# A walk in random forests

Erwan Scornet (LSTA - Paris 6),
supervised by Gérard Biau and Jean-Philippe Vert (Institut Curie)

Journées MAS 2016

# Background on random forests

Random forests are a class of algorithms used to solve regression and classification problems

- They are often used in applied fields since they handle high-dimensional settings.
- They have good predictive power and can outperform state-of-the-art methods.

# Background on random forests

Random forests are a class of algorithms used to solve regression and classification problems

- They are often used in applied fields since they handle high-dimensional settings.
- They have good predictive power and can outperform state-of-the-art methods.

But mathematical properties of random forests remain a bit magical.

1. Construction of random forests
2. Centred Forests
3. Median forests
4. Consistency of Breiman forests

# General framework of the presentation

## Regression setting

We are given a training set $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ where the pairs $(X_i, Y_i) \in [0, 1]^d \times \mathbb{R}$ are *i.i.d.* distributed as $(X, Y)$.

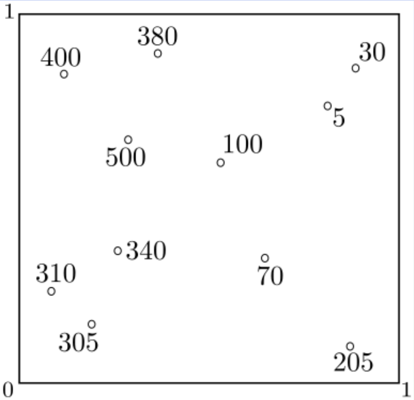
We assume that

$$Y = m(\mathbf{X}) + \varepsilon.$$

We want to build an estimate of the regression function $m$ using random forest algorithm.

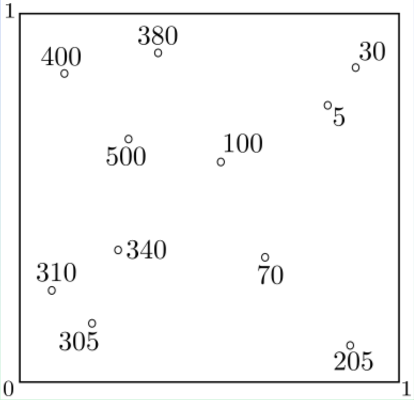
# How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

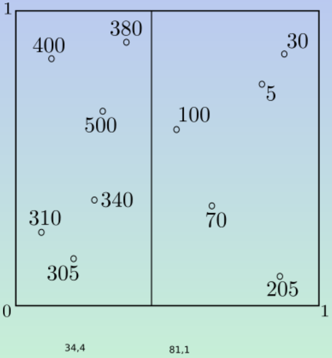


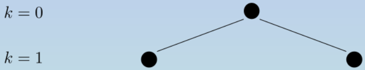
# How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.



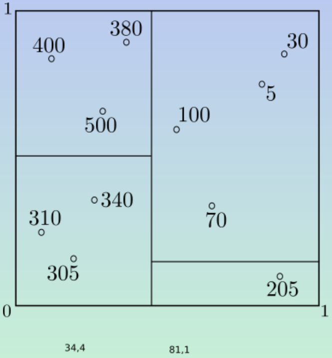

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

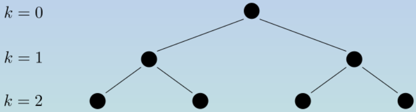


# How to build a tree?

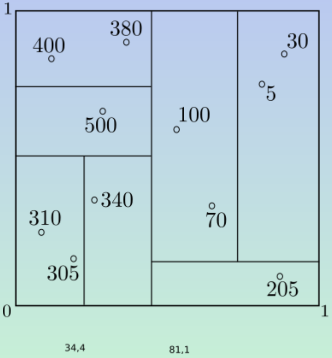
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

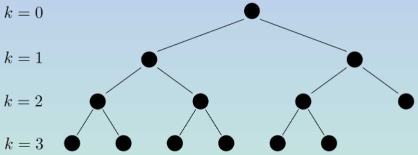


# How to build a tree?

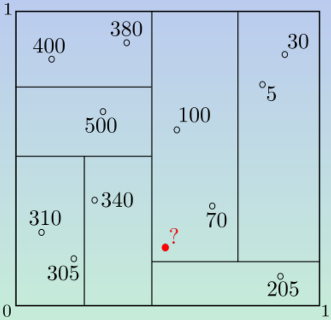
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

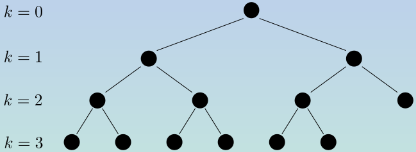


# How to build a tree?

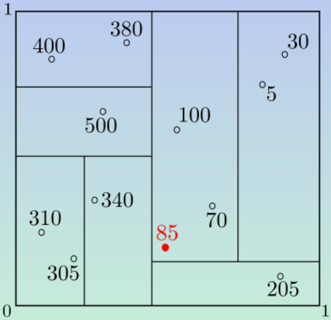
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

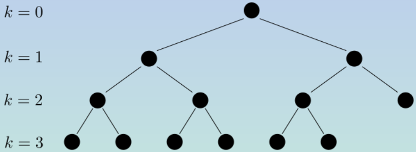


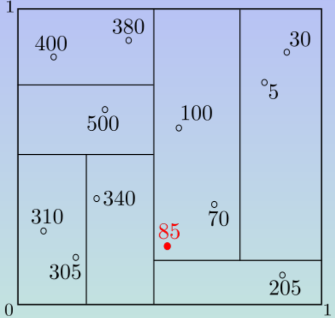
# How to build a tree?

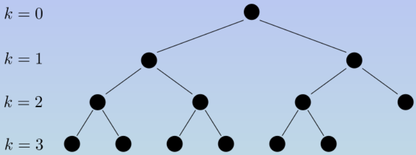
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.



# How to build a tree?



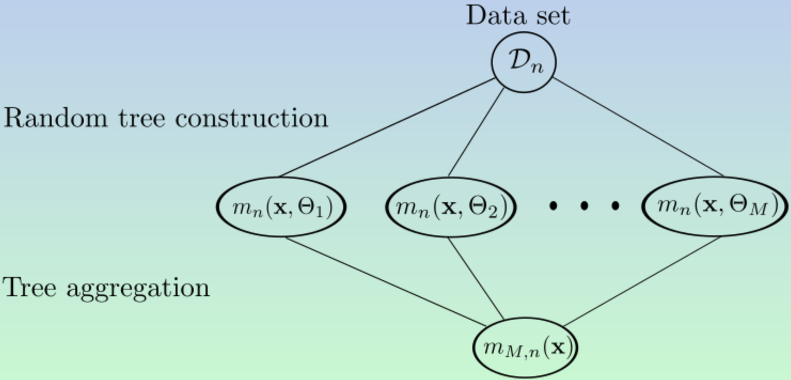

Breiman Random forests are defined by

1. A splitting rule : minimize the variance within the resulting cells.
2. A stopping rule : stop when each cell contains less than `nodesize` $= 2$ observations.

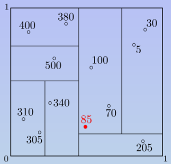
# Construction of random forests

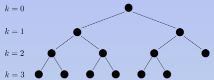
## Randomness in tree construction

- Resample the data set via bootstrap;
- At each node, preselect a subset of `mtry` variables eligible for splitting.



# Construction of Breiman forests




## Breiman tree

- Select $a_n$ observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- Stop when each cell contains less than `nodesize` observations.

# Literature

- Random forests were created by Breiman [2001].
- Many theoretical results focus on simplified version on random forests, whose construction is independent of the dataset.
  [Biau et al., 2008, Biau, 2012, Genuer, 2012, Zhu et al., 2012, Arlot and Genuer, 2014].
- Analysis of more data-dependent forests:
  - Asymptotic normality of random forests [Wager, 2014, Mentch and Hooker, 2015].
  - Variable importance [Louppe et al., 2013].
- Literature review on random forests:
  - Methodological review [Criminisi et al., 2011, Boulesteix et al., 2012].
  - Theoretical review [Biau and Scornet, 2016].

# Different types of forests

| Centred forest | | |
|---|---|---|
| | | |
| | | |

# Different types of forests

| Centred forest | | |
|---|---|---|
| Independent of $X_i$ and $Y_i$ | | |
| | | |

# Different types of forests

| Centred forest | | |
|---|---|---|
| Independent of $X_i$ and $Y_i$ | | |
|  | | |

# Different types of forests

| Centred forest | | Breiman's forests |
|---|---|---|
| Independent of $X_i$ and $Y_i$ | | |
|  | | |

# Different types of forests

| Centred forest | | Breiman's forests |
|---|---|---|
| Independent of $X_i$ and $Y_i$ | | Dependent on $X_i$ and $Y_i$ |
|  | | |

# Different types of forests

| Centred forest | | Breiman's forests |
|---|---|---|
| Independent of $X_i$ and $Y_i$ | | Dependent on $X_i$ and $Y_i$ |
|  | |  |

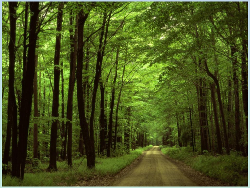
# Different types of forests

| Centred forest | Median forests | Breiman's forests |
|---|---|---|
| Independent of $X_i$ and $Y_i$ | | Dependent on $X_i$ and $Y_i$ |
|  | |  |

# Different types of forests

| Centred forest | Median forests | Breiman's forests |
|---|---|---|
| Independent of $X_i$ and $Y_i$ | Independent of $Y_i$ | Dependent on $X_i$ and $Y_i$ |
|  | |  |

# Different types of forests

| Centred forest | Median forests | Breiman's forests |
|---|---|---|
| Independent of $X_i$ and $Y_i$ | Independent of $Y_i$ | Dependent on $X_i$ and $Y_i$ |
|  |  |  |

# Tree consistency

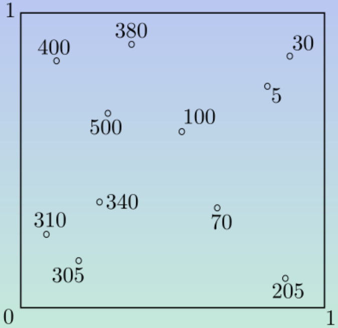
For a tree whose construction is independent of data, if

1. $\mathrm{diam}(A_n(\mathbf{X})) \to 0$, in probability;
2. $N_n(A_n(\mathbf{X})) \to \infty$, in probability;

then the tree is consistent, that is

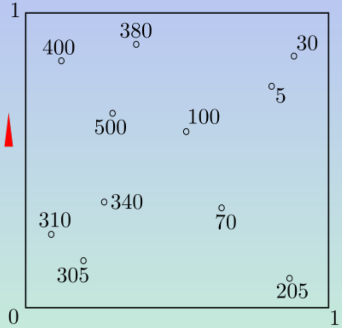
$$\lim_{n\to\infty} \mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

# Centered forests



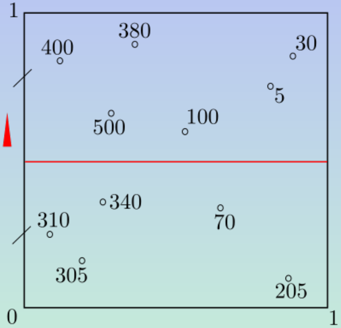

$k = 0$

# Centered forests



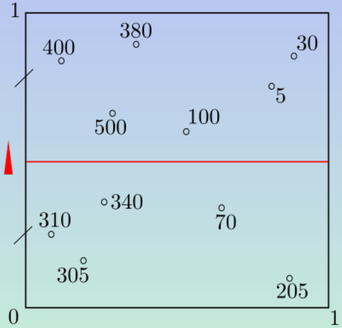

$k = 0$

# Centered forests



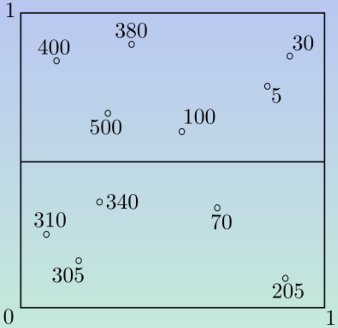

$k = 0$

# Centered forests



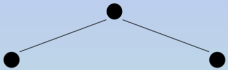

$k = 0$

# Centered forests

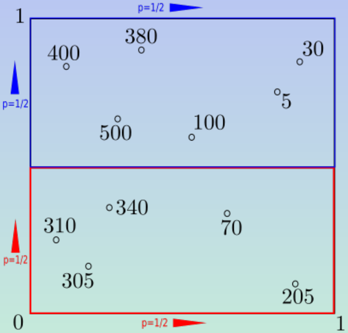



$k = 0$

$k = 1$

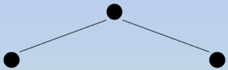


# Centered forests

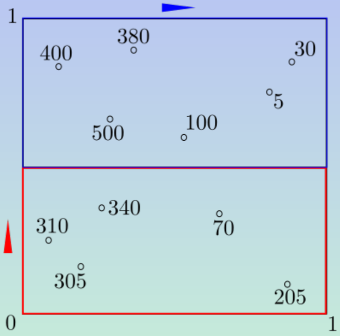


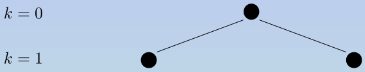

$k = 0$

$k = 1$

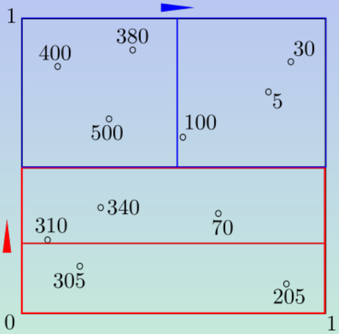


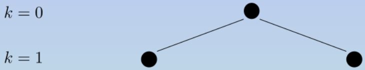
# Centered forests



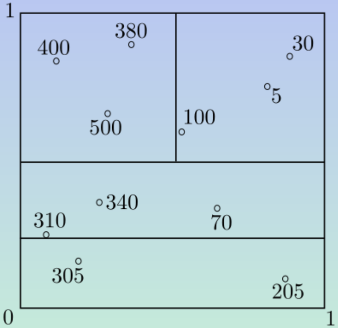



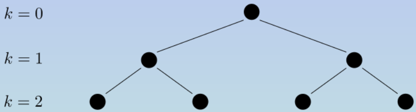
# Centered forests



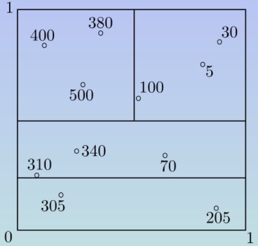



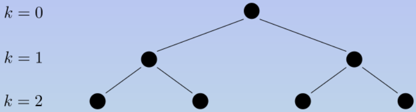
# Centered forests



# Centered forests

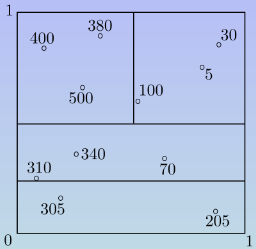


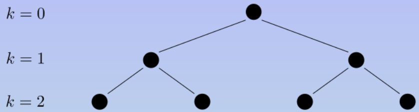




**Theorem (Biau [2012])**

*Under proper regularity hypothesis, provided $k \to \infty$ and $n/2^k \to \infty$, the centred random forest is consistent.*

# Centered forests




**Theorem (Biau [2012])**

*Under proper regularity hypothesis, provided $k \to \infty$ and $n/2^k \to \infty$, the centred random forest is consistent.*

$\rightarrow$ Forest consistency results from the consistency of each tree.

$\rightarrow$ Trees are not fully developed.

# Construction of Breiman/Median forests

## Breiman tree

- Select $a_n$ observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- Stop when each cell contains less than `nodesize` observations.

# Construction of Breiman/Median forests

**Breiman tree**

- Select $a_n$ observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- At each cell, select randomly mtry coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- Stop when each cell contains less than nodesize observations.

**Median tree**

- Select $a_n$ observations without replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- At each cell, select randomly $\texttt{mtry} = 1$ coordinate among $\{1, \ldots, d\}$.
- Split at the location of the empirical median of $X_i$.
- Stop when each cell contains exactly $\texttt{nodesize} = 1$ observation.

**Theorem**

Assume that **(H1)** is satisfied. Then, provided $a_n \to \infty$ and $a_n/n \to 0$, median forests are consistent, i.e.,

$$\lim_{n\to\infty} \mathbb{E}\left[m_n(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

**Remarks**

- Good trade-off between simplicity of centred forests and complexity of Breiman's forests.
- First consistency results for fully grown trees.
- Each tree is not consistent but the forest is, because of subsampling.

# Construction of Breiman forests

**Breiman tree**

- Select $a_n$ observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- Stop when each cell contains less than `nodesize` observations.

# Construction of Breiman forests

## Breiman tree

- Select $a_n$ observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- Stop when each cell contains less than `nodesize` observations.

## Modified Breiman tree

- Select $a_n$ observations without replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- At each cell, select randomly `mtry` coordinates among $\{1, \ldots, d\}$.
- Split at the location that minimizes the square loss.
- Stop when the number of cells is exactly $t_n$.

# Assumption **(H1)**

Additive regression model:

$$Y = \sum_{i=1}^{d} m_i(\mathbf{X}^{(i)}) + \varepsilon,$$

where

- $\mathbf{X}$ is uniformly distributed on $[0,1]^d$,
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\varepsilon$ independent of $\mathbf{X}$,
- Each model component $m_i$ is continuous.

**Theorem [Scornet et al., 2015]**

Assume that **(H1)** is satisfied. Then, provided $a_n \to \infty$ and $t_n(\log a_n)^9/a_n \to 0$, random forests are consistent, i.e.,

$$\lim_{n\to\infty} \mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

**Remarks**

- First consistency result for Breiman's original forest.
- Consistency of CART.

## Theorem [Scornet et al., 2015]

Assume that **(H1)** and **(H2.1)** are satisfied and let $t_n = a_n$. Then, provided $a_n \to \infty$ and $a_n \log n / n \to 0$, random forests are consistent, i.e.,

$$\lim_{n \to \infty} \mathbb{E}\left[m_{\infty,n}(\mathbf{X}) - m(\mathbf{X})\right]^2 = 0.$$

**Remarks**:

- First result for fully developed forest;
- Importance of subsampling;
- One major drawback: **(H2)** seems impossible to verify.

- Assume that

$$Y = \sum_{i=1}^{S} m_i(\mathbf{X}^{(i)}) + \varepsilon,$$

for some $S < d$.

- Denote by $j_{1,n}(\mathbf{X}), \dots, j_{k,n}(\mathbf{X})$ the first $k$ cut directions used to construct the cell containing $\mathbf{X}$.

**Proposition [Scornet et al., 2015]**

Let $k \in \mathbb{N}^\star$ and $\xi > 0$. Under appropriate assumptions, with probability $1 - \xi$, for all $n$ large enough, we have, for all $1 \leq q \leq k$,

$$j_{q,n}(\mathbf{X}) \in \{1, \dots, S\}.$$

# Conclusion

- Centred forests: their consistency results from the consistency of each tree.

  $\rightarrow$ No benefits from using a forest instead of a single tree.

- Median forests: the aggregation process can turn inconsistent trees into a consistent forest.

  $\rightarrow$ Benefits from using a random forest compared to a single tree.

- Breiman forests: consistent as well as CART procedure. The splitting criterion asymptotically selects relevant features.

  $\rightarrow$ Good performance in high-dimensional settings.

# Merci pour votre attention !

S. Arlot and R. Genuer. Analysis of purely random forests bias. 2014.

G. Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13:1063–1095, 2012.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

A.-L. Boulesteix, S. Janitza, J. Kruppa, and I.R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2:493–507, 2012.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7:81–227, 2011.

R. Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24: 543–562, 2012.

G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 431–439, 2013.

L. Mentch and G. Hooker. Ensemble trees and CLTs: Statistical inference for supervised learning. *Journal of Machine Learning Research, in press*, 2015.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43: 1716–1741, 2015.

S. Wager. Asymptotic theory for random forests. arXiv:1405.0352, 2014.

R. Zhu, D. Zeng, and M.R. Kosorok. Reinforcement learning trees. 2012.

# How to deal with fully developed Breiman forest?

Let

$$\psi_{i,j}(Y_i, Y_j) = \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \middle| \mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \ldots, \mathbf{X}_n, Y_i, Y_j\right]$$

$$\text{and} \quad \psi_{i,j} = \mathbb{E}\left[\mathbb{1}_{\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i} \mathbb{1}_{\mathbf{X} \overset{\Theta'}{\leftrightarrow} \mathbf{X}_j} \middle| \mathbf{X}, \Theta, \Theta', \mathbf{X}_1, \ldots, \mathbf{X}_n\right].$$

One assumption **(H2.1)**:

$$\lim_{n \to \infty} (\log a_n)^{2p-2} (\log n)^2 \mathbb{E}\left[\max_{\substack{i,j \\ i \neq j}} |\psi_{i,j}(Y_i, Y_j) - \psi_{i,j}|\right]^2 = 0.$$