)**

$k(x, x')$ $\qquad$ $\mathcal{H}_k$

ある $\phi(x) : \mathbb{R}^p \to \mathcal{H}_k$ が存在して,

- $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_k}$.
- カーネルトリック: $\langle \sum_{i=1}^n \alpha_i \phi(x_i), \phi(x) \rangle_{\mathcal{H}_k} = \sum_{i=1}^n \alpha_i k(x_i, x)$.

  $\to$カーネル関数の値さえ計算できれば良い.

再生核ヒルベルト空間 (Reproducing Kernel Hilbert Space, RKHS)

Definition (再生核ヒルベルト空間 (RKHS))

集合 $\mathcal{X}$ 上の（実数値）関数からなるヒルベルト空間 $\mathcal{H}$ が再生核ヒルベルト空間であるとは，$\forall x \in \mathcal{X}$ に対して $k_x \in \mathcal{H}$ が存在して

$$\text{再生性：} \quad \langle f, k_x \rangle_{\mathcal{H}} = f(x) \quad (\forall f \in \mathcal{H})$$

が成り立つこととする．

特に，$k(x, y) = \langle k_x, k_y \rangle_{\mathcal{H}}$ は正定値カーネルであり，$\mathcal{H}$ に付随した**再生核 (Reproducing kernel)** とよぶ．

Theorem (Moore-Aronszajn (Aronszajn, 1950))

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ は正定値カーネル．
$\Rightarrow$ k を再生核とする再生核ヒルベルト空間が一意的に存在する．

再生核ヒルベルト空間の表現

1. 入力データの分布：P_X
2. 対応する L_2 空間：$L_2(P_X) = \{f \mid \mathbb{E}_{X\sim P_X}[f(X)^2] < \infty\}$ (可分とする).

$\int k(x,x)\mathrm{d}P_X(x) < \infty$ なら，カーネル関数は以下のように分解できる (Steinwart and Scovel, 2012):

$$k(x,x') = \sum_{j=1}^{\infty} \mu_j e_j(x) e_j(x') \qquad (P_X \times P_X\text{-a.s.}).$$

- $(e_j)_{j=1}^{\infty}$ は $L_2(P_X)$ 内の正規直交基底: $\|e_j\|_{L_2(P_X)} = 1$, $\langle e_j, e_{j'}\rangle_{L_2(P_X)} = 0$ $(j \neq j')$.
- $\mu_j \geq 0$.

Theorem (再生核ヒルベルト空間 ($\mathcal{H}_k$) の表現)

- $\langle f, g\rangle_{\mathcal{H}_k} := \sum_{j=1}^{\infty} \frac{1}{\mu_j}\alpha_j\beta_j$ *for* $f = \sum_{j=1}^{\infty}\alpha_j e_j$, $g = \sum_{j=1}^{\infty}\beta_j e_j \in L_2(P_X)$.
- $\|f\|_{\mathcal{H}_k} := \sqrt{\langle f, f\rangle_{\mathcal{H}_k}}$.
- $\mathcal{H}_k := \{f \in L_2(P_X) \mid \|f\|_{\mathcal{H}_k} < \infty\}$ *equipped with* $\langle\cdot,\cdot\rangle_{\mathcal{H}_k}$.

再生性: $f \in \mathcal{H}_k$ に対して $f(x)$ は内積の形で「再生」される:

$$\langle f, k(x,\cdot)\rangle_{\mathcal{H}_k} = \sum_{j=1}^{\infty}\frac{1}{\mu_j}\alpha_j\mu_j e_j(x) = \sum_{j=1}^{\infty}\alpha_j e_j(x) = f(x).$$

再生核ヒルベルト空間の性質

カーネル関数に対応する積分作用素 $T_k : L_2(P_X) \to L_2(P_X)$:

$$T_k f := \int f(x) k(x, \cdot) \mathrm{d}P_X(x).$$

- 先のカーネル関数の分解は T_k のスペクトル分解に対応.
- 再生核ヒルベルト空間 $\mathcal{H}_k$ は以下のようにも書ける:

$$\mathcal{H}_k = T_k^{1/2} L_2(P_X).$$

再生核ヒルベルト空間の性質

カーネル関数に対応する積分作用素 $T_k : L_2(P_X) \to L_2(P_X)$:

$$T_k f := \int f(x)k(x,\cdot)\mathrm{d}P_X(x).$$

- 先のカーネル関数の分解は T_k のスペクトル分解に対応.
- 再生核ヒルベルト空間 $\mathcal{H}_k$ は以下のようにも書ける:

$$\mathcal{H}_k = T_k^{1/2} L_2(P_X).$$

- $\|f\|_{\mathcal{H}_k} = \inf\{\|h\|_{L_2(P_X)} \mid f = T_k^{1/2}h,\ h \in L_2(P_X)\}$.
  - $f \in \mathcal{H}_k$ を $f = T_k^{1/2} h$ for $h = \sum_j a_j e_j$ としたとき,

$$f(x) = \sum_{j=1}^{\infty} a_j \sqrt{\mu_j} e_j(x), \quad \|f\|_{\mathcal{H}_k} = \sqrt{\sum_{j=1}^{\infty} a_j^2} = \|h\|_{L_2(P_X)}.$$

  - $(e_j)_j$ は L_2 内の正規直交基底, $(\sqrt{\mu_j} e_j)_j$ は RKHS 内の完全正規直交基底.
  - $\phi_k(x) = k(x,\cdot) \in \mathcal{H}_k$ は**特徴写像**とみなせ, $\mathcal{H}_k$ の完全正規直交基底に関する係数で表現すると

$$\phi_k(x) = (\sqrt{\mu_1} e_1(x), \sqrt{\mu_2} e_2(x), \dots)^\top,$$
$$k(x,x') = \langle \phi_k(x), \phi_k(x') \rangle_{\mathcal{H}_k}.$$

カーネルリッジ回帰の再定式化

- カーネルリッジ回帰の再定式化

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

Theorem (表現定理)

最適解 $\hat{f} \in \mathcal{H}_k$ は n 個のカーネルの和で表現できる:

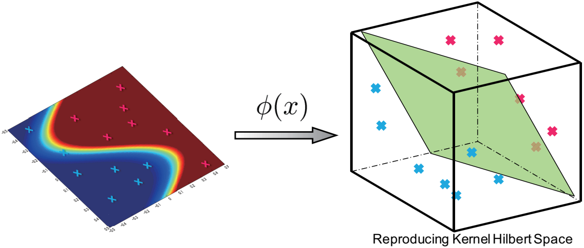
$$\exists \alpha_i \in \mathbb{R} \quad s.t. \quad \hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x).$$

(Proof) $f(x_i) = \langle f, k(x_i, \cdot)\rangle_{\mathcal{H}_k}$ より, $\hat{f}$ は $\mathrm{span}\{k(x_i, \cdot)\ (i \in [n])\} = \bar{\mathcal{H}}$ 上にいる. 実際, $f = \bar{f} + f_\perp$ ($f \in \bar{\mathcal{H}}$, $f_\perp \in \bar{\mathcal{H}}^\perp$) と分解すると, $\|f\|_{\mathcal{H}_k}^2 = \|\bar{f}\|_{\mathcal{H}_k}^2 + \|f_\perp\|_{\mathcal{H}_k}^2$ かつ $\langle f, k(x_i, \cdot)\rangle_{\mathcal{H}_k} = \langle \bar{f}, k(x_i, \cdot)\rangle_{\mathcal{H}_k}$ なので, $f_\perp = 0$ とした方が目的関数を小さくできる. □

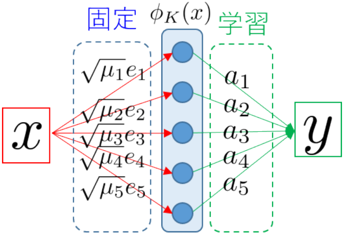
$\|\hat{f}\|_{\mathcal{H}_k}^2 = \sum_{i,j=1}^{n} \alpha_i\alpha_j k(x_i, x_j) = \boldsymbol{\alpha}^\top K \boldsymbol{\alpha}$ に注意すると,
さきほど天下り的に与えたカーネルリッジ回帰の定式化と一致.

再生核ヒルベルト空間のイメージ

- 非線形な推論を再生核ヒルベルト空間への非線形写像 ϕ を用いて行う.
- 再生核ヒルベルト空間では線形な処理をする.



- **カーネル法は第一層を固定し第二層目のパラメータを学習する横幅無限大の 2 層ニューラルネットワークともみなせる.**
  (“ 浅い ” 学習手法の代表例)



特徴写像を陽に用いたカーネルリッジ回帰

$$
\begin{aligned}
& \min_{f \in \mathcal{H}_k} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + C\|f\|_{\mathcal{H}_k}^2 \\
\Longleftrightarrow \quad & \min_{a \in \mathbb{R}^\infty} \quad \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{\infty} a_j \phi_j(x_i)\right)^2 + C\sum_{j=1}^{\infty} a_j^2
\end{aligned}
$$

ただし，$\phi_j(x) = \sqrt{\mu_j} e_j(x)$.

カーネルの例

- 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

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

- グラフカーネル，時系列カーネル，...

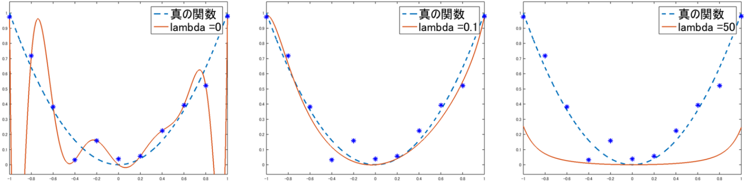
Outline

1. 再生核ヒルベルト空間の定義
2. カーネル法の推定精度
3. 再生核ヒルベルト空間における最適化

Example of (kernel) ridge regression

Polynomial regression （15th-order polynomial）

$$\min_{\theta \in \mathbb{R}^{15}} \frac{1}{n} \sum_{i=1}^{n} \{ y_i - (\theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_{15} x_i^{15}) \}^2 + \lambda \|\theta\|_2^2$$



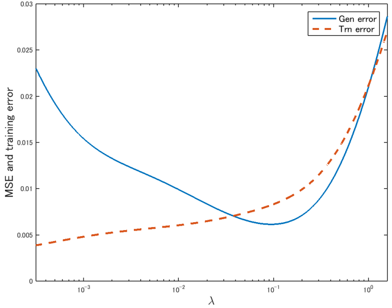
$\lambda = 0$ $\lambda = 0.1$ $\lambda = 50$

過学習 良い推定 過小学習

Regularization parameter and generalization error

Polynomial regression （15th-order polynomial）

$$\min_{\theta \in \mathbb{R}^{15}} \frac{1}{n} \sum_{i=1}^{n} \{y_i - (\theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_{15} x_i^{15})\}^2 + \lambda \|\theta\|_2^2$$



Horizontal axis: regularization parameter λ (log-scale).
Vertical axis: generalization error (blue), training error (red).

RKHSの「複雑さ」

- 積分作用素としての表現:

$$T_k : f \mapsto \int k(\cdot, x')f(x')\mathrm{d}P_X(x').$$

- カーネルの分解:

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j e_j(x) e_j(x'),$$

in $L_2(P(X) \times P(X))$.

Definition

再生核ヒルベルト空間の自由度 (degrees of freedom):

$$N_k(\lambda) := \mathrm{Tr}[(T_k + \lambda)^{-1} T_k] = \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda}.$$

$N_k(\lambda)$ は RKHS の「複雑さ」を計る.

カーネル法の推定精度

モデル:

$$y_i = f^{\mathrm{o}}(x_i) + \epsilon_i \quad (i = 1, \dots, n).$$

カーネルリッジ回帰:

$$\hat{f}_{\color{red}\lambda} = \operatorname*{argmin}_{\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + {\color{red}\lambda}\|f\|_{\mathcal{H}}^2.$$

Proposition (Caponnetto and de Vito (2007))

$f^{\mathrm{o}} \in \mathcal{H}$ であるなら,

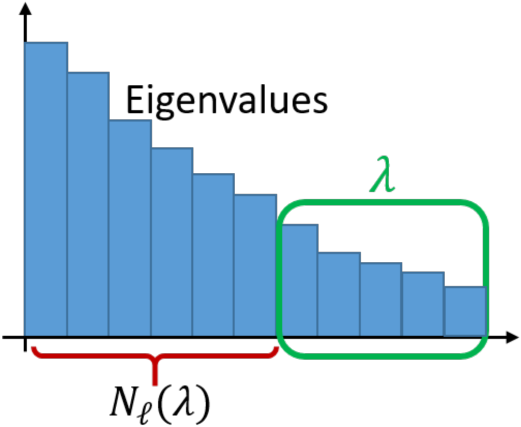
$$\|\hat{f}_\lambda - f^{\mathrm{o}}\|_{L_2(P_X)}^2 \leq C\Bigg(\underbrace{\lambda}_{\text{bias}} + \underbrace{\frac{N_k(\lambda)}{n}}_{\text{variance}}\Bigg),$$

が高い確率で成り立つ (**バイアス-バリアンスのトレードオフ**).

1. 基本的に $\frac{N_k(\lambda)}{n} = \lambda$ を満たす λ を選べば良い.
2. $\mu_i \lesssim i^{-\alpha}$ なら $N_k(\lambda) \lesssim \lambda^{-1/\alpha}$ で, $\lambda \simeq n^{-\frac{\alpha}{\alpha+1}}$ が最適:

$$\|\hat{f}_\lambda - f^{\mathrm{o}}\|_{L_2(P_X)}^2 \lesssim n^{-\frac{\alpha}{\alpha+1}}.$$

Minimax-optimalであることが知られている.



Rough sketch of $N_k(\lambda)$.

- Estimation error in $N_k(\lambda)$ dimensional space: $\frac{N_k(\lambda)}{n}$
- Bias (residual): λ

Outline

1. 再生核ヒルベルト空間の定義
2. カーネル法の推定精度
3. 再生核ヒルベルト空間における最適化

再生核ヒルベルト空間内の確率的最適化 (1)

問題設定:

$$y_i = f^{\mathrm{o}}(x_i) + \xi_i.$$

$(x_i, y_i)_{i=1}^n$ から f^{o} を推定したい．(f^{o} は $\mathcal{H}_k$ にほぼ入っている)

期待損失の変形:

$$\mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(f(X) - f^{\mathrm{o}}(X) - \xi)^2] = \mathbb{E}[(f(X) - f^{\mathrm{o}}(X))^2] + \sigma^2$$

$\rightarrow \min_{f \in \mathcal{H}_k} \mathbb{E}[(f(X) - Y)^2]$ を解けば f^{o} が求まる．

期待損失の Frechet 微分:

$K_x = k(x, \cdot) \in \mathcal{H}_k$ とする．$f(x) = \langle f, K_x \rangle_{\mathcal{H}_k}$ に気を付けると
$L(f) = \mathbb{E}[(f(X) - Y)^2] = \mathbb{E}[(\langle K_X, f \rangle_{\mathcal{H}_k} - Y)^2]$ の RKHS 内での Frechet 微分は以下の通り:

$$\begin{aligned}
\nabla L(f) &= 2\mathbb{E}[K_X(\langle K_X, f \rangle_{\mathcal{H}_k} - Y)] \\
&= 2(\underbrace{\mathbb{E}[K_X K_X^*]}_{=:\Sigma} f - \mathbb{E}[K_X Y]) \\
&= 2(\Sigma f - \mathbb{E}[K_X Y]).
\end{aligned}$$

再生核ヒルベルト空間内の確率的最適化 (2)

$L(f) = \mathbb{E}[(f(X) - Y)^2]$ の RKHS 内での Frechet 微分:

$$\nabla L(f) = 2\mathbb{E}[K_X(\langle K_X, f\rangle_{\mathcal{H}_k} - Y)] = 2(\underbrace{\mathbb{E}[K_X K_X^*]}_{=:\Sigma} f - \mathbb{E}[K_X Y]) = 2(\Sigma f - \mathbb{E}[K_X Y]).$$

- 期待損失の勾配法:

$$f_t^* = f_{t-1}^* - \eta 2(\Sigma f_{t-1}^* - \mathbb{E}[K_X Y]).$$

- 経験損失の勾配法 ($\widehat{\mathbb{E}}[\cdot]$ は標本平均):

$$\hat{f}_t = \hat{f}_{t-1} - \eta 2(\widehat{\Sigma}\hat{f}_{t-1} - \widehat{\mathbb{E}}[K_X Y]).$$

- 確率的勾配による更新:

$$g_t = g_{t-1} - \eta 2(K_{x_{i_t}} K_{x_{i_t}}^* g_{t-1} - K_{x_{i_t}} y_{i_t}).$$

※ $(x_{i_t}, y_{i_t})_{t=1}^{\infty}$ は $(x_i, y_i)_{i=1}^{n}$ から i.i.d. 一様に取得.

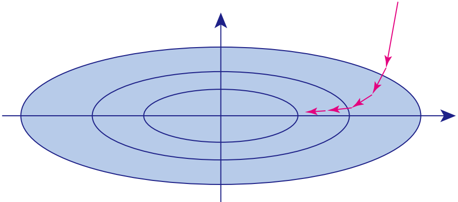
勾配のスムージングとしての見方

関数値の更新式:

$$
\begin{aligned}
f_t^*(x) &= f_{t-1}^*(x) - \eta 2(\Sigma f_{t-1}^* - \mathbb{E}[K_X Y])(x) \\
&= f_{t-1}^*(x) - 2\eta \int k(x,X) \underbrace{(f_{t-1}^*(X) - Y)}_{\to f_{t-1}^*(X) - f^{\mathrm{o}}(X)} \mathrm{d}P(X,Y) \\
&= f_{t-1}^*(x) - 2\eta T_k(f_{t-1}^* - f^{\mathrm{o}})(x).
\end{aligned}
$$

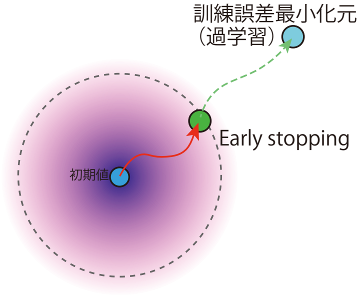
積分作用素 T_k は高周波成分を抑制する作用がある．

- RKHS 内の勾配は L_2 内の関数勾配を T_k によって平滑化したものになっている．(実際は T_k のサンプルからの推定値を使う)
- 高周波成分が出てくる前に止めれば過学習を防げる．
  $\to$ **Early stopping**
- 迂闊に Newton 法などを使うと危険.



Early stopping による正則化

Early stopping による正則化



バイアス-バリアンス分解

$$\underbrace{\|f^{\mathrm{o}} - \hat{f}\|_{L_2(P_X)}}_{\text{Estimation error}} \leq \underbrace{\|f^{\mathrm{o}} - \check{f}\|_{L_2(P_X)}}_{\substack{\text{Approximation error} \\ \text{(bias)}}} + \underbrace{\|\check{f} - \hat{f}\|_{L_2(P_X)}}_{\substack{\text{Sample deviation} \\ \text{(variance)}}}$$

訓練誤差最小化元に達する前に止める (early stopping) ことで正則化が働く.
無限次元モデル (RKHS) は過学習しやすいので気を付ける必要がある.

解析に用いる条件

通常，以下の条件を考える．(統計理論でも同様の仮定を課す定番の仮定)
(Caponnetto and de Vito, 2007, Dieuleveut et al., 2016, Pillaud-Vivien et al., 2018)

- $\boxed{\mu_i = O(i^{-\alpha})}$ for $\alpha > 1$.
  α は RKHS$\mathcal{H}_k$ の複雑さを特徴づける．(小さい α: 複雑，大きい α: 単純)
- $\boxed{f^{\mathrm{o}} \in T_k^r(L_2(P_X))}$ for $r > 0$.
  f^{o} が RKHS からどれだけ "はみ出ているか" を特徴づけ.
  $r = 1/2$ は $f^{\mathrm{o}} \in \mathcal{H}_k$ に対応．($r < 1/2$: はみ出てる, $r \geq 1/2$: 含まれる)
- $\boxed{\|f\|_{L_\infty(P_X)} \lesssim \|f\|_{L_2(P_X)}^{1-\mu} \|f\|_{\mathcal{H}_k}^{\mu} \ (\forall f \in \mathcal{H}_k)}$ for $\mu \in (0, 1]$.
  $\mathcal{H}_k$ に含まれている関数の滑らかさを特徴づけ. (小さい μ: 滑らか)

※ 最後の条件について: $f \in W^m([0,1]^d)$ (Sobolev 空間) かつ P_X の台が $[0,1]^d$ で密度関数を持ち，その密度が下からある定数 $c > 0$ で抑えられていれば，$\mu = d/(2m)$ でなりたつ.

収束レート

バイアス-バリアンスの分解:

$$\|f^{\mathrm{o}} - g_t\|_{L_2(P_X)}^2 \lesssim \underbrace{\|f^{\mathrm{o}} - f_t^*\|_{L_2(P_X)}^2}_{(a):\ \textbf{Bias}} + \underbrace{\|f_t^* - \hat{f}_t\|_{L_2(P_X)}^2}_{(b):\ \textbf{Variance}} + \underbrace{\|\hat{f}_t - g_t\|_{L_2(P_X)}^2}_{(c):\ \textbf{SGD deviation}}$$

$$\text{(a) } (\eta t)^{-2r}, \quad \text{(b) } \frac{(\eta t)^{1/\alpha} + (\eta t)^{\mu - 2r}}{n}, \quad \text{(c) } \eta(\eta t)^{1/\alpha - 1}$$

(a) 勾配法の解のデータに関する期待値と真の関数とのズレ (Bias).
(b) 勾配法の解の分散 (Variance).
(c) 確率的勾配を用いることによる変動.

更新数 t を大きくすると Bias は減るが Variance が増える．これらをバランスする必要がある (Early stopping).

Theorem (Multi-pass SGD の収束レート (Pillaud-Vivien et al., 2018))

$\eta = 1/(4 \sup_x k(x,x)^2)$ とする．

- $\mu\alpha < 2r\alpha + 1 < \alpha$ の時，$t = \Theta(n^{\alpha/(2r\alpha+1)})$ とすれば，

$$\mathbb{E}[L(g_t)] - L(f^{\mathrm{o}}) = O(n^{-2r\alpha/(2r\alpha+1)}).$$

- $\mu\alpha \geq 2r\alpha + 1$ の時，$t = \Theta(n^{\frac{1}{\mu}} (\log n)^{\frac{1}{\mu}})$ とすれば，$\mathbb{E}[L(g_t)] - L(f^{\mathrm{o}}) = O(n^{-2r/\mu})$.

Natural gradient の収束

Natural gradient (自然勾配法):

$$\hat{f}_t = \hat{f}_{t-1} - \eta(\Sigma + \lambda \mathrm{I})^{-1}(\widehat{\Sigma}\hat{f}_{t-1} - \widehat{\mathbb{E}}[K_X Y]).$$

(unlabeled data が沢山あり Σ は良く推定できる設定; GD の解析 (Murata and Suzuki, 2021))

Theorem (Natural gradient の収束 (Amari et al., 2021))

$$\mathbb{E}[\|\hat{f}_t - f^{\mathrm{o}}\|_{L_2(P_X)}^2] \lesssim B(t) + V(t),$$

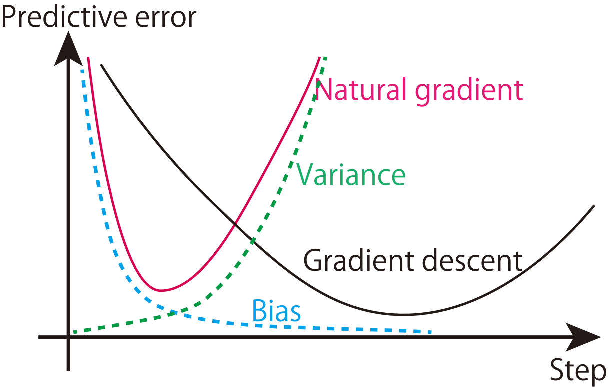
ただし, $B(t) = \exp(-\eta t) \vee (\lambda/(\eta t))^{2r}$,

$$V(t) = (1 + \eta t)\frac{\lambda^{-1}B(t) + \lambda^{-\frac{1}{\alpha}}}{n} + (1 + t\eta)^4 \frac{(1 \vee \lambda^{2r-\mu})\lambda^{-\frac{1}{\alpha}}}{n}.$$

特に, $\lambda = n^{-\frac{\alpha}{2r\alpha+1}}$, $t = \Theta(\log(n))$ で $\mathbb{E}[\|\hat{f}_t - f^{\mathrm{o}}\|_{L_2(P_X)}^2] = O(n^{-\frac{2r\alpha}{2r\alpha+1}}\log(n)^4)$.

※ バイアスは急速に収束するが, バリアンスも速く増大する.
→ Preconditioning のため高周波成分が早めに出現する. より早めに止めないと過学習する.

収束の様子


Predictive error
Natural gradient
Variance
Gradient descent
Bias
Step

作用素 Bernstein の不等式

- $\Sigma = \mathbb{E}_x[K_x K_x^*]$: $\Sigma f = \int k(\cdot, x) f(x) \mathrm{d}P_x(x)$
- $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n K_{x_i} K_{x_i}^*$: $\widehat{\Sigma} f = \frac{1}{n}\sum_{i=1}^n k(\cdot, x_i) f(x_i)$

$\Sigma_\lambda := \Sigma + \lambda \mathrm{I}$, $\mathcal{F}_\infty(\lambda) := \sup_x K_x^* \Sigma_\lambda^{-1} K_x$ とする．以下のような評価が必要:

$$\|\Sigma_\lambda^{-1}(\Sigma - \widehat{\Sigma})\Sigma_\lambda^{-1}\| \lesssim \sqrt{\frac{\mathcal{F}_\infty(\lambda)\beta}{n}} + \frac{(1 + \mathcal{F}_\infty(\lambda))\beta}{n}$$

with prob. $1 - \delta$. ただし，$\beta = \log(\frac{4\mathrm{Tr}[\Sigma\Sigma_\lambda^{-1}]}{\delta})$.
$\rightarrow$ 経験分布と真の分布のずれをバウンド.

Theorem (自己共役作用素の Bernstein の不等式 (Minsker, 2017))

$(X_i)_{i=1}^n$ は独立な自己共役作用素の確率変数で $\mathbb{E}[X_i] = 0$ かつ,
$\sigma^2 \geq \|\sum_{i=1}^n \mathbb{E}[X_i^2]\|$, $U \geq \|X_i\|$ とする．$r(A) = \mathrm{Tr}[A]/\|A\|$ として,

$$P\left(\left\|\sum_{i=1}^n X_i\right\| \geq t\right) \leq 14 r(\textstyle\sum_{i=1}^n \mathbb{E}[X_i^2]) \exp\left(-\frac{t^2}{2(\sigma^2 + tU/3)}\right).$$

$X_i = \Sigma_\lambda^{-1} K_{x_i} K_{x_i}^* \Sigma_\lambda^{-1}$ とする． (Tropp (2012) も参照)

S. Amari, J. Ba, R. B. Grosse, X. Li, A. Nitanda, T. Suzuki, D. Wu, and J. Xu. When does preconditioning help or hurt generalization? In *International Conference on Learning Representations*, 2021.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

A. Caponnetto and E. de Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

A. Dieuleveut, F. Bach, et al. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363–1399, 2016.

S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

T. Murata and T. Suzuki. Gradient descent in rkhs with importance labeling. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1981–1989. PMLR, 2021.

L. Pillaud-Vivien, A. Rudi, and F. Bach. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In *Advances in Neural Information Processing Systems*, pages 8114–8124, 2018.

I. Steinwart and C. Scovel. Mercer ' s theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

J. A. Tropp. User-friendly tools for random matrices: An introduction. Technical report, 2012.