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
@九州大学集中講義

Outline

Outline

線形回帰

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

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

リッジ回帰（**Tsykonov 正則化**）

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

線形回帰

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

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

リッジ回帰（**Tsykonov 正則化**）

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

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.

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

※ $(XX^\top)_{ij} = x_i^\top x_j$ より，観測値 x_i と x_j の内積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）

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

※ $(XX^\top)_{ij} = x_i^\top x_j$ 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）

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

※ $(XX^\top)_{ij} = x_i^\top x_j$ 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 $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ 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: $\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j k(x_i, x_j) \geq 0, \quad (\forall \{x_i\}_{i=1}^m, \ \{\alpha_i\}_{i=1}^m, \ m)$.

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

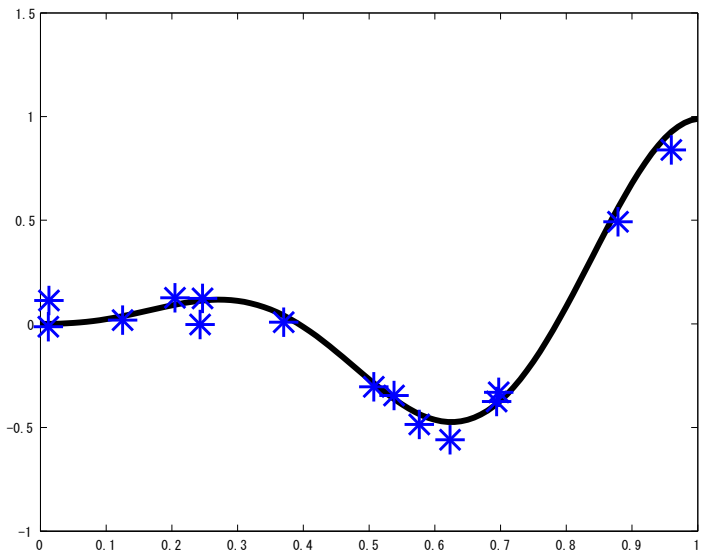
カーネルリッジ回帰: $K = (k(x_i, x_j))_{i,j=1}^n$ として,

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

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.



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,

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

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.

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

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: $\langle \sum_{i=1}^n \alpha_i \phi(x_i), \phi(x) \rangle_{\mathcal{H}_k} = \sum_{i=1}^n \alpha_i k(x_i, x)$.
  →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)

入力データの分布：P_X，対応する L_2 空間：$L_2(P_X) = \{f \mid \mathbb{E}_{X\sim P_X}[f(X)^2] < \infty\}$.
カーネル関数は以下のように分解できる (Steinwart and Scovel, 2012):

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

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

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

- $\langle f, g\rangle_{\mathcal{H}_k} := \sum_{j=1}^{\infty} \frac{1}{\mu_j}\alpha_j\beta_j$ 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:

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

再生核ヒルベルト空間の性質

$$\phi_k(x) = k(x, \cdot) \in \mathcal{H}_k$$

と書けば，$k(x, x') = \langle \phi_k(x), \phi_k(x') \rangle_{\mathcal{H}_k}$ と書ける．この ϕ_k を特徴写像とも言う．

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:

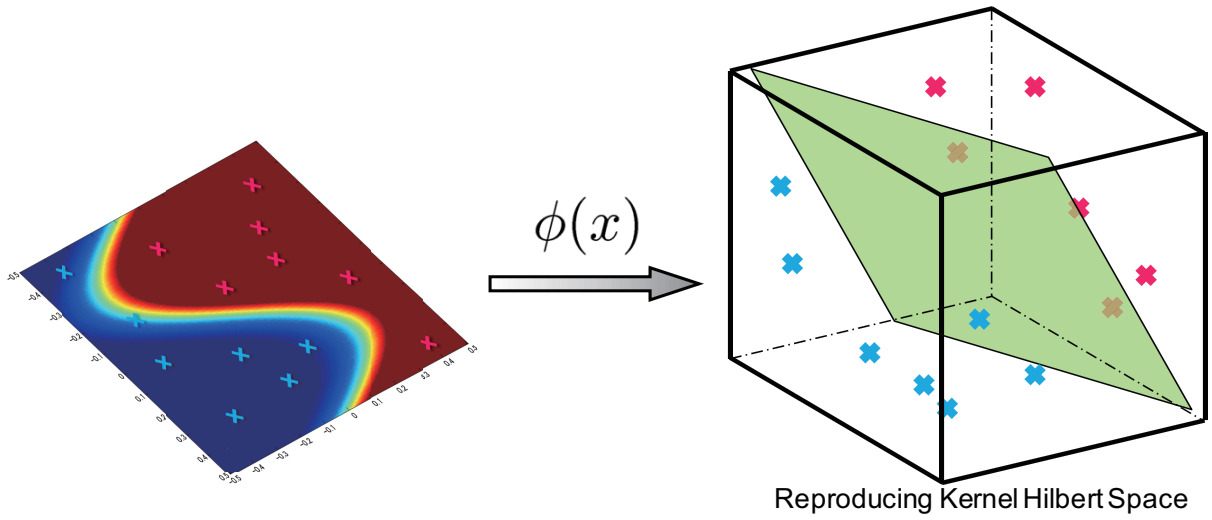
$$\mathcal{H}_k = T_k^{1/2} L_2(P_X).$$

- $\|f\|_{\mathcal{H}_k} = \inf\{\|h\|_{L_2(P_X)} \mid f = T_k^{1/2} h,\ h \in L_2(P_X)\}$.
  - $f \in \mathcal{H}_k$ は $f(x) = \sum_{j=1}^{\infty} a_j \sqrt{\mu_j} e_j(x)$ と書けて，$\|f\|_{\mathcal{H}_k} = \sqrt{\sum_{j=1}^{\infty} a_j^2}$.
  - $(e_j)_j$ 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．
  - 特徴写像 $\phi_k(x) = k(x, \cdot) \in \mathcal{H}_k$ を完全正規直交基底に関する係数で表現すると

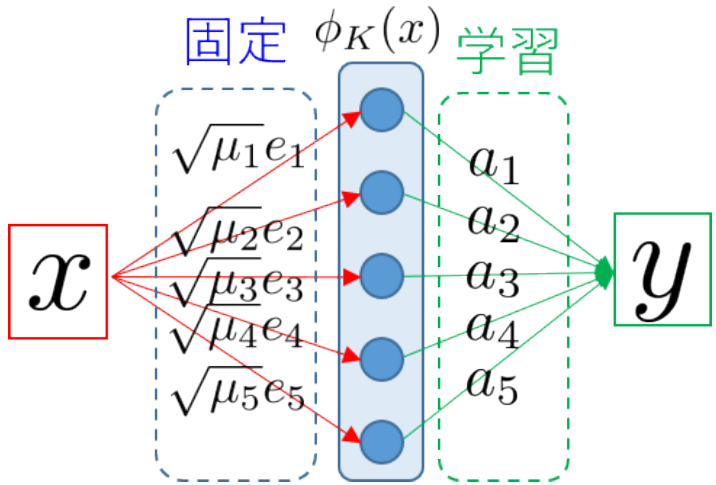
$$\phi_k(x) = (\sqrt{\mu_1} e_1(x), \sqrt{\mu_2} e_2(x), \dots)^\top$$

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)



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

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

- 表現定理

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

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

さきほどのカーネルリッジ回帰の定式化と一致.

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

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:

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

$\to \min_{f \in \mathcal{H}_k} \mathbb{E}[(f(X) - Y)^2]$ 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.

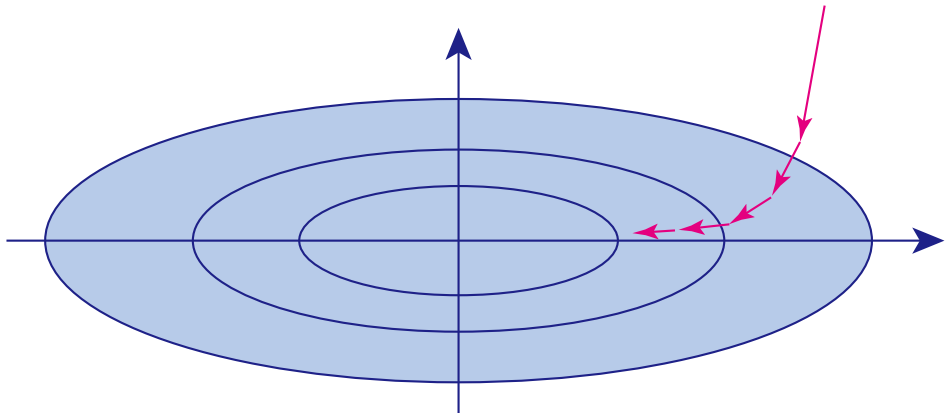
勾配のスムージングとしての見方

関数値の更新式:

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

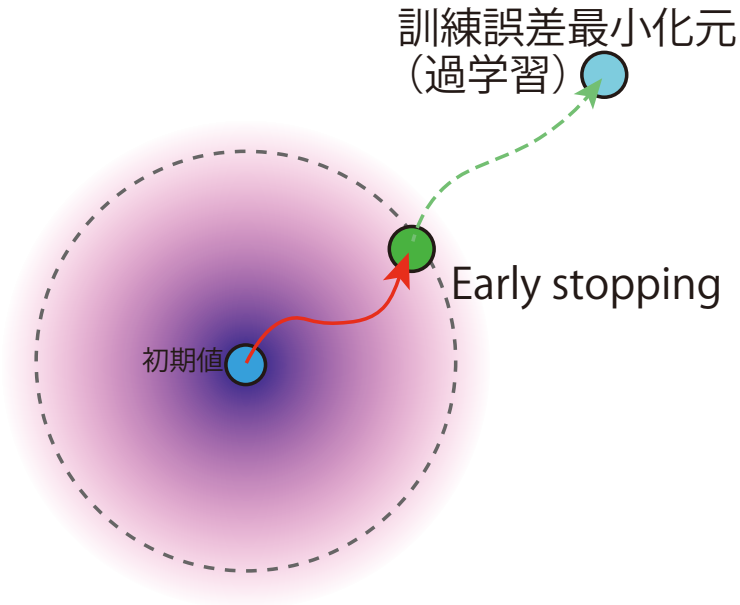
積分作用素 T_k は高周波成分を抑制する作用がある.

- 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.
  → **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)

- $\mu_i = O(i^{-\alpha})$ 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)
- $f^{\mathrm{o}} \in T^r(L_2(P_X))$ 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)
- $\|f\|_{L_\infty(P_X)} \lesssim \|f\|_{L_2(P_X)}^{1-\mu} \|f\|_{\mathcal{H}_k}^{\mu}$ $(\forall f \in \mathcal{H}_k)$ 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:

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

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

(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, 2020))

Theorem (Natural gradient の収束 (Amari et al., 2020))

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

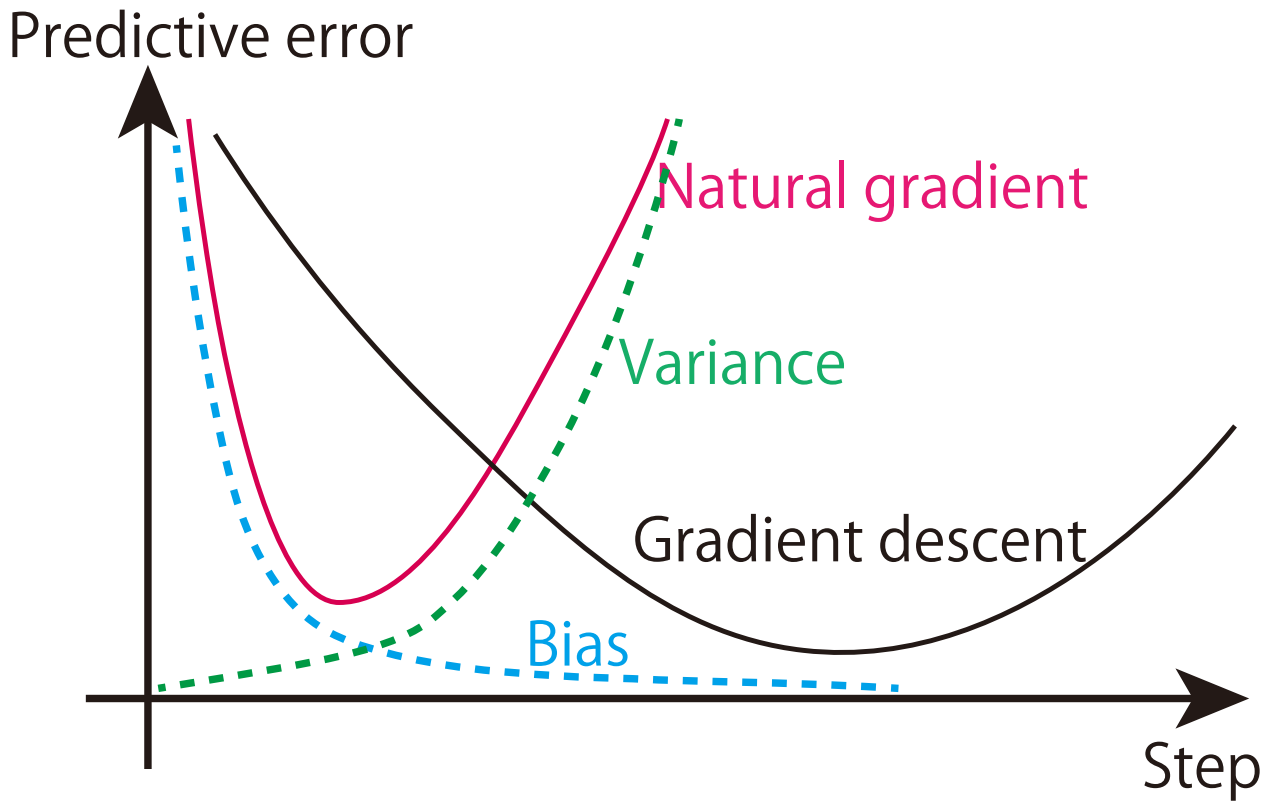
ただし, $B(t) = \exp(-\eta t) \vee (\lambda/(\eta t))^{2r}$,

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

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

※ 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，$\beta = \log(\frac{4\mathrm{Tr}[\Sigma\Sigma_\lambda^{-1}]}{\delta})$.
→ 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)

正則化ありの確率的最適化

二乗損失を拡張して，一般の滑らかな凸損失関数 ℓ を考える．(判別問題など)
正則化ありの期待損失最小化:

$$\min_{f \in \mathcal{H}_k} \quad \mathbb{E}[\ell(Y, f(X))] + \lambda \|f\|_{\mathcal{H}_k}^2 =: L_\lambda(f).$$

これを SGD で解く．目的関数が λ-強凸であることを利用.

$$g_{t+1} = g_t - \eta_t(\ell'(y_t, g_t(x_t)) + \lambda g_t).$$

$$\bar{g}_{T+1} = \sum_{t=1}^{T+1} \frac{2(c_0+t-1)}{(2c_0+T)(T+1)} g_t \quad \text{(多項式平均)}.$$

仮定：(i) ℓ は γ-平滑，$\|\ell'\|_\infty \le M$, (ii) $k(x,x) \le 1$.　　$g_\lambda = \mathrm{argmin}_{g \in \mathcal{H}_k} L_\lambda(g)$.

Theorem (Nitanda and Suzuki (2019))

適切な $c_0 > 0$ に対して $\eta_t = 2/(\lambda(c_0 + t))$ とすれば,

$$\mathbb{E}[L_\lambda(\bar{g}_{T+1}) - L_\lambda(g_\lambda)] \lesssim \frac{M^2}{\lambda(c_0+T)} + \frac{\gamma+\lambda}{T+1}\|g_1 - g_\lambda\|_{\mathcal{H}_k}^2.$$

さらにマルチンゲール確率集中不等式より High probability bound も得られる．
判別問題なら strong low noise condition のもと判別誤差の**指数収束**も示せる．

マルチンゲール Hoeffding の不等式

Theorem (マルチンゲール Hoeffding 型集中不等式 (Pinelis, 1994))

確率変数列: $D_1, \ldots, D_T \in \mathcal{H}_k$. $\mathbb{E}[D_t] = 0$, $\|D_t\|_{\mathcal{H}_k} \leq R_t$ *(a.s.)* とする.
$\forall \epsilon > 0$ に対し

$$P\left[\max_{1 \leq t \leq T} \|\sum_{s=1}^{t} D_s\|_{\mathcal{H}_k} \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{t=1}^{T} R_t^2}\right).$$

$$D_t = \mathbb{E}[\bar{g}_{T+1}|Z_1, \ldots, Z_t] - \mathbb{E}[\bar{g}_{T+1}|Z_1, \ldots, Z_{t-1}],$$

ただし $Z_t = (x_t, y_t)$ とすれば, $\sum_{t=1}^{T} D_t = \bar{g}_{T+1} - \mathbb{E}[\bar{g}_{T+1}]$ となり, 期待値と実現値のずれを抑えられる.

マルチンゲール Hoeffding の不等式

Theorem (マルチンゲール Hoeffding 型集中不等式 (Pinelis, 1994))

確率変数列: $D_1, \ldots, D_T \in \mathcal{H}_k$. $\mathbb{E}[D_t] = 0$, $\|D_t\|_{\mathcal{H}_k} \leq R_t$ *(a.s.)* とする.
$\forall \epsilon > 0$ に対し

$$P\left[\max_{1 \leq t \leq T} \|\sum_{s=1}^{t} D_s\|_{\mathcal{H}_k} \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{t=1}^{T} R_t^2}\right).$$

$$D_t = \mathbb{E}[\bar{g}_{T+1}|Z_1, \ldots, Z_t] - \mathbb{E}[\bar{g}_{T+1}|Z_1, \ldots, Z_{t-1}],$$

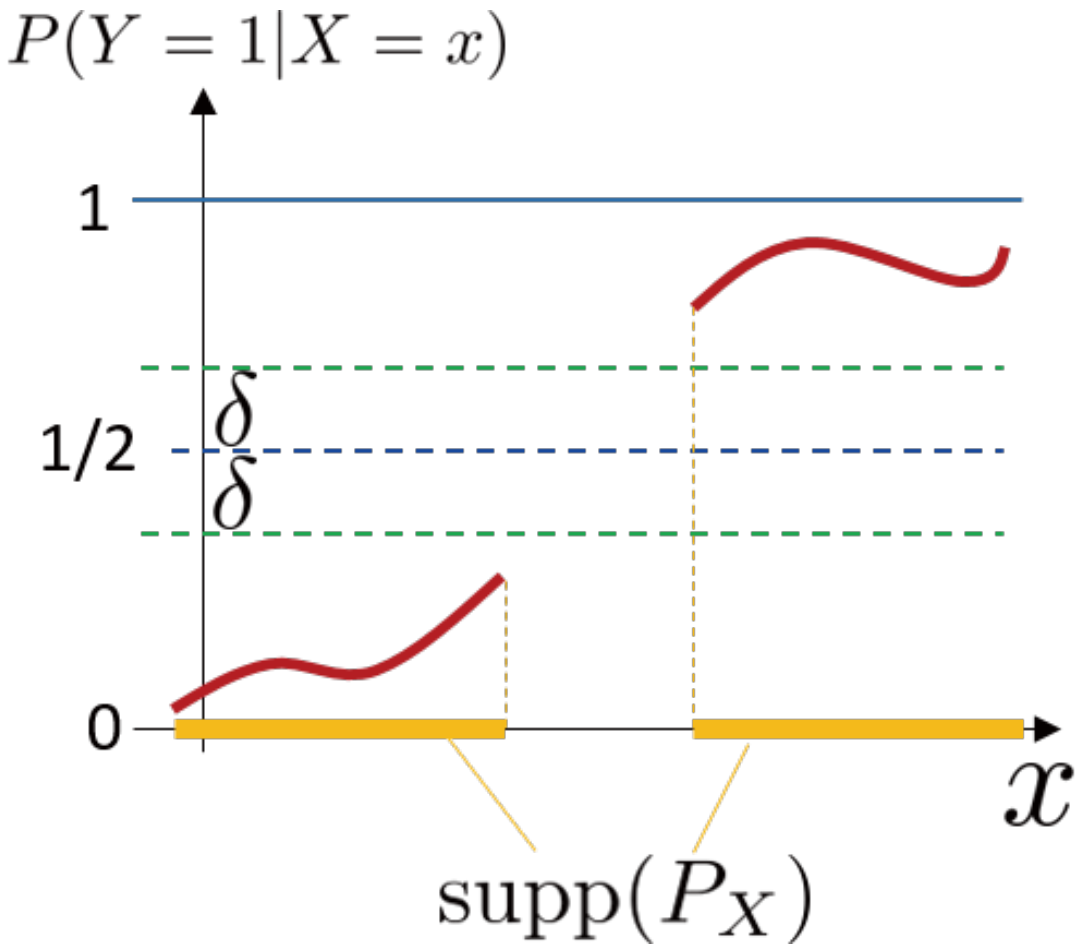
ただし $Z_t = (x_t, y_t)$ とすれば, $\sum_{t=1}^{T} D_t = \bar{g}_{T+1} - \mathbb{E}[\bar{g}_{T+1}]$ となり, 期待値と実現値のずれを抑えられる.

(補足) $\mathcal{L}_\lambda$ は RKHS ノルムに関して λ-強凸であることより,

$$\|\bar{g}_{T+1} - g_\lambda\|_{\mathcal{H}_k} \leq O(\frac{1}{\lambda^2 T})$$

が高い確率で成り立つ. 実は $\|\cdot\|_\infty \leq \|\cdot\|_{\mathcal{H}_k}$ でもあるので,
$|P(Y=1|X) - P(Y=-1|X)| \geq \delta$ なるマージン条件 (strong low noise condition) のもと, 完全な判別が高い確率でできるようになる.

(参考) Strong low noise condition



Outline

Integral representation

Definition: η and ψ are *admissible* if

$$\int \frac{\widehat{\psi}(\zeta)\widehat{\eta}(\zeta)}{|\zeta|^d}\mathrm{d}\zeta < \infty.$$

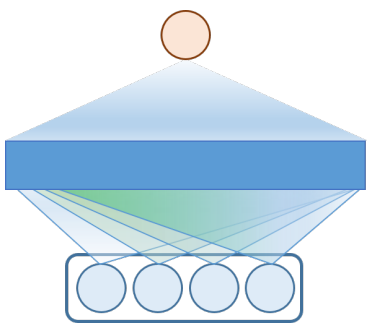
(where $\widehat{\psi}, \widehat{\eta}$ are the Fourie traonsform of ψ, η).

Theorem (Sonoda and Murata (2015))

If $f : \mathbb{R}^d \to \mathbb{R}$ and its Fourie transorm are in $L_1(\mathbb{R}^d)$, and η, ψ are admissible (e.g., η is ReLU), then

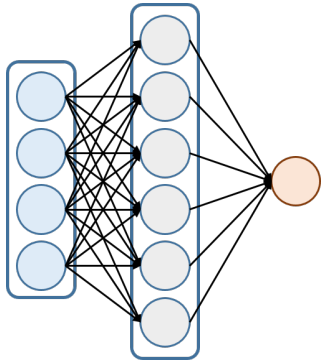
$$T(w, b) = \int f(x)\psi(w^\top x - b)\|x\|\mathrm{d}x,$$

$$f(x) = \int T(w, b)\|w\|^{-1}\eta(w^\top x - b)\mathrm{d}w\mathrm{d}b \quad (\textbf{integral form}).$$



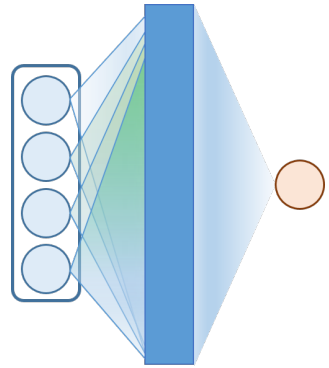
Integral representation of deep neural network

Finite sum form

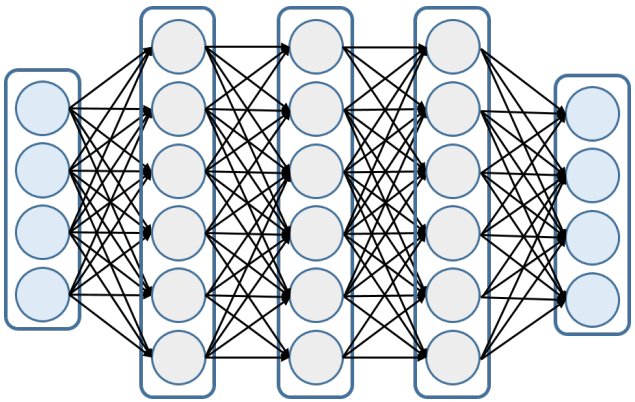


$$\hat{f}(x) = \sum_{j=1}^{m} v_j \eta(w_j^\top x + b_j)$$

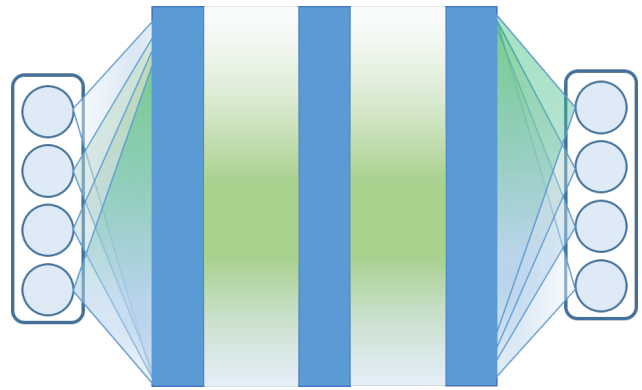
Integral form



$$f^{\mathrm{o}}(x) = \int h(w,b)\eta(w^\top x + b)\mathrm{d}w\mathrm{d}b$$



$$\hat{f}(x) = W_L \eta(W_{L-1}\eta(W_{L-2}\dots \eta(W_1 x + b_1) + b_2 \dots)))$$

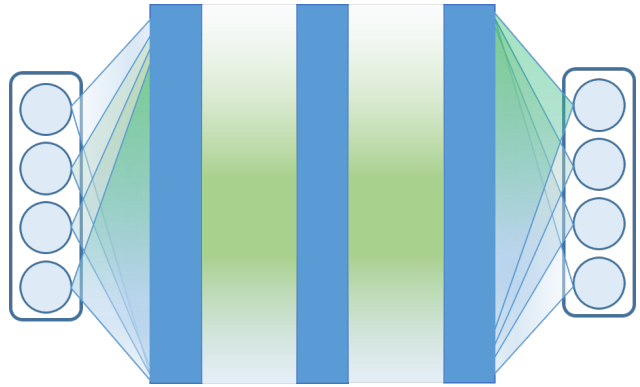


$$f^{\mathrm{o}}(x) = f_L^{\mathrm{o}} \circ f_{L-1}^{\mathrm{o}} \circ \cdots \circ f_1^{\mathrm{o}}(x)$$

$$f_\ell^{\mathrm{o}}[F](\tau, x) = \int h_\ell^{\mathrm{o}}(\tau,\tau')\eta(F(\tau', x))\mathrm{d}Q_\ell(\tau') + b_\ell^{\mathrm{o}}(\tau).$$

Still universal approximator.

Detail of the integral form of DNN



Output to the ℓ-th layer:

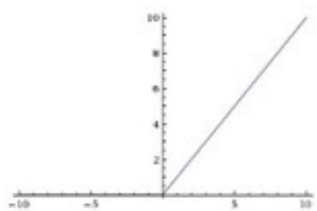
$$F_\ell(\tau, x) = \int_{\mathcal{Y}_\ell} \underbrace{h_\ell^{\circ}(\tau, \tau')}_{\text{Weight}} \eta(F_{\ell-1}(\tau', x)) \mathrm{d}Q_\ell(\tau') + \underbrace{b_\ell^{\circ}(\tau)}_{\text{Bias}}.$$

This measures how much the input x contains the feature τ at the ℓ-th layer.

- $\mathcal{Y}_\ell$: the feature index space at the ℓ-th layer (Generally continuous space).
- Q_ℓ: prob. measure on $\mathcal{Y}_\ell$

Examples of activation functions:

- **ReLU:** $\eta(u) = \max\{u, 0\}$
- **Sigmoid:** $\eta(u) = \frac{1}{1+\exp(-u)}$



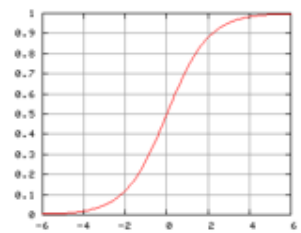
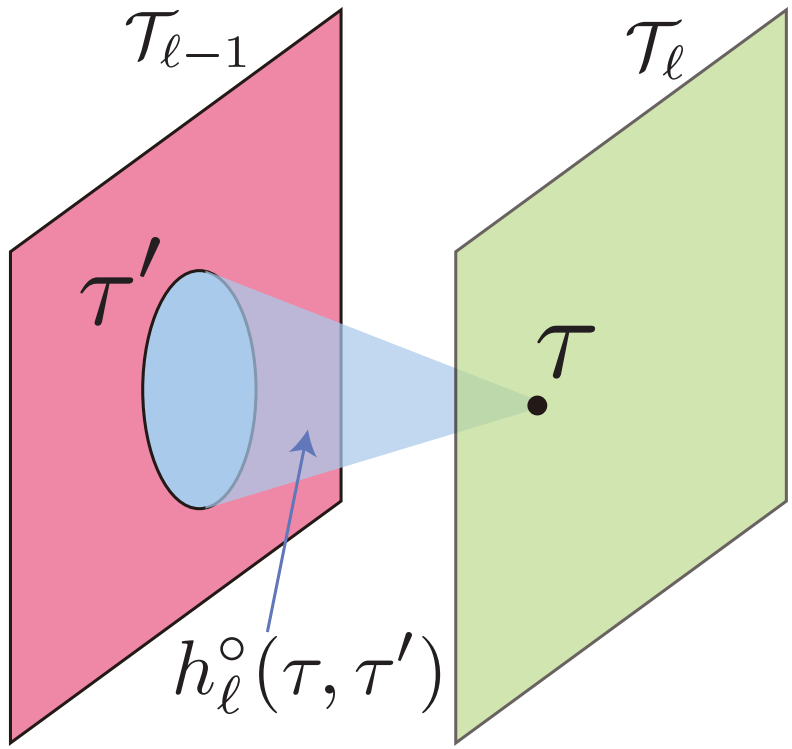


Illustration of continuous feature space

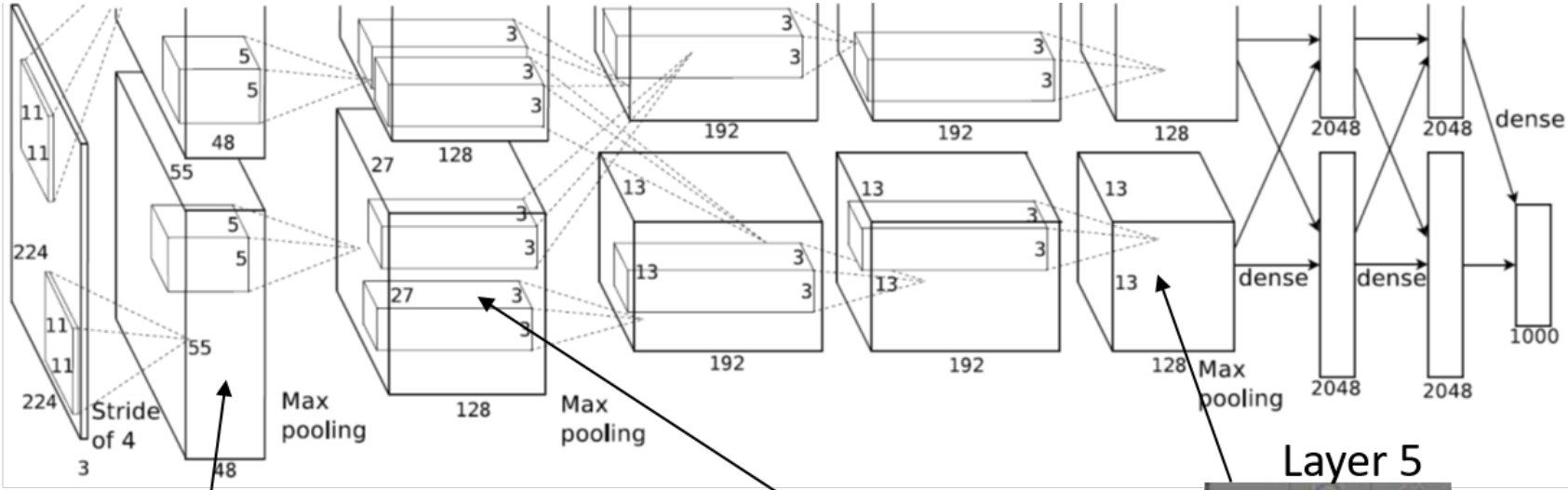


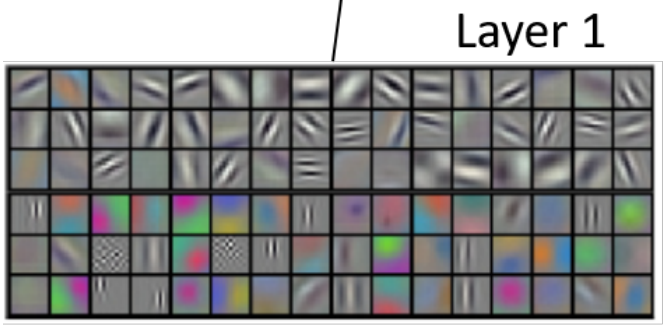
$$F_\ell(\tau, x) = \int_{\mathcal{Y}_\ell} \underbrace{h_\ell^\circ(\tau, \tau')}_{\text{Weight}} \eta(F_{\ell-1}(\tau', x)) \mathrm{d}Q_\ell(\tau') + \underbrace{b_\ell^\circ(\tau)}_{\text{Bias}}.$$

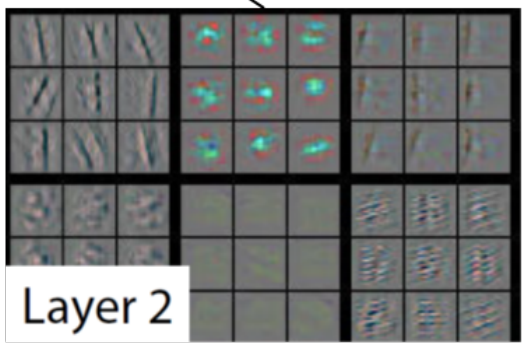
The shape of the space $\mathcal{Y}_\ell$ could be arbitrary.
(could be discrete and could be continuous)

Continuous feature in real

Distributed representation in a real DNN







Reproducing kernel Hilbert space on the ℓ-th layer

Construct an RKHS on each layer.
$\rightarrow$ We can employ the theory of the kernel method.

$F_\ell(\tau, x)$: an output from the ℓ-th layer to the feature τ in the next layer.
$(F_\ell(\tau, x) = \int_{\mathcal{Y}_\ell} h_\ell^{\circ}(\tau, \tau')\eta(F_{\ell-1}(\tau', x))\mathrm{d}Q_\ell(\tau') + b_\ell^{\circ}(\tau).)$

$$\mathrm{k}_\ell(x, x') = \int \eta(F_{\ell-1}(\tau, x))\eta(F_{\ell-1}(\tau, x'))\mathrm{d}Q_\ell(\tau)$$

- k_ℓ defines an RKHS $\mathcal{H}_\ell$.
- For all $f \in \mathcal{H}_\ell$, there exits $h \in L_2(Q_\ell)$ and $g \in L_2(P(X))$ such that

$$f(x) = \int_{\mathcal{Y}_\ell} h(\tau')\eta(F_{\ell-1}(\tau', x))\mathrm{d}Q_\ell(\tau') = \int \mathrm{k}_\ell(x, x')g(x')\mathrm{d}P(x')$$
$$\|f\|_{\mathcal{H}_\ell} = \|h\|_{L_2(Q_\ell)} = \|g\|_{L_2(P(X))}$$

(c.f., Bach (2015)).

Complexity of RKHS

- Let
$$T_\ell : f \mapsto \int \mathrm{k}_\ell(\cdot, x') f(x') \mathrm{d}P(x').$$
- Let the spectrum decomposition of k_ℓ be
$$\mathrm{k}_\ell(x, x') = \sum_{j=1}^{\infty} \mu_j^{(\ell)} \phi_j^{(\ell)}(x) \phi_j^{(\ell)}(x')$$
in $L_2(P(X) \times P(X))$.

Definition

The *degree of freedom* of $\mathcal{F}_\ell$ is defined as
$$N_\ell(\lambda) := \mathrm{Tr}[(T_\ell + \lambda)^{-1} T_\ell] = \sum_{j=1}^{\infty} \frac{\mu_j^{(\ell)}}{\mu_j^{(\ell)} + \lambda}.$$

$N_\ell(\lambda)$ measures *complexity* of the RKHS.

This is very much related to the notion of *covering number* of the RKHS.

Degree of freedom in kernel method

The degree of freedom appears to characterize the generalization error of kernel ridge regression.

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

where $\mathcal{H}$ is an RKHS with a bounded kernel k.

Proposition (Caponnetto and de Vito (2007))

If $f^\circ \in \mathcal{H}$, then it holds that

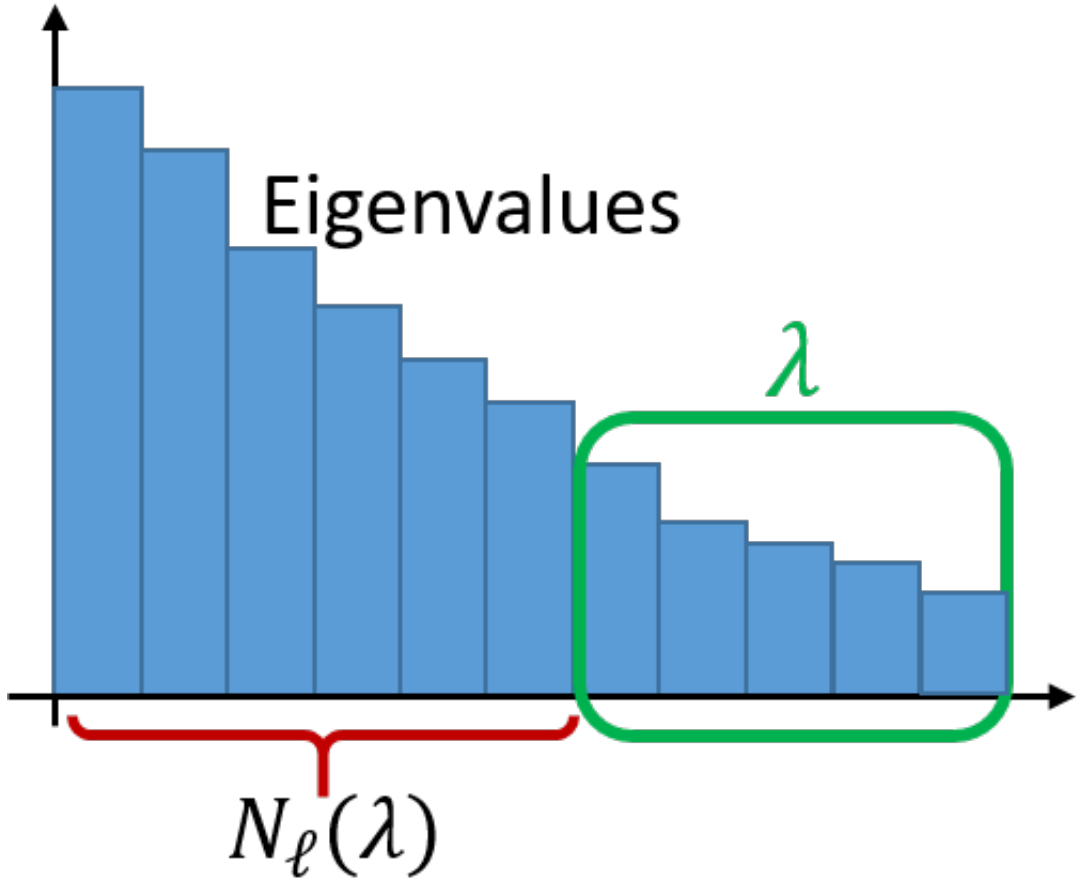
$$\|\hat{f}_\lambda - f^\circ\|_{L_2(P_X)}^2 \leq C\Big(\underbrace{\lambda}_{\text{bias}} + \underbrace{\frac{N(\lambda)}{n}}_{\text{variance}}\Big),$$

with high probability. ($N(\lambda) := \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j+\lambda}$ where $(\mu_j)_{j=1}^{\infty}$ are the eigenvalues of the kernel)

Basically, λ satisfying

$$\frac{N(\lambda)}{n} = \lambda$$

gives the optimal rate.



Rough sketch of $N_\ell(\lambda)$.

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

Finite approximation via kernel quadrature

Theorem (Approximation error in RKHS $\mathcal{H}_\ell$)

For $\lambda > 0$, suppose that

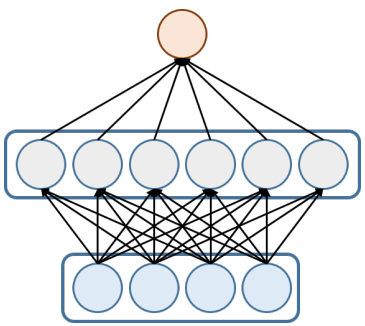
$$m_\ell \geq 5N_\ell(\lambda)\log(64N_\ell(\lambda)),$$

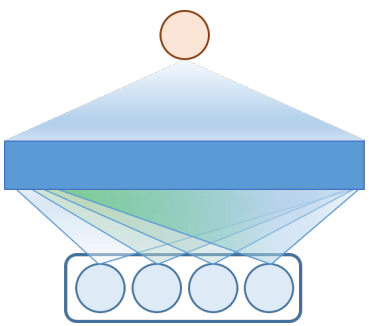
then there exist positive reals $\{\tau_i\}_{i=1}^{m_\ell} \subset \mathcal{Y}_\ell$ and $(q_j)_{j=1}^{m_\ell}$ with $\sum_{j=1}^{m_\ell} \frac{1}{q_j} \leq 2m_\ell$ such that

$$\sup_{f:\|f\|_{\mathcal{H}_\ell}\leq 1} \inf_{\|\beta\|_2^2 \leq \frac{4}{m_\ell}} \left\| f - \sum_{j=1}^{m_\ell} \beta_j q_j^{-1/2} \eta(F_{\ell-1}(\tau_j, \cdot)) \right\|_{L_2(P(X))}^2 \leq 4\lambda.$$

Proof is given by a modification of Bach (2015).

The true function g can be approximated with precision λ by a finite sum ($m_\ell = O(N_\ell(\lambda)\log(N_\ell(\lambda)))$).





$$F_\ell(\tau, x) = \int_{\mathcal{Y}_\ell} h_\ell^\circ(\tau, \tau')\eta(F_{\ell-1}(\tau', x))\mathrm{d}Q_\ell(\tau') + b_\ell^\circ(\tau).$$

$$N_\ell(\lambda) = \sum_{j=1}^{\infty} \frac{\mu_j^{(\ell)}}{\mu_j^{(\ell)} + \lambda}$$

Finite approximation via kernel quadrature

Theorem (Approximation error in RKHS $\mathcal{H}_\ell$)

For $\lambda > 0$, suppose that

$$m_\ell \geq 5 N_\ell(\lambda) \log(64 N_\ell(\lambda)),$$

then there exist positive reals $\{\tau_i\}_{i=1}^{m_\ell} \subset \mathcal{Y}_\ell$ and $(q_j)_{j=1}^{m_\ell}$ with $\sum_{j=1}^{m_\ell} \frac{1}{q_j} \leq 2m_\ell$ such that if η is scale invariant ($\eta(au) = a\eta(u)$ $(\forall a > 0)$), then

$$\sup_{f:\|f\|_{\mathcal{H}_\ell} \leq 1} \inf_{\|\boldsymbol{\beta}\|_2^2 \leq \frac{4}{m_\ell}} \left\| f - \sum_{j=1}^{m_\ell} \beta_j \eta(q_j^{-1/2} F_{\ell-1}(\tau_j, \cdot)) \right\|_{L_2(P(X))}^2 \leq 4\lambda.$$

Proof is given by a modification of Bach (2015).
The true function g can be approximated with precision λ by a finite sum ($m_\ell = O(N_\ell(\lambda) \log(N_\ell(\lambda)))$).

This reduces the complexity very much!

Assumption on norms

Now, we move to the deep neural network, and assume the following norm bound.

$$F_\ell(\tau, x) = \int_{\mathcal{Y}_\ell} \underbrace{h_\ell^{\mathrm{o}}(\tau, \tau')}_{\text{Weight}} \eta(F_{\ell-1}(\tau', x)) \mathrm{d}Q_\ell(\tau') + \underbrace{b_\ell^{\mathrm{o}}(\tau)}_{\text{Bias}}.$$

- $\sup_{\tau \in \mathcal{Y}_{\ell+1}} \|h_\ell^{\mathrm{o}}(\tau, \cdot)\|_{L_2(Q_\ell)} \leq R \quad (\forall \ell)$
  $(\Rightarrow \|F_\ell(\tau, \cdot)\|_{\mathcal{H}_\ell} \leq R)$
- $\|b_\ell^{\mathrm{o}}\|_\infty \leq R_b \quad (\forall \ell)$

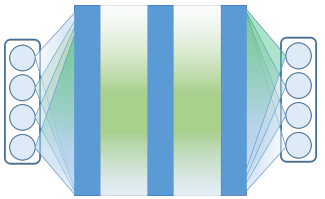
Approximation error of deep NN

(degree of freedom) $N_\ell(\lambda) = \sum_{j=1}^{\infty} \frac{\mu_j^{(\ell)}}{\mu_j^{(\ell)}+\lambda}$.

Integral form:

$$f_\ell^{\mathrm{o}}(g) = \int h_\ell^{\mathrm{o}}(\tau,\tau')\eta(g(\tau'))\mathrm{d}Q_\ell(\tau') + b_\ell^{\mathrm{o}}(\tau),$$

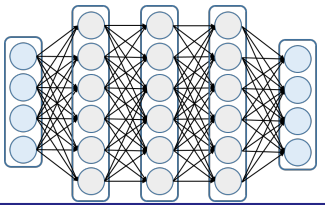
$$f^{\mathrm{o}}(x) = f_L^{\mathrm{o}} \circ f_{L-1}^{\mathrm{o}} \circ \cdots \circ f_1^{\mathrm{o}}(x),$$



Finite dimensional model:

$$f_\ell^*(g) = W^{(\ell)}\eta(g) + b^{(\ell)},$$

$$f^*(x) = f_L^* \circ f_{L-1}^* \circ \cdots \circ f_1^*(x).$$



Theorem (Approximation error of deep NN)

For any $\lambda_\ell > 0,\ \delta > 0,$

$$m_\ell \geq 5N_\ell(\lambda_\ell)\log(32N_\ell(\lambda_\ell)/\delta) \quad \text{(width of } \ell\text{-th layer)}$$

$$\Rightarrow \quad \exists\{W^{(\ell)}, b^{(\ell)}\}_{\ell=1}^{L} \quad \text{s.t.} \quad \|f^{\mathrm{o}} - f^*\|_{L_2(P(X))} \leq \sum_{\ell=2}^{L} 2\sqrt{\hat{c}_\delta^{L-\ell-1}}R^{L-\ell}\sqrt{\lambda_\ell}$$

where $\hat{c}_\delta = \frac{4}{1-\delta}$,

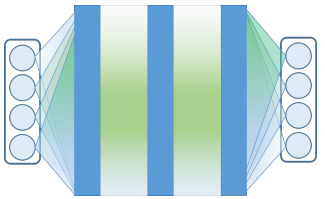
Approximation error of deep NN

(degree of freedom) $N_\ell(\lambda) = \sum_{j=1}^{\infty} \frac{\mu_j^{(\ell)}}{\mu_j^{(\ell)}+\lambda}$.

Integral form:

$$f_\ell^{\mathrm{o}}(g) = \int h_\ell^{\mathrm{o}}(\tau,\tau')\eta(g(\tau'))\mathrm{d}Q_\ell(\tau') + b_\ell^{\mathrm{o}}(\tau),$$

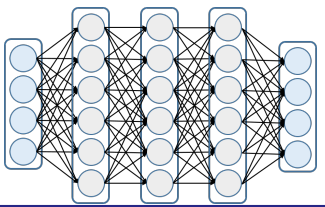
$$f^{\mathrm{o}}(x) = f_L^{\mathrm{o}} \circ f_{L-1}^{\mathrm{o}} \circ \cdots \circ f_1^{\mathrm{o}}(x),$$



Finite dimensional model:

$$f_\ell^*(g) = W^{(\ell)}\eta(g) + b^{(\ell)},$$

$$f^*(x) = f_L^* \circ f_{L-1}^* \circ \cdots \circ f_1^*(x).$$



Theorem (Approximation error of deep NN)

For any $\lambda_\ell > 0,\ \delta > 0,$

$$m_\ell \geq 5N_\ell(\lambda_\ell)\log(32N_\ell(\lambda_\ell)/\delta) \quad \text{(width of } \ell\text{-th layer)}$$

$$\Rightarrow \quad \exists\{W^{(\ell)}, b^{(\ell)}\}_{\ell=1}^{L} \quad \text{s.t.} \quad \|f^{\mathrm{o}} - f^*\|_{L_2(P(X))} \leq \sum_{\ell=2}^{L} 2\sqrt{\hat{c}_\delta^{L-\ell-1}}R^{L-\ell}\sqrt{\lambda_\ell}$$

where $\hat{c}_\delta = \frac{4}{1-\delta}$, *moreover* $\|W^{(\ell)}\|_F \leq \hat{c}_\delta R, \|b^{(\ell)}\| \leq R_b.$

S. Amari, J. Ba, R. Grosse, X. Li, A. Nitanda, T. Suzuki, D. Wu, and J. Xu. When does preconditioning help or hurt generalization?, 2020.

F. Bach. On the equivalence between kernel quadrature rules and random feature expansions. arXiv preprint arXiv:1502.06800, 2015.

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, 2020.

A. Nitanda and T. Suzuki. Stochastic gradient descent with exponential convergence rates of expected classification errors. In K. Chaudhuri and M. Sugiyama, editors, Proceedings of Machine Learning Research, volume 89 of Proceedings of Machine Learning Research, pages 1417–1426. PMLR, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/nitanda19a.html.

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. Pinelis. Optimum bounds for the distributions of martingales in banach spaces. The Annals of Probability, pages 1679–1706, 1994.

S. Sonoda and N. Murata. Neural network with unbounded activation functions is universal approximator. Applied and Computational Harmonic Analysis, 2015.

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.