# Hilbertraum

## Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

## Zusammenfassung

Ein Hilbertraum ist ein completeer Skalarprodukt-Raum: jede Cauchy sequence seiner Elemente hat einen Grenzwert in diesem Raum. Drei Beispiele in der machine learning (ML) sind: ein endlichdimensionaler Euclidean space, ein Raum von Zufallsvariablen (ZV) mit endlicher Varianz auf einem gemeinsamen probability space und ein reproducing kernel Hilbert space (RKHS). Die lineare Regression verwendet einen Euclidean space als ihren Merkmalsraum und zur Parametrisierung ihres Hypothesenraum. Die optimale lineare und nichtlineare Schätzung einer ZV sind orthogonale Projektionen im Hilbertraum von Zufallsvariablen mit endlicher Varianz. Eine Kernel-Methode verwendet einen RKHS als ihren transformierten Merkmalsraum und als ihren Hypothesenraum.

## Definition

Man betrachte die Gradientenabstieg für die lineare Regression oder eine beliebige iterative Training-Methode, deren Hypothesenraum ein metric space ist: Sie erzeugt die Folge $h^{(1)}, h^{(2)}, \ldots$ von Hypothesen. Damit die Folge gegen eine optimale Hypothese $h^{\star}$ konvergiert, müssen die Hypothesen eine Cauchy sequence bilden (Rudin, 1976, Thm. 3.11, Teil (a)): Für jeden Abstand $\epsilon > 0$, wie klein auch immer, gibt es einen Index $N$, sodass der Abstand zwischen $h^{(r)}$ und $h^{(r')}$ für alle $r, r' \geq N$ unter $\epsilon$ liegt (Rudin, 1976, Def. 3.8). Eine Cauchy sequence zu sein, ist allein nicht ausreichend. Wenn der zugrunde liegende metric space complete ist, dann konvergiert jede Cauchy sequence in diesem Raum gegen einen Punkt in diesem Raum. Ein Beispiel für einen completeen metric space ist ein Euclidean space $\mathbb{R}^d$ endlicher Dimension $d$.

Ein Hilbertraum ist eine Verallgemeinerung eines Euclidean space auf möglicherweise unendliche Dimensionen. Insbesondere ist ein Hilbertraum $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ ein Skalarprodukt-Raum, der complete ist, d.h. in dem jede Cauchy sequence seiner Elemente einen Grenzwert hat, der wiederum zu $\mathcal{H}$ gehört (Rudin, 1987, Def. 4.4) (siehe Abb. 1).

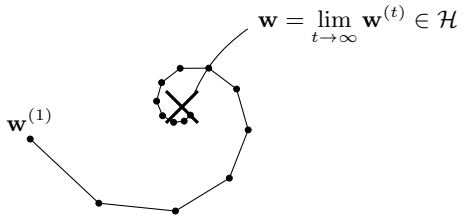


Abbildung 1: Vollständigkeit. Die Elemente $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ einer Cauchy sequence (Punkte) liegen immer näher beieinander, und ihr Grenzwert $\mathbf{w} = \lim_{t\to\infty} \mathbf{w}^{(t)}$ (Kreuz) gehört wieder zum Hilbertraum. Daten generiert von `pythondemos/hilbertspace.py`

Die Elemente von $\mathcal{H}$ werden Vektoren genannt, ob es sich um Arrays von Zahlen, Zufallsvariablen (ZV) oder Funktionen handelt: Was sie zu Vektoren macht, ist, dass sie addiert und skaliert werden können (siehe Vektorraum). Das Skalarprodukt induziert eine Norm $\|\mathbf{u}\|_{\mathcal{H}} := \sqrt{\langle \mathbf{u}, \mathbf{u}\rangle}$ und wiederum eine Metrik $d(\mathbf{u}, \mathbf{v}) := \|\mathbf{u} - \mathbf{v}\|_{\mathcal{H}}$ (Rudin, 1987, Abschn. 4.2–4.3, Def. 4.4) (siehe Skalarprodukt), sodass eine Cauchy sequence von $\mathcal{H}$ eine ist, deren Vektoren schließlich innerhalb eines beliebigen vorgegebenen Norm Abstand voneinander liegen.

Drei Beispiele von Hilberträumen werden in der gesamten machine learning (ML) verwendet: ein Euclidean space $\mathbb{R}^d$ als Merkmalsraum, ein Raum von Zufallsvariablen mit endlicher Varianz auf einem gemeinsamen probability space und ein reproducing kernel Hilbert space (RKHS) von Funktionen. Der erste von ihnen trägt das Standard-Skalarprodukt $\langle \mathbf{u}, \mathbf{v}\rangle = \mathbf{u}^\top \mathbf{v}$, und seine Vollständigkeit folgt aus der der reellen Zahlen (Rudin, 1976, Thm. 3.11). Die lineare Regression verwendet den Euclidean space $\mathbb{R}^d$ in zwei Rollen. Die Merkmalsvektoren $\mathbf{x} \in \mathbb{R}^d$ der Datenpunkte, auf die sie angewendet wird, sind Vektoren dieses Raumes, ebenso wie die Gewichte $\mathbf{w} \in \mathbb{R}^d$, die ihren Hypothesenraum $\{h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle : \mathbf{w} \in \mathbb{R}^d\}$ parametrisieren. Die Gradientenabstieg, die nach diesen Gewichte sucht, erzeugt eine Folge $\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \ldots$ von Modellparameter, die in Abb. 1 dargestellte Folge. Die Modellparameter sind in der zweiten Rolle Vektoren, sodass die Bedingung des einleitenden Absatzes hier erfüllt ist: $\mathbb{R}^d$ ist complete, und eine Cauchy sequence von Modellparameter konvergiert gegen Modellparameter.

Das zweite Beispiel ist die Menge aller Zufallsvariablen $x$ mit endlicher Varianz, definiert auf einem gemeinsamen probability space (Gray and Davisson, 2004, Abschn. 5.8.1). Hier werden zwei Zufallsvariablen identifiziert, wenn der Erwartungswert ihrer quadrierten Differenz Null ist, $\mathbb{E}\{(x - x')^2\} = 0$, d.h. wenn $x = x'$ mit Wahrscheinlichkeit eins. Mit dieser Identifikation ist der Erwartungswert $\langle x, x'\rangle := \mathbb{E}\{xx'\}$ ein Skalarprodukt (Gray and Davisson, 2004, Abschn. 5.8.1), und der resultierende Raum ist complete (Gray and Davisson, 2004, Lemma 5.1).

Im Hilbertraum von Zufallsvariablen mit endlicher Varianz auf einem ge-

meinsamen probability space ist die optimale lineare Schätzung eine orthogonale Projektion (siehe Abb. 2). Ein linearer Schätzer $\widehat{y}$ einer ZV $y$ aus einer beobachteten ZV $x$ ist ein Element des von $x$ aufgespannten Unterraum $\{ax \,:\, a \in \mathbb{R}\}$. Der entsprechende estimation error $y - \widehat{y}$ wird durch die induzierte Norm $\sqrt{\langle y - \widehat{y}, y - \widehat{y}\rangle}$ gemessen. Der kleinste Fehler wird durch den linearen Schätzer erzielt, dessen Fehler orthogonal zu $x$ ist (Gray and Davisson, 2004, Thm. 4.9),

$$\begin{aligned} 0 &= \langle y - \widehat{y}, x\rangle \\ &= \mathbb{E}\{(y - \widehat{y})x\}. \end{aligned}$$

Wenn die Zufallsvariablen $y, x$ null-Mittelwert sind, bedeutet diese Orthogonalität, dass der Fehler $y - \widehat{y}$ unkorreliert mit $x$ ist.

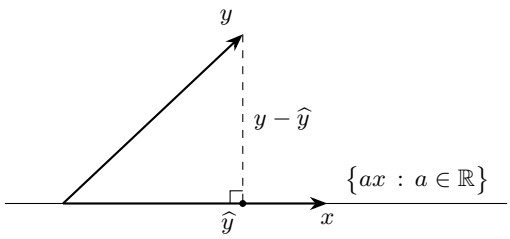


Abbildung 2: Optimale lineare Schätzung als orthogonale Projektion. Die ZV $y$, die beobachtete ZV $x$ und der Schätzer $\widehat{y}$ sind Vektoren des Hilbertraums von Zufallsvariablen mit endlicher Varianz. Die Schätzer, die linear in $x$ sind, bilden den von $x$ aufgespannten Unterraum $\{ax : a \in \mathbb{R}\}$ (die gerade Linie durch ihn), und der estimation error $y - \widehat{y}$ ist am kleinsten für das $\widehat{y}$, dessen Fehler orthogonal zu $x$ ist, der rechte Winkel, der durch das kleine Quadrat bei $\widehat{y}$ markiert ist.

Die optimale nichtlineare Schätzung kann ebenfalls als orthogonale Projektion dargestellt werden, jedoch auf einem größeren Unterraum. Die Projektion von $y$ auf die Zufallsvariablen, die Funktionen von $x$ sind, anstatt nur auf die Vielfachen $ax$, ergibt den conditional expectation $\mathbb{E}\{y \mid x\}$: Er hat den kleinsten quadrierten Fehler unter allen Schätzer, die aus $x$ berechnet werden (Gray and Davisson, 2004, Thm. 4.5).

Das dritte Beispiel ist ein RKHS. Es ist ein Hilbertraum $\mathcal{H}$ von Funktionen $h : \mathcal{X} \to \mathbb{R}$, dessen Skalarprodukt Punktauswertungen reproduziert. Insbesondere ist jeder RKHS mit einem Kernel $k(\cdot, \cdot)$ assoziiert, sodass $k(\mathbf{x}, \cdot) \in \mathcal{H}$ für jedes $\mathbf{x} \in \mathcal{X}$ und

$$h(\mathbf{x}) = \langle h, k(\mathbf{x}, \cdot)\rangle \quad \text{für jedes } h \in \mathcal{H}.$$

Somit ist jeder Vektor $h \in \mathcal{H}$ selbst eine Hypothese-Abbildung $\mathcal{X} \to \mathbb{R}$, die durch die Bildung eines Skalarprodukt mit $k(\mathbf{x}, \cdot)$ ausgewertet wird. Diese Konstruktion einer Hypothese-Abbildung verallgemeinert das lineares Modell, das

das Standard-Skalarprodukt des Euclidean space verwendet. Die Abbildung $\mathbf{x} \mapsto k(\mathbf{x}, \cdot)$ ist eine Merkmalstransformation, die den Merkmalsvektor eines Datenpunkt in $\mathcal{H}$ überführt. Eine Kernel-Methode verwendet einen RKHS in zwei Rollen. Es ist der transformierte Merkmalsraum, in dem ein Datenpunkt durch $k(\mathbf{x}, \cdot)$ repräsentiert wird, und es ist der Hypothesenraum, aus dem eine Hypothese gelernt wird (Hastie et al., 2009, Abschn. 12.3.3; Schölkopf and Smola, 2002). Die beiden Rollen sind auch die von $\mathbb{R}^d$ in der lineare Regression, mit einem Unterschied: Dort parametrisiert ein Vektor eine Hypothese, hier ist ein Vektor selbst eine.

Die drei Beispiele unterscheiden sich darin, was ihre Vektoren sind, nicht in ihrer Geometrie. Jeder Hilbertraum hat eine orthonormale Basis (Rudin, 1987, Thm. 4.22), und zwei Hilberträume, deren orthonormale Basen die gleiche Anzahl von Elementen haben, sind Kopien voneinander: Die Abbildung einer Basis auf die andere erweitert sich zu einer lineare Abbildung, die Skalarprodukte und damit Abstände bewahrt. Eine solche lineare Abbildung ist daher eine isometry (Rudin, 1987, Thms. 4.17 und 4.18, Abschn. 4.19).

Siehe auch: Skalarprodukt, Vektorraum, Norm, Cauchy sequence, Euclidean space, RKHS, orthogonale Projektion, isometry.

## Querverweise

## Literatur


1. Gray and Davisson (2004). An Introduction to Statistical Signal Processing. Cambridge Univ. Press, Cambridge, U.K.

2. Schölkopf and Smola (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

3. Rudin (1987). Real and Complex Analysis. 3rd ed., McGraw-Hill, New York, NY, USA.

4. Rudin (1976). Principles of Mathematical Analysis. 3rd ed., McGraw-Hill, New York, NY, USA.

5. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.