Early stopping points for gradient desccent

An application to least square regression

Mohammed HSSEIN

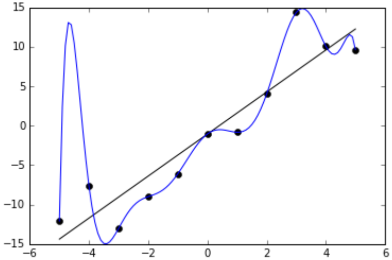
[1]Centrale Lille Institut
Villeneuve d'Ascq, France

Promo : 2021

Plan

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

Context



Context

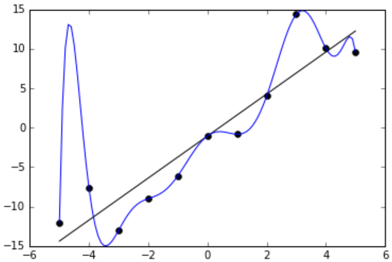


Figure: Overfitting phenomenon

Context

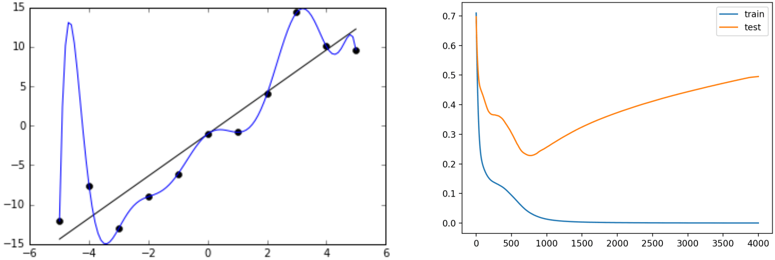


Figure: Overfitting phenomenon

Context

Common problems :

- Running infinite gradient descent iterations, lead to over-fitting !

Context

Common problems :

- Running infinite gradient descent iterations, lead to over-fitting !
- Solution : **Regularization** (Lasso, $\mathcal{L}^1$, ...) !!

But :

Context

Common problems :

- Running infinite gradient descent iterations, lead to over-fitting !
- Solution : **Regularization** (Lasso, $\mathcal{L}^1$, ...) !!

But :

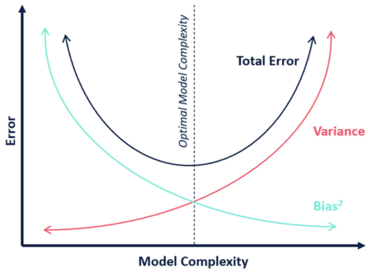
- Cost increases (Time, complexity, ...)
- Alternative : **Early stopping** : find the number of iterations $\hat{T}$, to perform before interrupting the training procedure.
- Motivated by the **Bias-Variance** balance :

Bias-Variance trade-off

Bias variance principle

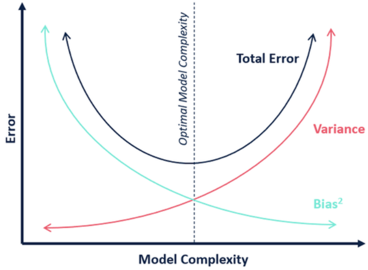
Bias-Variance trade-off

Bias variance principle



Bias-Variance trade-off

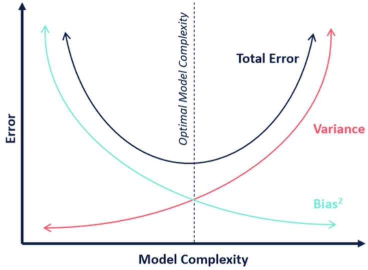
Bias variance principle

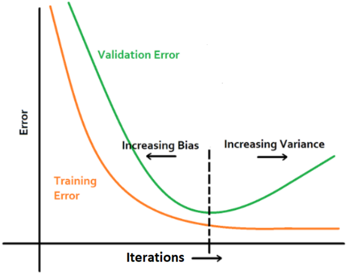


$$\mathbb{E}_{\mathcal{D}}(y-\hat{f})^2 = (y-\mathbb{E}_{\mathcal{D}}\hat{f})^2 + \mathbb{E}_{\mathcal{D}}(\hat{f}-\mathbb{E}_{\mathcal{D}}\hat{f})^2 + \sigma^2$$

Bias-Variance trade-off

Bias variance principle

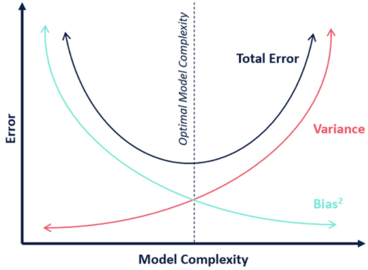


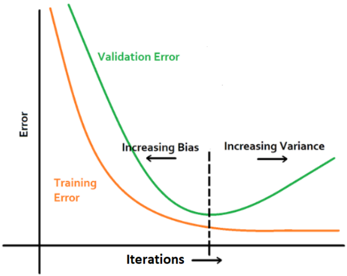


$$\mathbb{E}_{\mathcal{D}}(y - \hat{f})^2 = (y - \mathbb{E}_{\mathcal{D}}\hat{f})^2 + \mathbb{E}_{\mathcal{D}}(\hat{f} - \mathbb{E}_{\mathcal{D}}\hat{f})^2 + \sigma^2$$

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 term and variance term behave in opposite sens
- Controling their evolution may lead to consistant rules

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

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

Using RKHS setting gives broad class of functions and algebraic properties

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

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 :

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

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.

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 !

- 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.
- **Not computable in practice** !!!

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

Construct a stopping rule from the bias variance tradeoff

The bias variance balance principle gives a way to construct stopping rules

Construct a stopping rule from the bias variance tradeoff

The bias variance balance principle gives a way to construct stopping rules

- The bias term involves $f^* \in \arg\min_{f \in \mathcal{H}} \mathcal{L}(f)$ and thus unknown !

Construct a stopping rule from the bias variance tradeoff

The bias variance balance principle gives a way to construct stopping rules

- The bias term involves $f^* \in \arg\min_{f \in \mathcal{H}} \mathcal{L}(f)$ and thus unknown !
- The variance term involves randomness of the model

Construct a stopping rule from the bias variance tradeoff

The bias variance balance principle gives a way to construct stopping rules

- The bias term involves $f^* \in \arg\min_{f \in \mathcal{H}} \mathcal{L}(f)$ and thus unknown !
- The variance term involves randomness of the model
- **Idea** :

Construct a stopping rule from the bias variance tradeoff

The bias variance balance principle gives a way to construct stopping rules

- The bias term involves $f^* \in \arg\min_{f \in \mathcal{H}} \mathcal{L}(f)$ and thus unknown !
- The variance term involves randomness of the model
- **Idea** :
  - Upper bound bias term carefully

Construct a stopping rule from the bias variance tradeoff

The bias variance balance principle gives a way to construct stopping rules

- The bias term involves $f^* \in \arg\min_{f \in \mathcal{H}} \mathcal{L}(f)$ and thus unknown !
- The variance term involves randomness of the model
- **Idea** :
  - Upper bound bias term carefully
  - Control the variance term carefully using the randomness of the model

Construct a stopping rule from the bias variance tradeoff

The bias variance balance principle gives a way to construct stopping rules

- The bias term involves $f^* \in \arg\min_{f\in\mathcal{H}} \mathcal{L}(f)$ and thus unknown !
- The variance term involves randomness of the model
- **Idea** :
  - Upper bound bias term carefully
  - Control the variance term carefully using the randomness of the model
  - If the bounds are computable : stop when the terms are equivalent

Construct a stopping rule from the bias variance tradeoff

The bias variance balance principle gives a way to construct stopping rules

- The bias term involves $f^* \in \arg\min_{f \in \mathcal{H}} \mathcal{L}(f)$ and thus unknown !
- The variance term involves randomness of the model
- **Idea** :
  - Upper bound bias term carefully
  - Control the variance term carefully using the randomness of the model
  - If the bounds are computable : stop when the terms are equivalent
- **RKHS** mathematical setting

RKHS setting : consequences

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

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

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

RKHS setting : consequences

Representation theorem

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

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

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

- Functions expansion $\forall x \in \mathcal{X}$: $f(x) = \sum_{j=1}^{\infty} \sqrt{\mu_j} a_j \phi_j(x)$ with $a_k = \frac{1}{\sqrt{\mu_k}} \langle f, \phi_k \rangle_{\mathcal{H}}$

RKHS setting : consequences

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

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

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

- 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 $\langle f, g \rangle_{\mathcal{H}} = \sum_{j=1}^{\infty} a_j b_j$

RKHS setting : consequences

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

- 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 :$\langle f, g \rangle_{L^2(\mathcal{X})} = \int_{\mathcal{X}} f(x) g(x) d\mathbb{P}(x)$

RKHS consequences

- Define the **Local Rademacher upper bound** :
  $\mathcal{R}_K(\epsilon) = \left[\frac{1}{n}\sum_{i=1}^{n} \min(\hat{\lambda}_i, \epsilon^2)\right]^{\frac{1}{2}}$ where $K = UAU^T$ the empirical kernel matrix, $r = \mathrm{rank}(K)$ its rank, and finally :
  $\mathrm{Sp}_{\mathbb{R}}(K) = \{\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3, ..., \geq \hat{\lambda}_r\}$

RKHS consequences

- Define the **Local Rademacher upper bound** :
  $\mathcal{R}_K(\epsilon) = \left[\frac{1}{n}\sum_{i=1}^{n}\min(\hat{\lambda}_i, \epsilon^2)\right]^{\frac{1}{2}}$ where $K = UAU^T$ the empirical kernel matrix, $r = \text{rank}(K)$ its rank, and finally : $\text{Sp}_{\mathbb{R}}(K) = \{\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3, ..., \geq \hat{\lambda}_r\}$
- Define the empirical radius :$\hat{\epsilon_n} = \inf\left\{\epsilon > 0 \;|\mathcal{R}_K(\epsilon) \leq \frac{\epsilon^2}{2e\sigma}\right\}$

RKHS consequences

- Define the **Local Rademacher upper bound** :
  $\mathcal{R}_K(\epsilon) = \left[\frac{1}{n}\sum_{i=1}^{n} \min(\hat{\lambda}_i, \epsilon^2)\right]^{\frac{1}{2}}$ where $K = UAU^T$ the empirical kernel matrix, $r = \mathrm{rank}(K)$ its rank, and finally : $\mathrm{Sp}_{\mathbb{R}}(K) = \{\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \hat{\lambda}_3, ..., \geq \hat{\lambda}_r\}$
- Define the empirical radius :$\hat{\epsilon_n} = \inf\left\{\epsilon > 0 \; |\mathcal{R}_K(\epsilon) \leq \frac{\epsilon^2}{2e\sigma}\right\}$
- Define the stopping time :

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

where η_t is the the sum of step sizes (*learning rates*) untill time $t - 1$

Theorem : regression case (fixed design)

For regression problems, [2] have proved that :

Theorem : Raskutti, Wainwright [2]

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 $t > \hat{T}$:

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

Remarks

Remarks

- Radius $\hat{\epsilon}_n$ depends certainly on the data, yet it depends only on the entries, not on the outputs $\{y_i\}_{i=1}^n$

Remarks

- Radius $\hat{\epsilon}_n$ depends certainly on the data, yet it depends only on the entries, not on the outputs $\{y_i\}_{i=1}^n$
- The results apply for a specific loss function, which is the **mean squared loss**, and **regression analysis** !

Remarks

- Radius $\hat{\epsilon}_n$ depends certainly on the data, yet it depends only on the entries, not on the outputs $\{y_i\}_{i=1}^n$
- The results apply for a specific loss function, which is the **mean squared loss**, and **regression analysis** !
- Generalisation error is not bounded !

Remarks

- Radius $\hat{\epsilon}_n$ depends certainly on the data, yet it depends only on the entries, not on the outputs $\{y_i\}_{i=1}^n$
- The results apply for a specific loss function, which is the **mean squared loss**, and **regression analysis** !
- Generalisation error is not bounded !
- Claiming results in high probability is true only if the decay of the empirical radius is of a minimum $n^{\frac{\alpha-1}{2}}$ for $\alpha > 0$

Remarks

- Radius $\hat{\epsilon}_n$ depends certainly on the data, yet it depends only on the entries, not on the outputs $\{y_i\}_{i=1}^n$
- The results apply for a specific loss function, which is the **mean squared loss**, and **regression analysis** !
- Generalisation error is not bounded !
- Claiming results in high probability is true only if the decay of the empirical radius is of a minimum $n^{\frac{\alpha-1}{2}}$ for $\alpha > 0$
- Norms control : the paper uses the property : $\|f\|_{\mathcal{H}} \leq B$ for some $B > 0$ for all $f \in \mathbb{B}_{\mathcal{H}}(f^*, 1)$ and as consequence $\|f\|_\infty \leq B$ for all $f \in \mathbb{B}_{\mathcal{H}}(f^*, 1)$.

Numerical illustration

$f^*(x) = |x - \frac{1}{2}| - \frac{1}{2}$ and $(x, y) \in [0, 1] \times [0, 1]$ and $x_i = \frac{i}{n}$ $i = 0, ..., n - 1$

Numerical illustration

$f^*(x) = |x - \frac{1}{2}| - \frac{1}{2}$ and $(x, y) \in [0, 1] \times [0, 1]$ and $x_i = \frac{i}{n}\ i = 0, \ldots, n-1$

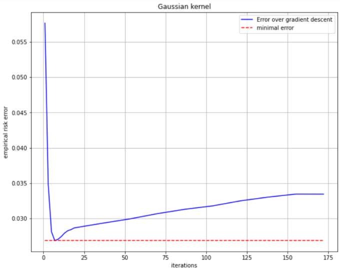


Figure: Gaussian kernel

Numerical illustration

$f^*(x) = |x - \frac{1}{2}| - \frac{1}{2}$ and $(x, y) \in [0,1] \times [0,1]$ and $x_i = \frac{i}{n}\ i = 0, \ldots, n-1$

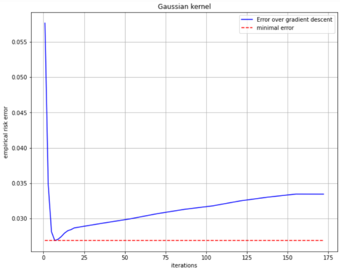


Figure: Gaussian kernel

- Gaussian kernel $T = 9$ iterations (infinitely differentiable functions)

Numerical illustration

$f^*(x) = |x - \frac{1}{2}| - \frac{1}{2}$ and $(x, y) \in [0, 1] \times [0, 1]$ and $x_i = \frac{i}{n}\ i = 0, ..., n - 1$

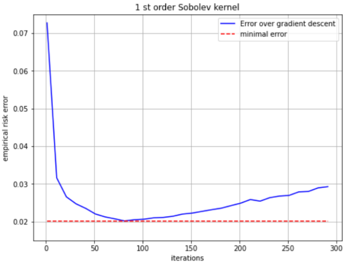


Figure: First order Sobolev kernels

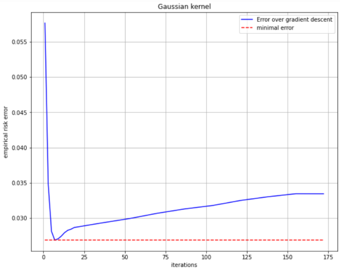


Figure: Gaussian kernel

- Gaussian kernel $T = 9$ iterations (infinitely differentiable functions)

Numerical illustration

$f^*(x) = |x - \frac{1}{2}| - \frac{1}{2}$ and $(x, y) \in [0, 1] \times [0, 1]$ and $x_i = \frac{i}{n}$ $i = 0, ..., n - 1$

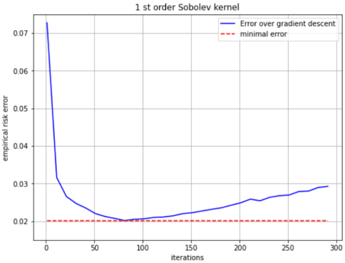


Figure: First order Sobolev kernels

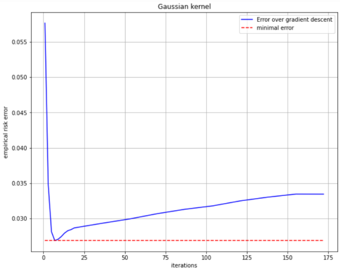


Figure: Gaussian kernel

- Gaussian kernel $T = 9$ iterations (infinitely differentiable functions)
- Sobolev kernel $T = 70$ iterations (Lipchitz functions)

Numerical illustration

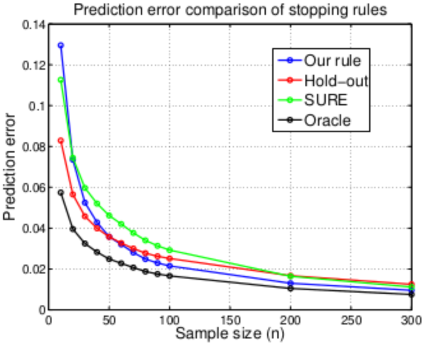


Figure: Different rules

Conclusion

Bibliography

- 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.

Bibliography

- 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*.