Early stopping points for gradient desccent

A survey

Mohammed HSSEIN

[1]Centrale Lille Institut
Villeneuve d'Ascq, France

Promo : 2021

Plan

1. Introduction
   - Context
   - Settings
   - Kernels
2. Stopping rules
3. Conclusion

Context

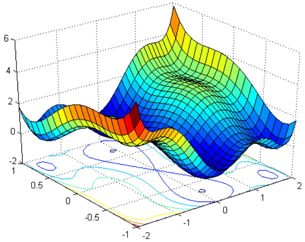


Figure: Local minimums

Context

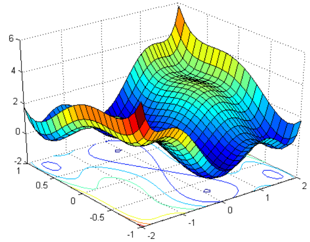


Figure: Local minimums

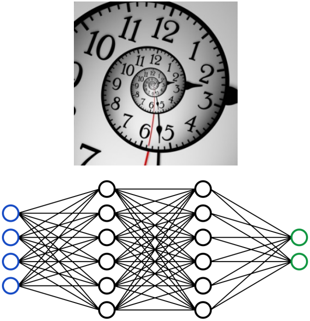


Figure: Time

Context

Running infinite iterations, lead to over-fitting !

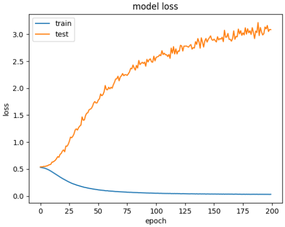


Figure: regression problem deep nets

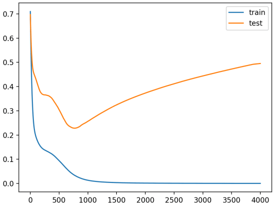


Figure: classification problem deep nets

- Time
- local minimums and over-fitting

Settings

- Regression model
- data points $\mathcal{D}_{train} = \{(x_i, y_i), i = 1, 2, ..., n\}$ and $\mathcal{D}_{test} = \{(x_i, y_i), i = 1, 2, ..., m\}$

Settings

- Regression model
- data points $\mathcal{D}_{train} = \{(x_i, y_i), i = 1, 2, ..., n\}$ and $\mathcal{D}_{test} = \{(x_i, y_i), i = 1, 2, ..., m\}$
- **Non parametric scenario :**

centralelille

Settings

- Regression model
- data points $\mathcal{D}_{train} = \{(x_i, y_i), i = 1, 2, ..., n\}$ and $\mathcal{D}_{test} = \{(x_i, y_i), i = 1, 2, ..., m\}$
- **Non parametric scenario :**
  - $y_i = f^*(x_i) + w_i, \ i = 1, 2, ..., n$

Settings

- Regression model
- data points $\mathcal{D}_{train} = \{(x_i, y_i), i = 1, 2, ..., n\}$ and $\mathcal{D}_{test} = \{(x_i, y_i), i = 1, 2, ..., m\}$
- **Non parametric scenario :**
  - $y_i = f^*(x_i) + w_i, \; i = 1, 2, ..., n$
  - $w_i \sim \mathcal{N}(0, 1)$ iid

Settings

- Regression model
- data points $\mathcal{D}_{train} = \{(x_i, y_i), i = 1, 2, ..., n\}$ and $\mathcal{D}_{test} = \{(x_i, y_i), i = 1, 2, ..., m\}$
- **Non parametric scenario :**
  - $y_i = f^*(x_i) + w_i, \ i = 1, 2, ..., n$
  - $w_i \sim \mathcal{N}(0, 1)$ iid
  - $f^* \in \mathcal{H}$

Settings

- Regression model
- data points $\mathcal{D}_{train} = \{(x_i, y_i), i = 1, 2, ..., n\}$ and $\mathcal{D}_{test} = \{(x_i, y_i), i = 1, 2, ..., m\}$
- **Non parametric scenario :**
  - $y_i = f^*(x_i) + w_i, \; i = 1, 2, ..., n$
  - $w_i \sim \mathcal{N}(0, 1)$ iid
  - $f^* \in \mathcal{H}$
- Fix $\mathcal{H}$ a function space

context

- **Goal :** fit function space $\mathcal{H}$ to the model via gradient descent

context

- **Goal :** fit function space $\mathcal{H}$ to the model via gradient descent

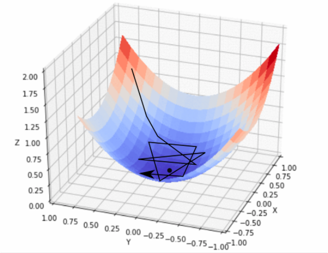


Figure: gradient descent

context

- **Goal :** fit function space $\mathcal{H}$ to the model via gradient descent

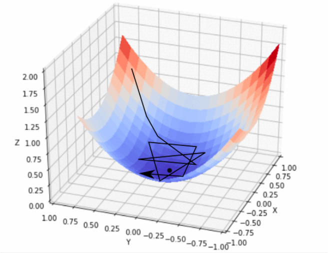


Figure: gradient descent

$$f_{t+1} = f_t + \alpha \nabla \mathcal{L}(f_t), \quad f^* \in \arg\min_{f \in \mathcal{H}} \mathcal{L}(f), \quad \mathcal{L}(f) = \mathbb{E}_y \frac{1}{n} \sum_{i=1}^{n} \phi(y_i, f(x_i))$$

Context

Kernels in use

- Many problems in statistics involve optimizing over function spaces.

Context

Kernels in use

- Many problems in statistics involve optimizing over function spaces.
- Kernels and their associated Reproducing Kernel Hilbert Spaces, give a broad class of functions

Context

Kernels in use

- Many problems in statistics involve optimizing over function spaces.
- Kernels and their associated Reproducing Kernel Hilbert Spaces, give a broad class of functions
- Have geometric properties similar to real euclidean spaces

Context

Kernels in use

- Many problems in statistics involve optimizing over function spaces.
- Kernels and their associated Reproducing Kernel Hilbert Spaces, give a broad class of functions
- Have geometric properties similar to real euclidean spaces
- Give some sort of machinery to manipulate such functions

Context

Kernels in use

- Many problems in statistics involve optimizing over function spaces.
- Kernels and their associated Reproducing Kernel Hilbert Spaces, give a broad class of functions
- Have geometric properties similar to real euclidean spaces
- Give some sort of machinery to manipulate such functions
- What is an **RKHS** ?

kernels : RKHS

Definition

Now let $\mathcal{H}$ be a function Hilbert space, of functions with values on the set $\mathbb{K}$. $\mathcal{H}$ is said to be a **Reproducing Kernel Hilbert Space** if there exists a kernel k over $\mathcal{H}$ such that :

- $\forall x \in \mathbb{K} : k(.,x) \in \mathcal{H}$
- **Reproducing property** :
  $\forall (f,x) \in \mathcal{H} \times \mathbb{K} : f(x) = \langle f, k(.,x) \rangle_{\mathcal{H}}$, with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product over $\mathcal{H}$.

kernels : RKHS

Definition

Now let $\mathcal{H}$ be a function Hilbert space, of functions with values on the set $\mathbb{K}$. $\mathcal{H}$ is said to be a **Reproducing Kernel Hilbert Space** if there exists a kernel k over $\mathcal{H}$ such that :

- $\forall x \in \mathbb{K} : k(.,x) \in \mathcal{H}$
- **Reproducing property** :
  $\forall (f,x) \in \mathcal{H} \times \mathbb{K} : f(x) = \langle f, k(.,x) \rangle_{\mathcal{H}}$, with $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product over $\mathcal{H}$.

Definition

A symmetric bivariate function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is Positive semidefinite (PSD) if for all integers $n \geq 1$ and all $\{x_i\}_{i=1}^{n} \subseteq \mathcal{X}$, the $n \times n$ matrix $\mathbf{K}$ with entries $\mathbf{K}_{i,j} = \mathcal{K}(x_i, x_j)$ is Positive semidefinite.

Kernels : theorems

Connection between RKHS and Kernels

Theorem : RKHS from PSD kernels

Given any PSD kernel $\mathcal{K}$, there is a unique Hilbert space $\mathcal{H}$, in which the kernel $\mathcal{K}$ satisfies the **reproducing property**. $\mathcal{H}$ is said to be an **RKHS** associated to kernel $\mathcal{K}$.

Kernels : theorems

Connection between RKHS and Kernels

Theorem : RKHS from PSD kernels

Given any PSD kernel $\mathcal{K}$, there is a unique Hilbert space $\mathcal{H}$, in which the kernel $\mathcal{K}$ satisfies the **reproducing property**. $\mathcal{H}$ is said to be an **RKHS** associated to kernel $\mathcal{K}$.

Theorem : Kernel from RKHS

Given a function Hilbert space $\mathcal{H}$, suppose the evaluation operator (linear) $L_x : f \in \mathcal{H} \to f(x) \in \mathbb{R}$ is uniformlly bounded, ie there is some universal constant $M > 0$ such that : for all $x \in \mathcal{X}$ and for all $f \in \mathcal{H}$ $|L_x(f)| \leq M \|f\|_{\mathcal{H}}$, then there is a **unique** PSD kernel that satisfies the **reproducing property**.

Kernels : mercer's expansion formula

- We already know that for PSD matrices : $\mathbf{K} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$
- Is there a generalisation for kernels ?

suppose kernel satisfies :

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \mathcal{K}(x, y)^2 d\mathbb{P}(x) d\mathbb{P}(y) < \infty$$

Kernels : mercer's expansion formula

- We already know that for PSD matrices : $\mathbf{K} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T$
- Is there a generalisation for kernels ?

suppose kernel satisfies :

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \mathcal{K}(x,y)^2 d\mathbb{P}(x) d\mathbb{P}(y) < \infty$$

Mercer's theorem

There exists a sequence of eigenfunctions $(\phi_j)_{j=1}^{\infty}$ that form an orthonormal basis of $L^2(\mathcal{X}, \mathbb{P})$ and non-negative eigenvalues $(\mu_i)_{i=1}^{\infty}$ such that :

$$\mathcal{K}(x,y) = \sum_{i=1}^{\infty} \mu_i \phi_i(x) \phi_i(y)$$

Kernels : consequences

- Functions expansion $\forall x \in \mathcal{X}$: $f(x) = \sum_{j=1}^{\infty} \sqrt{\mu_j} a_j \phi_j(x)$ with $a_k = \frac{1}{\sqrt{\mu_k}} \langle f, \phi_k \rangle_{\mathcal{H}}$
- we have the inner products $\langle f, g \rangle_{L^2(\mathcal{X})} = \sum_{j=1}^{\infty} \mu_j a_j b_j$ and $\langle f, g \rangle_{\mathcal{H}} = \sum_{j=1}^{\infty} a_j b_j$

Kernels : consequences

- Functions expansion $\forall x \in \mathcal{X}$: $f(x) = \sum_{j=1}^{\infty} \sqrt{\mu_j} a_j \phi_j(x)$ with $a_k = \frac{1}{\sqrt{\mu_k}} \langle f, \phi_k \rangle_{\mathcal{H}}$
- we have the inner products $\langle f, g \rangle_{L^2(\mathcal{X})} = \sum_{j=1}^{\infty} \mu_j a_j b_j$ and $\langle f, g \rangle_{\mathcal{H}} = \sum_{j=1}^{\infty} a_j b_j$
- recall the fact that :$\langle f, g \rangle_{L^2(\mathcal{X})} = \int_{\mathcal{X}} f(x) g(x) d\mathbb{P}(x)$

Kernels : consequences

- Functions expansion $\forall x \in \mathcal{X}: \; f(x) = \sum_{j=1}^{\infty} \sqrt{\mu_j} a_j \phi_j(x)$ with $a_k = \frac{1}{\sqrt{\mu_k}} \langle f, \phi_k \rangle_{\mathcal{H}}$
- we have the inner products $\langle f, g \rangle_{L^2(\mathcal{X})} = \sum_{j=1}^{\infty} \mu_j a_j b_j$ and $\langle f, g \rangle_{\mathcal{H}} = \sum_{j=1}^{\infty} a_j b_j$
- recall the fact that :$\langle f, g \rangle_{L^2(\mathcal{X})} = \int_{\mathcal{X}} f(x) g(x) d\mathbb{P}(x)$

Representation theorem

Consider a $\mathcal{H}$ to be a **RKHS** defined with a kernel $\mathbb{K}$ over a domain $\mathcal{X}$. let $(x_1, x_2, ..., x_n) \in \mathcal{X}^n$. Let a functional $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ increasing wrt (with respect to) its last variable. Then

$$\min_{f \in \mathcal{F}} \Psi(f(x_1), ..., f(x_n), \|f\|_{\mathcal{F}}^2)$$

is reached at some $f = \sum_{i=1}^{i=n} \alpha_i \mathbb{K}(x_i, .)$

Kernels : consequences

- **Empirical loss** is defined as :$\mathcal{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, f(x_i))$ with $\phi(x, y) = (x - y)^2$
- Gradient descent iterates $f^{t+1}(x_n^1) = f^t(x_n^1) - \alpha_t K(f^t(x_n^1) - y_1^n) = \Big(I_n - \alpha_t K\Big) f^t(x_n^1) + \alpha_t K y_1^n$ with K the empirical matrix $K = \frac{1}{n}\mathbb{K}[x_i, x_j]$ ($Gramm\ matrix$)

Kernels : consequences

- **Empirical loss** is defined as :$\mathcal{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, f(x_i))$ with $\phi(x,y) = (x-y)^2$
- Gradient descent iterates $f^{t+1}(x_n^1) = f^t(x_n^1) - \alpha_t K(f^t(x_n^1) - y_1^n) = \Big(I_n - \alpha_t K\Big) f^t(x_n^1) + \alpha_t K y_1^n$ with K the empirical matrix $K = \frac{1}{n}\mathbb{K}[x_i, x_j]$ (*Gramm matrix*)
- Let $r = \text{rank}(K)$ and $K = UAU^T$ and $S^t = \prod_{\tau=1}^{t-1}\Big(I_n - \alpha_\tau A\Big)$

Kernels : consequences

- **Empirical loss** is defined as :$\mathcal{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i, f(x_i))$ with $\phi(x,y) = (x-y)^2$
- Gradient descent iterates $f^{t+1}(x_n^1) = f^t(x_n^1) - \alpha_t K(f^t(x_n^1) - y_1^n) = \Big(I_n - \alpha_t K\Big) f^t(x_n^1) + \alpha_t K y_1^n$ with K the empirical matrix $K = \frac{1}{n}\mathbb{K}[x_i, x_j]$ (*Gramm matrix*)
- Let $r = \text{rank}(K)$ and $K = UAU^T$ and $S^t = \prod_{\tau=1}^{t-1}\Big(I_n - \alpha_\tau A\Big)$
- Recall A is of form $\text{diag}(\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_r, 0, 0, ..., 0)$ where we have supposed $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq ... \geq \hat{\lambda}_r$
- Denote $\eta_t = \sum_{\tau=0}^{t-1} \alpha_\tau$ the sum of the learning rates.

1 Introduction

2 Stopping rules
- Naive stopping rules
- Bias variance balance : To a sophisticated stopping rule
- Analysis

3 Conclusion

Oracle

Denote by : $R_{OR}(f_t) = \|f^* - f_t\|_n^2$. We might attempt for the **Oracle** rule :

$$\hat{T}_{OR} = \arg\min \left\{ t \in \mathbb{N}, \mid R_{OR}(f_{t+1}) > R_{OR}(f_t) \right\} - 1$$

Oracle

Denote by : $R_{OR}(f_t) = \|f^* - f_t\|_n^2$. We might attempt for the **Oracle** rule :

$$\hat{T}_{OR} = \arg\min \left\{ t \in \mathbb{N}, | \; R_{OR}(f_{t+1}) > R_{OR}(f_t) \right\} - 1$$

But !

Oracle

Denote by : $R_{OR}(f_t) = \|f^* - f_t\|_n^2$. We might attempt for the **Oracle** rule :

$$\hat{T}_{OR} = \arg\min \left\{ t \in \mathbb{N}, \mid R_{OR}(f_{t+1}) > R_{OR}(f_t) \right\} - 1$$

But !

- no mathematical argument showing that the function $t \xrightarrow{\Phi} R_{OR}(f_t) = \|f^* - f_t\|_n^2$, is *convex*

Oracle

Denote by : $R_{OR}(f_t) = \|f^* - f_t\|_n^2$. We might attempt for the **Oracle** rule :

$$\hat{T}_{OR} = \arg\min \left\{ t \in \mathbb{N}, \mid R_{OR}(f_{t+1}) > R_{OR}(f_t) \right\} - 1$$

But !

- no mathematical argument showing that the function $t \xrightarrow{\Phi} R_{OR}(f_t) = \|f^* - f_t\|_n^2$, is *convex*
- **data independent** rule : with $\mathcal{D}_{train} \neq \mathcal{D}'_{train}$ we have the same performance.

Hold out

- Let's suppose that the size of the full data $\{x_i\}_{i=1}^{n}$ is even. S_{te}, and S_{tr} the train/test sets .
- at each iteration, the training data is used to estimate the risk $R_{HO}(f_t) = \frac{1}{n}\sum_{i\in S_{te}}(y_i - f_{tr,t}(x_i))^2$.

Hold out

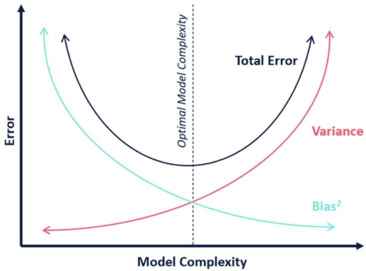
- Let's suppose that the size of the full data $\{x_i\}_{i=1}^{n}$ is even. S_{te}, and S_{tr} the train/test sets .
- at each iteration, the training data is used to estimate the risk $R_{HO}(f_t) = \frac{1}{n}\sum_{i \in S_{te}}(y_i - f_{tr,t}(x_i))^2$.
- Possible rule $\hat{T}_{HO} = \arg\min\left\{t \in \mathbb{N}, R_{HO}(f_{tr,t+1}) > R_{HO}(f_{tr,t})\right\} - 1$

Bias-Variance trade-off

Bias variance principle

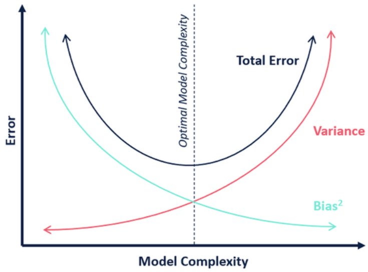
Bias-Variance trade-off

Bias variance principle



Bias-Variance trade-off

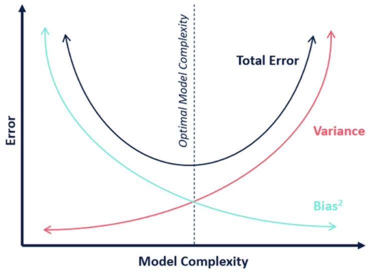
Bias variance principle

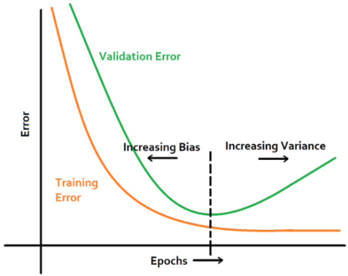


$$\mathbb{E}_{\mathcal{D}}(y - \hat{f})^2 = (y - \mathbb{E}_{\mathcal{D}}\hat{f})^2 + \mathbb{E}_{\mathcal{D}}(\hat{f} - \mathbb{E}_{\mathcal{D}}\hat{f})^2 + \sigma^2$$

Bias-Variance trade-off

Bias variance principle

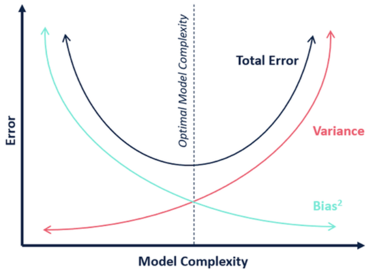


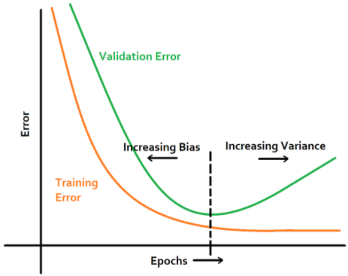


$$\mathbb{E}_{\mathcal{D}}(y - \hat{f})^2 = (y - \mathbb{E}_{\mathcal{D}}\hat{f})^2 + \mathbb{E}_{\mathcal{D}}(\hat{f} - \mathbb{E}_{\mathcal{D}}\hat{f})^2 + \sigma^2$$

Bias-Variance trade-off

Bias variance principle





$$\mathbb{E}_{\mathcal{D}}(y - \hat{f})^2 = (y - \mathbb{E}_{\mathcal{D}}\hat{f})^2 + \mathbb{E}_{\mathcal{D}}(\hat{f} - \mathbb{E}_{\mathcal{D}}\hat{f})^2 + \sigma^2$$

- Bias-Variance decomposition seems good idea !
- How to concertize it ???

Construct a stopping rule

Using basic algebraic manipulations :

Lemma

$$\forall t > 0 : \|f_t - f^*\|_n^2 \leq B_t^2 + V_t$$

where :

$$B_t^2 = \frac{2}{n}\sum_{j=1}^{j=r}(S^t)_{j,j}^2[U^T f^*(x_n^1)]_j^2 + \frac{2}{n}\sum_{j=r+1}^{j=n}[U^T f^*(x_n^1)]_j^2$$

and

$$V_t = \frac{2}{n}\sum_{j=1}^{j=r}(1 - S_{j,j}^t)^2[U^T w]_j^2$$

Properties of Shrinkage matrices

We have the following properties of matrices S^t:

$$0 \leq (S^t)^2_{j,j} \leq \frac{1}{2e\eta_t \hat{\lambda}_j}$$

and

$$\frac{1}{2}\min(1, \eta_t \hat{\lambda}_j) \leq (1 - (S^t)_{j,j})^2 \leq \min(1, \eta_t \hat{\lambda}_j)$$

Properties of Shrinkage matrices

We have the following properties of matrices S^t:

$$0 \leq (S^t)^2_{j,j} \leq \frac{1}{2e\eta_t\hat{\lambda}_j}$$

and

$$\frac{1}{2}\min(1,\eta_t\hat{\lambda}_j) \leq (1-(S^t)_{j,j})^2 \leq \min(1,\eta_t\hat{\lambda}_j)$$

With these properties we can prove that :

Lemma

for all iterations $t = 1, 2, ...$ we have the upper bound :

$$B_t^2 \leq \frac{1}{e\eta_t}$$

Now the Bias term is controled !

How to proceed

- Remains to control to variance term and exploit it's dependency on w_i
- With basic computations we have the following lemma :

Lemma

$$V_t = \frac{2}{n} Tr(UQU^T ww^T)$$

and

$$\mathbb{E} V_t = \frac{2}{n} Tr(Q)$$

where $Q = diag((1 - S^t_{j,j})^2)_{1 \leq j \leq n}$

centralelille

Controling the mean

Using the properties of Shrinkage matrices, we have easily :

Lemma

for all iteration $t > 0$:

$$\frac{\sigma^2}{4}\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2 \leq \mathbb{E}V_t \leq 2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$$

where $\mathcal{R}_K(\epsilon) = \left[\frac{1}{n}\sum_{i=1}^{n}\min(\hat{\lambda}_i, \epsilon^2)\right]^{\frac{1}{2}}$

Controling the mean

Using the properties of Shrinkage matrices, we have easily :

Lemma

for all iteration $t > 0$:

$$\frac{\sigma^2}{4}\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2 \leq \mathbb{E}V_t \leq 2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$$

where $\mathcal{R}_K(\epsilon) = \left[\frac{1}{n}\sum_{i=1}^{n} \min(\hat{\lambda}_i, \epsilon^2)\right]^{\frac{1}{2}}$

- For now all the elements are there !

Controling the mean

Using the properties of Shrinkage matrices, we have easily :

Lemma

for all iteration $t > 0$:

$$\frac{\sigma^2}{4}\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2 \leq \mathbb{E}V_t \leq 2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$$

where $\mathcal{R}_K(\epsilon) = \left[\frac{1}{n}\sum_{i=1}^{n}\min(\hat{\lambda}_i, \epsilon^2)\right]^{\frac{1}{2}}$

- For now all the elements are there !
- Concentration inequality to control V_t suitably

Controling the mean

Using the properties of Shrinkage matrices, we have easily :

Lemma

for all iteration $t > 0$:

$$\frac{\sigma^2}{4}\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2 \leq \mathbb{E}V_t \leq 2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$$

where $\mathcal{R}_K(\epsilon) = \left[\frac{1}{n}\sum_{i=1}^{n}\min(\hat{\lambda}_i, \epsilon^2)\right]^{\frac{1}{2}}$

- For now all the elements are there !
- Concentration inequality to control V_t suitably
- we can wright : $V_t = \sum_{i,j} a_{i,j}(Z_iZ_j - \mathbb{E}(Z_iZ_j))$ with $A = \frac{2}{n}UQU^T$ and $Z_i = w_i$

Controling the mean

Wright in 1973 proved that :

Lemma

For $\mathcal{Q} = \sum_{i,j} a_{i,j}(Z_i Z_j - \mathbb{E}(Z_i Z_j))$ with Z_i are iid Sub-Gaussian random variables, then :

$$\mathbb{P}(|V_t - \mathbb{E}V_t| \geq \delta) \leq \exp\left\{ - c \min\left(\frac{\delta}{\|A\|_{op}}, \frac{\delta^2}{\|A\|_F} \right) \right\}$$

Controling the mean

Wright in 1973 proved that :

Lemma

For $\mathcal{Q} = \sum_{i,j} a_{i,j}(Z_i Z_j - \mathbb{E}(Z_i Z_j))$ with Z_i are iid Sub-Gaussian random variables, then :

$$\mathbb{P}(|V_t - \mathbb{E}V_t| \geq \delta) \leq \exp\left\{ -c \min\left(\frac{\delta}{\|A\|_{op}}, \frac{\delta^2}{\|A\|_F} \right) \right\}$$

As consequence :

$$V_t \leq \mathbb{E}V_t + \delta \ (*)$$

with probability at least

$$1 - \exp\left\{ -4cn\delta \min\left(1, \left(\eta_t \mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^{-1}\right)\right\}$$

Combining the pieces

Now conditioning on the event $(*)$

$$
\begin{aligned}
\|f_t - f^*\|_n^2 \leq B_t^2 + V_t \leq \mathbb{E} V_t + \delta + \frac{1}{e\eta_t} \\
\leq \frac{1}{e\eta_t} + \delta + 2\sigma^2 \eta_t \left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2
\end{aligned}
$$

Combining the pieces

Now conditioning on the event $(*)$

$$\|f_t - f^*\|_n^2 \leq B_t^2 + V_t \leq \mathbb{E} V_t + \delta + \frac{1}{e\eta_t}$$
$$\leq \frac{1}{e\eta_t} + \delta + 2\sigma^2 \eta_t \left(\mathcal{R}_K \left(\frac{1}{\sqrt{\eta_t}} \right) \right)^2$$

putting $\delta = 3\sigma^2 \eta_t \left(\mathcal{R}_K \left(\frac{1}{\sqrt{\eta_t}} \right) \right)^2$, yields :

Combining the pieces

Now conditioning on the event $(*)$

$$\|f_t - f^*\|_n^2 \leq B_t^2 + V_t \leq \mathbb{E}V_t + \delta + \frac{1}{e\eta_t}$$
$$\leq \frac{1}{e\eta_t} + \delta + 2\sigma^2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$$

putting $\delta = 3\sigma^2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$, yields :

$$\|f_t - f^*\|_n^2 \leq \frac{1}{e\eta_t} + 5\sigma^2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2 \qquad (1)$$

as a high probability claim

Combining the pieces

How to link the two quantities $5\sigma^2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$ and $\frac{1}{e\eta_t}$???

Combining the pieces

How to link the two quantities $5\sigma^2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$ and $\frac{1}{e\eta_t}$???

- Empirical radius :

$$\hat{\epsilon_n} = \inf\left\{\epsilon > 0 \;|\mathcal{R}_K(\epsilon) \leq \frac{\epsilon^2}{2e\sigma}\right\}$$

Combining the pieces

How to link the two quantities $5\sigma^2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$ and $\frac{1}{e\eta_t}$???

- Empirical radius :

$$\hat{\epsilon_n} = \inf\left\{\epsilon > 0 \;|\mathcal{R}_K(\epsilon) \leq \frac{\epsilon^2}{2e\sigma}\right\}$$

- Define :

$$\hat{T} := \arg\min\left\{t \in \mathbb{N} \mid \mathcal{R}_k\left(\frac{1}{\sqrt{\eta_k}}\right) > \frac{1}{2e\sigma\eta_k}\right\} - 1$$

Combining the pieces

How to link the two quantities $5\sigma^2\eta_t\left(\mathcal{R}_K\left(\frac{1}{\sqrt{\eta_t}}\right)\right)^2$ and $\frac{1}{e\eta_t}$???

- Empirical radius :

$$\hat{\epsilon_n} = \inf\left\{\epsilon > 0 \,|\mathcal{R}_K(\epsilon) \leq \frac{\epsilon^2}{2e\sigma}\right\}$$

- Define :

$$\hat{T} := \arg\min\left\{t \in \mathbb{N} \mid \mathcal{R}_k\left(\frac{1}{\sqrt{\eta_k}}\right) > \frac{1}{2e\sigma\eta_k}\right\} - 1$$

yields :

$$\|f_t - f^*\|_n^2 \leq \frac{4}{e\eta_t}$$

with probability at least $1 - \exp\{-cn\hat{\epsilon_n}^2\}$

Nice theorem :-)

Theorem

Suppose wa have a **valid step-size**. Then define $\hat{T}$ as previous. There are universal positive constants (c_1, c_2), such that, the following events hold with probability at least $1 - c_1 \exp\left(-c_2 n \hat{\epsilon}_n^2\right)$:

(a) : for all iterations t = 1, 2, ..., $\hat{T}$:

$$\|f_t - f^*\|_n^2 \leq \frac{4}{e\eta_t}$$

(b) : At the iteration $\hat{T}$ we have :

$$\|f_t - f^*\|_n^2 \leq 12\hat{\epsilon}_n^2$$

(c) : Moreover, for all $t > \hat{T}$:

$$\mathbb{E}\left[\|f_t - f^*\|_n^2\right] \geq \frac{\sigma^2}{4}\eta_t \hat{R}_k^2\left(\frac{1}{\eta_k}\right)$$

Numerical illustration

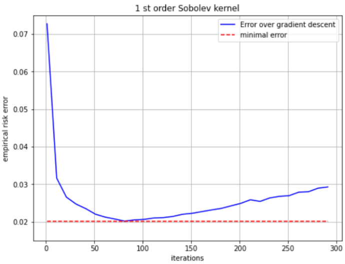


Figure: First order Sobolev kernels

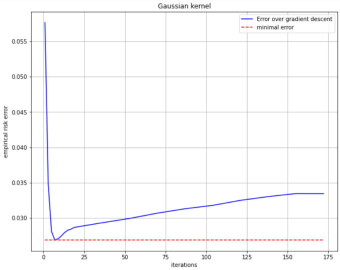


Figure: Gaussian kernel

- Gaussian kernel $T = 9$ iterations theoritically
- Sobolev kernel $T = 70$ iterations theoritically

Numerical illustration

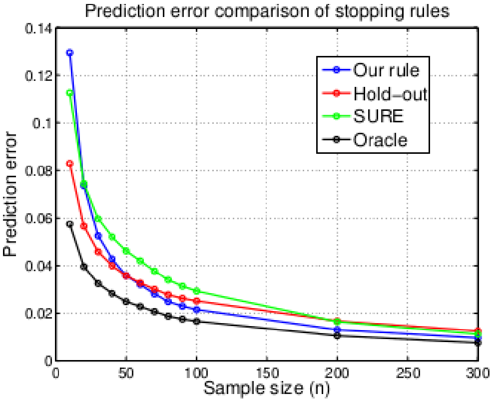


Figure: Different rules

Conclusion

- Garvesh Raskutti, Martin J. Wainwright, Bin Yu, "**Early Stopping and Non-parametric Regression: An Optimal Data-dependent Stopping Rule**", https://jmlr.org/papers/volume15/raskutti14a/raskutti14a.pdf
- Martin J. Wainwright, **High dimensional statistics**, *Cambridge series in statistical and probabilistic mathematics*, Cambridge University press, February 2019.
- Yuting Wei, Fanny Yang, Martin J. Wainwright **Early stopping for kernel boosting algorithms: A general analysis with localized complexities**. https://arxiv.org/abs/1707.01543.

Conclusion

- Michel Ledoux, **The concentraion of measure phenomenon**, *American Mathematical Society*.
- Shahar Mendelson, **Geometric Parameters of Kernel Machines**, *Proceedings of the Conference on Learning Theory (COLT)*. `https://maths-people.anu.edu.au/~mendelso/papers/published/conference/MenKer02.pdf`.
- Roman Vershynin, **High dimensional probability**, *Cambridge University Press*.
- Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar **Foundations of machine learning**, *the MIT press*.

Conclusion