**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**NHẬP MÔN HỌC MÁY**

**SỬ DỤNG PHƯƠNG PHÁP HỒI QUY RIDGE ĐỂ DỰ ĐOÁN THỂ LOẠI PHIM YÊU THÍCH**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | **: THS. NGUYỄN THỊ THANH TÂN** |
| **Sinh viên thực hiện** | **: NGUYỄN THỊ HẠNH AN**  **NGUYỄN THỊ THÙY LINH**  **ĐÀO THÀNH GIA HUY** | |
| **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | |
| **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | |
| **Lớp học phần** | **: D17CNPM3** | |
| **Khóa** | **: 2022-2027** | |

**PHIẾU CHẤM ĐIỂM**

Sinh viên thực hiện:

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Điểm** |
| Nguyễn Thị Thùy Linh |  |  |
| Nguyễn Thị Hạnh An |  |  |
| Đào Thành Gia Huy |  |  |

Giảng viên chấm:

|  |  |  |
| --- | --- | --- |
| **Họ tên** | **Chữ Ký** | **Ghi chú** |
|  |  |  |
|  |  |  |

**LỜI CẢM ƠN**

Trước tiên, chúng em xin được gửi lời cảm ơn chân thành và sâu sắc nhất đến quý thầy cô trong Khoa Công nghệ Thông tin - Trường Đại học Điện lực. Trong suốt quá trình học tập tại trường, các thầy cô đã không ngừng truyền đạt cho chúng em những kiến thức quý báu, không chỉ về mặt lý thuyết mà còn cả những kỹ năng thực tiễn cần thiết cho ngành nghề mà chúng em đang theo đuổi. Sự tận tâm và nhiệt huyết của quý thầy cô đã góp phần thắp sáng trong chúng em niềm đam mê với lĩnh vực Công nghệ Thông tin, đồng thời là hành trang vững chắc để chúng em tự tin bước vào con đường sự nghiệp trong tương lai.

Đặc biệt, chúng em xin bày tỏ lòng biết ơn sâu sắc tới Th.S Nguyễn Thị Thanh Tân, người đã trực tiếp hướng dẫn chúng em trong suốt quá trình thực hiện báo cáo chuyên đề môn Nhập môn học máy. Cô không chỉ tận tình giảng dạy, mà còn luôn sẵn lòng chỉ bảo, đưa ra những ý kiến đóng góp quý báu, giúp chúng em hiểu sâu hơn về vấn đề, hoàn thiện hơn từng nội dung trong báo cáo. Sự hỗ trợ và đồng hành của thầy chính là nguồn động lực to lớn giúp chúng em hoàn thành tốt nhiệm vụ được giao.

Mặc dù chúng em đã nỗ lực hết mình với tinh thần trách nhiệm cao, nhưng do hạn chế về kiến thức và kinh nghiệm thực tiễn, chắc chắn đề tài vẫn không tránh khỏi những thiếu sót nhất định. Chúng em rất mong nhận được những ý kiến đóng góp quý báu từ quý thầy cô và các bạn sinh viên, để chúng em có thể tiếp tục hoàn thiện, nâng cao chất lượng đề tài và hướng tới việc áp dụng hiệu quả vào thực tế.

**MỤC LỤC**

[LỜI MỞ ĐẦU 4](#_Toc197299595)

[CHƯƠNG 1: TỔNG QUAN HỌC MÁY 5](#_Toc197299596)

[1.1 Giới thiệu học máy 5](#_Toc197299597)

[1.2 Ứng dụng của học máy 6](#_Toc197299598)

[1.3 Phân loại học máy 6](#_Toc197299599)

[1.3.1 Học giám sát (Supervised Learning) 7](#_Toc197299600)

[1.3.2 Học không giám sát (Unsupervised Learning) 7](#_Toc197299601)

[1.3.3 Học bán giám sát (Semi-supervised Learning) 8](#_Toc197299602)

[1.3.4 Học tăng cường (Reinforcement Learning) 9](#_Toc197299603)

[CHƯƠNG 2 : PHƯƠNG PHÁP HỒI QUY RIDGE TRONG HỌC MÁY 10](#_Toc197299604)

[2.1 Tính tổng quát của mô hình 10](#_Toc197299605)

[2.2 Bài toán hồi quy tuyến tính 10](#_Toc197299606)

[2.3 Sự thay đổi của hàm mất mát trong hồi qui Ridge 13](#_Toc197299607)

[2.4 Nghiệm tối ưu của hồi qui ridge 14](#_Toc197299608)

[2.5 Sự đảm bảo lời giải của hồi qui ridge 15](#_Toc197299609)

[CHƯƠNG 3: CÀI ĐẶT THỰC NGHIỆM 17](#_Toc197299610)

[3.1 Nhập dữ liệu 17](#_Toc197299611)

[3.1.1 Download bộ dữ liệu 17](#_Toc197299612)

[3.1.2 Khám phá bộ dữ liệu 17](#_Toc197299613)

[3.2 Biến đổi dữ liệu 15](#_Toc197299614)

[3.3 Phân chia dữ liệu học 17](#_Toc197299615)

[3.4 Huấn luyện mô hình 17](#_Toc197299616)

[3.5 Tính tỉ lệ 18](#_Toc197299617)

[3.6 Biểu diễn 19](#_Toc197299618)

[KẾT LUẬN 21](#_Toc197299619)

[DANH MỤC TÀI LIỆU THAM KHẢO 22](#_Toc197299620)

**DANH MỤC HÌNH ẢNH**

[Hình 1. 1: Giới thiệu học máy 6](#_Toc197299791)

[Hình 1. 2:Cách thức học của mô hình Học có giám sát 8](#_Toc197299792)

[Hình 1. 3: Cách thức học của mô hình Học không giám sát 9](#_Toc197299793)

[Hình 3. 1: Biểu đồ biểu diễn điểm đánh giá 19](#_Toc197300049)

[Hình 3. 2: Biểu đồ biểu diễn điểm đánh giá 13](#_Toc197300050)

[Hình 3. 3: Biểu đồ biểu diễn MSE theo các giá trị alpha trong hồi quy Ridge 19](#_Toc197300051)

[Hình 3. 4: Đồ thị biểu diễn hệ số ước lượng theo các giá trị alpha 20](#_Toc197300052)

# 

# LỜI MỞ ĐẦU

Trong thời đại số, khi lượng dữ liệu người dùng và nội dung giải trí ngày càng tăng mạnh, các hệ thống gợi ý đóng vai trò quan trọng trong việc cá nhân hóa trải nghiệm, giúp người dùng tiếp cận nhanh chóng các nội dung phù hợp với sở thích. Một trong những bài toán nền tảng trong hệ thống gợi ý là dự đoán thể loại phim yêu thích của người dùng, từ đó cung cấp những đề xuất chính xác và mang tính cá nhân cao.

Học máy cung cấp các phương pháp hiệu quả để khai thác dữ liệu người dùng và thuộc tính nội dung, từ đó xây dựng mô hình phân loại chính xác. Trong số các kỹ thuật học có giám sát, hồi quy Ridge là một biến thể của hồi quy tuyến tính được điều chuẩn nhằm hạn chế hiện tượng quá khớp (overfitting), đặc biệt hiệu quả trong bối cảnh dữ liệu có đa cộng tuyến hoặc số lượng đặc trưng lớn.

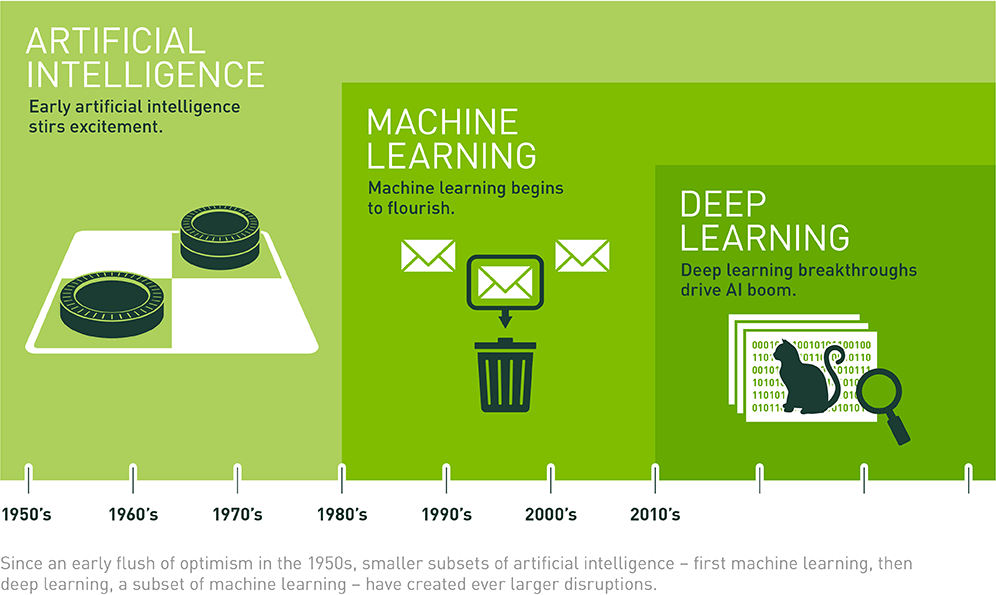
Việc áp dụng hồi quy Ridge trong bài toán phân loại thể loại phim yêu thích không chỉ giúp cải thiện độ chính xác dự đoán mà còn thể hiện rõ tiềm năng ứng dụng học máy trong ngành công nghiệp giải trí số, nơi cá nhân hóa trải nghiệm người dùng ngày càng trở thành yêu cầu thiết yếu.

# CHƯƠNG 1: TỔNG QUAN HỌC MÁY

## 1.1 Giới thiệu học máy

Machine learning là một lĩnh vực thuộc trí tuệ nhân tạo (AI), xoay quanh các thuật toán nhằm xác định mẫu hình và mối liên hệ trong dữ liệu. “Máy” ở đây đồng nghĩa với chương trình máy tính, và “học” miêu tả cách mà các thuật toán học máy trở nên chính xác hơn sau một quá trình tiếp nhận dữ liệu bổ sung.

Machine learning cho phép máy tính giải quyết các tác vụ mà trước đây chỉ con người mới thực hiện được – từ phiên dịch cho đến lái xe, học máy đã và đang tạo tiền đề cho sự bùng nổ của trí tuệ nhân tạo, giúp phần mềm nhận thức được thế giới thực vốn rất phức tạp và không thể dự đoán được.

Khái niệm học máy không hề mới mẻ, nhưng nó chỉ thực sự được ứng dụng trong doanh nghiệp khi internet trở nên phổ biến cùng những thành tựu gần đây trong lĩnh vực phân tích dữ liệu lớn và điện toán đám mây. Nguyên nhân là bởi việc huấn luyện thuật toán học máy tìm kiếm mẫu hình trong dữ liệu đòi hỏi rất nhiều tài nguyên điện toán và phải tiếp cận được dữ liệu lớn.

Hình 1. 1: Giới thiệu học máy

## 1.2 Ứng dụng của học máy

Có rất nhiều ứng dụng thực tế khác nhau của học máy. Hai lĩnh vực ứng dụng lớn nhất của học máy là khai phá dữ liệu (data mining) và nhận dạng mẫu (pattern recognition):

- Khai phá dữ liệu là ứng dụng kỹ thuật học máy vào các cơ sở dữ liệu hoặc các tập dữ liệu lớn để phát hiện quy luật hay tri thức trong dữ liệu đó hoặc để dự đoán các thông tin quan tâm trong tương lai. Ví dụ, từ tập hợp hóa đơn bán hàng có thể phát hiện ra quy luật “những người mua bánh mì thường mua bơ”.

- Nhận dạng mẫu là ứng dụng các kỹ thuật học máy để phát hiện các mẫu có tính quy luật trong dữ liệu, thường là dữ liệu hình ảnh, âm thanh. Bài toán nhận dạng mẫu cụ thể thường là xác định nhãn cho đầu vào cụ thể, ví dụ cho ảnh chụp mặt người, cần xác định đó là ai.

Cần lưu ý, khai phá dữ liệu và nhận dạng mẫu có nhiều điểm trùng nhau cả trong phạm vi nghiên cứu và ứng dụng. Điểm khác nhau chủ yếu liên quan tới lĩnh vực ứng dụng và kỹ thuật sử dụng, theo đó khai phá dữ liệu liên quan tới dữ liệu thương mại trong khi nhận dạng mẫu liên quan nhiều tới dữ liệu âm thanh, hình ảnh và được dùng nhiều trong kỹ thuật. Sau đây là một số ví dụ ứng dụng cụ thể của học máy:

- Nhận dạng ký tự: phân loại hình chụp ký tự thành các loại, mỗi loại ứng với một ký tự tương ứng.

- Phát hiện và nhận dạng mặt người: phát hiện vùng có chứa mặt người trong ảnh, xác định đó là mặt người nào trong số những người đã có ảnh trước đó, tức là phân chia ảnh thành những loại tương ứng với những người khác nhau.

- Lọc thư rác, phân loại văn bản: dựa trên nội dung thư điện tử, chia thư thành loại “thư rác” hay “thư bình thường”; hoặc phân chia tin tức thành các thể loại khác nhau như “xã hội”, “kinh tế”.

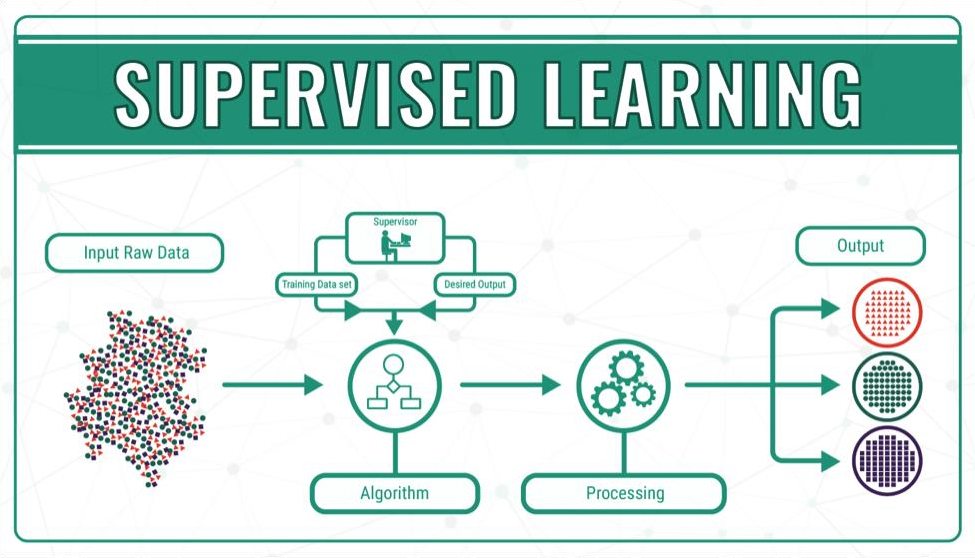
- Dịch tự động : dựa trên dữ liệu huấn luyện dưới dạng các văn bản song ngữ, hệ thống dịch tự động học cách dịch từ ngôn ngữ này sang ngôn ngữ khác.Hệ thống dịch tự động tiêu biểu dạng này là Google Translate.

- Chẩn đoán y tế: học cách dự đoán người bệnh có mắc hay không mắc một số bệnh nào đó dựa trên triệu chứng quan sát được.

## 1.3 Phân loại học máy

### 1.3.1 Học giám sát (Supervised Learning)

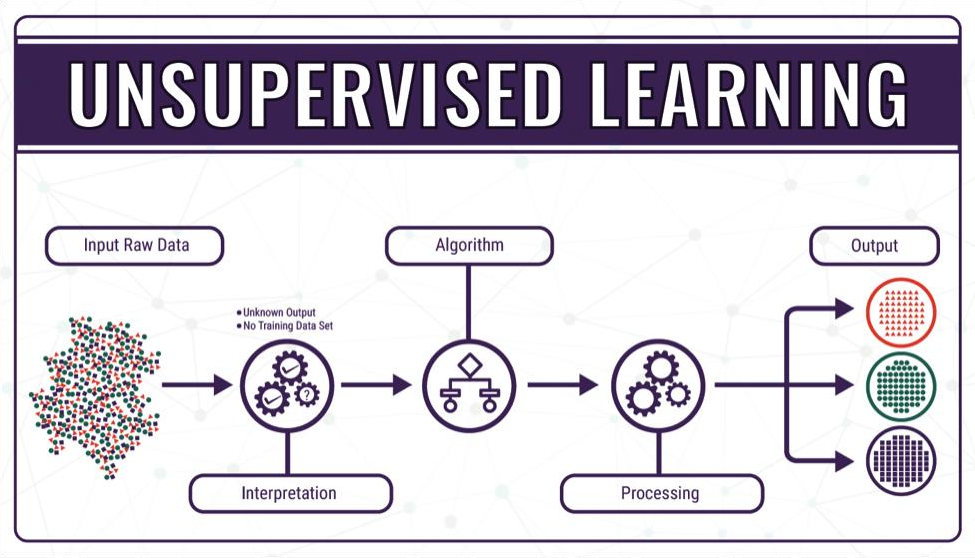
Supervised Learning (Học có giám sát) là một nhóm thuật toán sử dụng dữ liệu được gán nhãn nhằm mô hình hóa mối quan hệ giữa biến đầu vào (x) và biến đầu ra (y). Hai nhóm bài toán cơ bản trong học có giám sát là classification (phân loại) và regression (hồi quy), trong đó biến đầu ra của bài toán phân loại có các giá trị rời rạc trong khi biến đầu ra của bài toán hồi quy có các giá trị liên tục. Với Supervised Learning, bên cạnh xây dựng các mô hình mạnh, việc thu thập và gán nhãn dữ liệu tốt và hợp lý cũng đóng vai trò then chốt để giải quyết các bài toán trong thực tế.

Hình 1. 2:Cách thức học của mô hình Học có giám sát

### 1.3.2 Học không giám sát (Unsupervised Learning)

Ngược lại, Unsupervised Learning (Học không giám sát) là một nhóm thuật toán sử dụng dữ liệu không có nhãn. Các thuật toán theo cách tiếp cận này hướng đến việc mô hình hóa được cấu trúc hay thông tin ẩn trong dữ liệu. Hay nói cách khác, sử dụng các phương pháp này thiên về việc mô tả tính chất hay đặc tính của dữ liệu. Thông thường, các thuật toán này dựa trên những thông tin sau:

* + Mối quan hệ tương tự (similarity) giữa các ví dụ (được gọi là instance) trong dữ liệu như trong các thuật toán clustering (phân cụm)
  + Xác suất đồng xuất hiện của các đối tượng như trong Association mining
  + Các phép biến đổi ma trận để trích xuất các đặc trưng như PCA, SVD.

Hình 1. 3: Cách thức học của mô hình Học không giám sát

### 1.3.3 Học bán giám sát (Semi-supervised Learning)

Kỹ thuật này dựa vào một lượng nhỏ dữ liệu gắn nhãn và một lượng lớn dữ liệu chưa gắn nhãn để huấn luyện hệ thống. Dữ liệu gắn nhãn được sử dụng để huấn luyện cơ bản một mô hình học máy, sau đó mô hình đã qua huấn luyện cơ bản này sẽ được sử dụng để gắn nhãn dữ liệu chưa gắn nhãn – một quy trình gọi là “gắn nhãn giả”. Mô hình sau đó sẽ được huấn luyện bằng tổng hợp dữ liệu gắn nhãn và gắn nhãn giả.

Hiệu suất học nửa giám sát thời gian qua đã được cải thiện khá nhiều nhờ các Mạng đối nghịch tạo sinh (GAN), tức các hệ thống học máy có thể sử dụng dữ liệu gắn nhãn để tạo ra dữ liệu hoàn toàn mới, từ đó lại sử dụng dữ liệu mới để giúp huấn luyện một mô hình học máy.

### 1.3.4 Học tăng cường (Reinforcement Learning)

Học tăng cường (RL) là kỹ thuật máy học (ML) giúp đào tạo phần mềm đưa ra quyết định nhằm thu về kết quả tối ưu nhất. Kỹ thuật này bắt chước quy trình học thử và sai mà con người sử dụng để đạt được mục tiêu đã đặt ra. RL giúp phần mềm tăng cường các hành động hướng tới mục tiêu, đồng thời bỏ qua các hành động làm xao lãng mục tiêu.

Thuật toán RL sử dụng mô hình khen thưởng và trừng phạt trong quy trình xử lý dữ liệu. Các thuật toán này tiếp thu ý kiến phản hồi của từng hành động và tự khám phá ra con đường xử lý tốt nhất để thu về kết quả cuối cùng. Thuật toán RL còn có khả năng trì hoãn khen thưởng. Chiến lược tổng thể tốt nhất có thể đòi hỏi phải đánh đổi một vài lợi ích trước mắt, vì vậy cách tiếp cận tốt nhất mà RL khám phá ra có thể bao gồm một số trừng phạt hoặc giai đoạn quay lui. RL là phương pháp hiệu suất cao giúp hệ thống trí tuệ nhân tạo (AI) đạt kết quả tối ưu trong môi trường chưa biết.

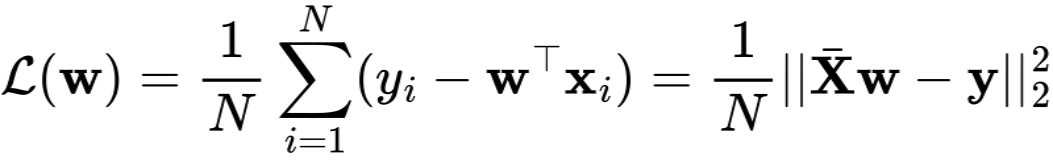
# CHƯƠNG 2 : PHƯƠNG PHÁP HỒI QUY RIDGE TRONG HỌC MÁY

## 2.1 Tính tổng quát của mô hình

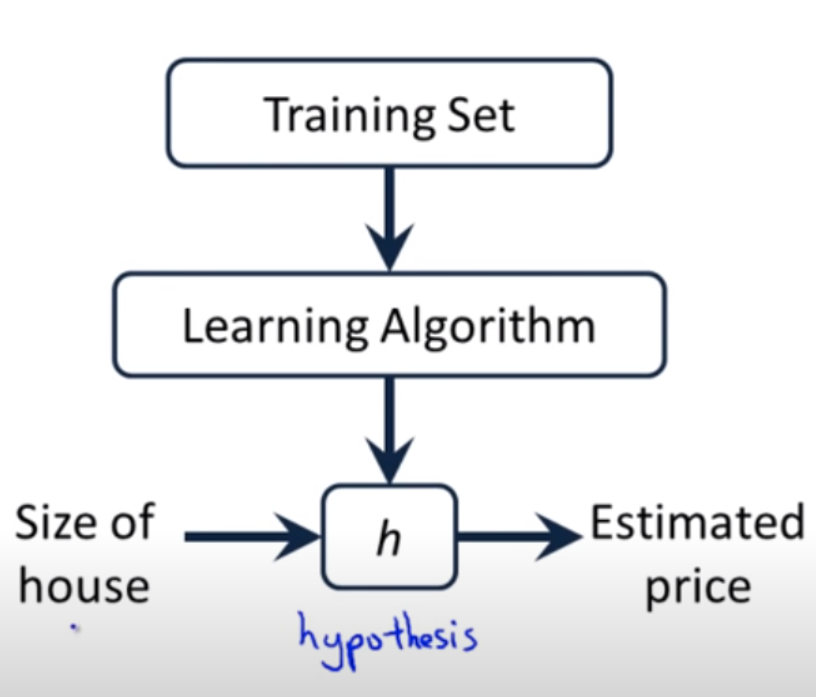
Một mục tiêu tiên quyết để có thể áp dụng được mô hình vào thực tiến đó là chúng ta cần giảm thiểu hiện tượng *quá khớp*. Để thực hiện được mục tiêu đó, mô hình được huấn luyện được kì vọng sẽ nắm bắt được qui luật tổng quát từ *tập huấn luyện* (*train dataset*) mà qui luật đó phải đúng trên những dữ liệu mới mà nó chưa được học. Thông thường tập dữ liệu mới đó được gọi là *tập kiểm tra* (*test dataset*). Đây là một tập dữ liệu độc lập được sử dụng để đánh giá mô hình.

## 2.2 Bài toán hồi quy tuyến tính

Giả định dữ liệu đầu vào bao gồmwpsquan sát là những cặp các biến đầu vào và biến mục tiêu wps, wps. Quá trình hồi qui mô hình sẽ tìm kiếm một véc tơ hệ số ước lượng wpssao cho tối thiểu hoá hàm mất mát dạng MSE:



Nhắc lại một chút về khái niệm hàm mất mát. Trong các mô hình học có giám sát của machine learning, từ dữ liệu đầu vào, thông qua phương pháp học tập (*learning algorithm*), chúng ta sẽ đặt ra một hàm giả thuyết *h* (*hypothesis function*) mô tả mối quan hệ dữ liệu giữa biến đầu vào và biến mục tiêu.



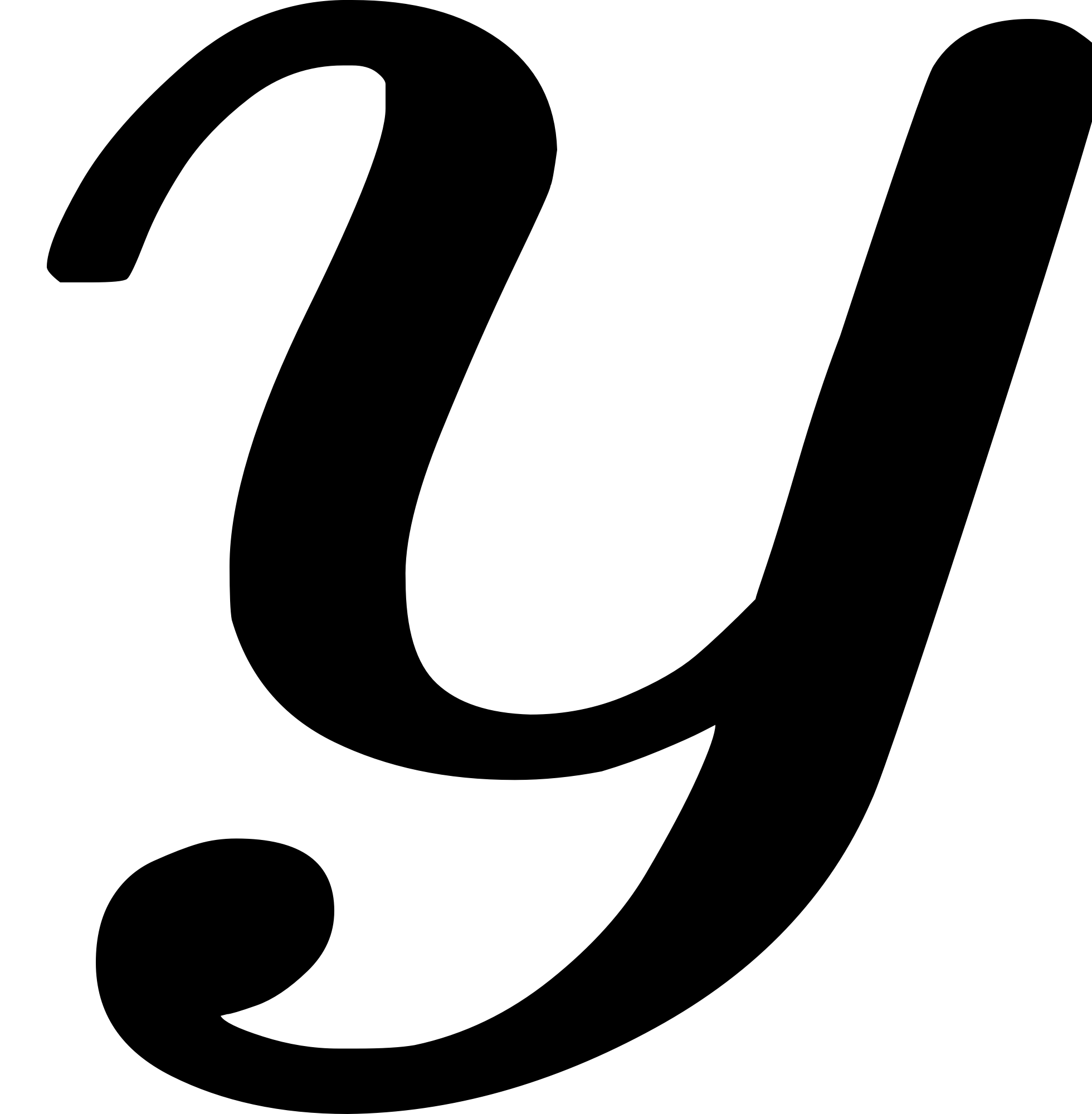
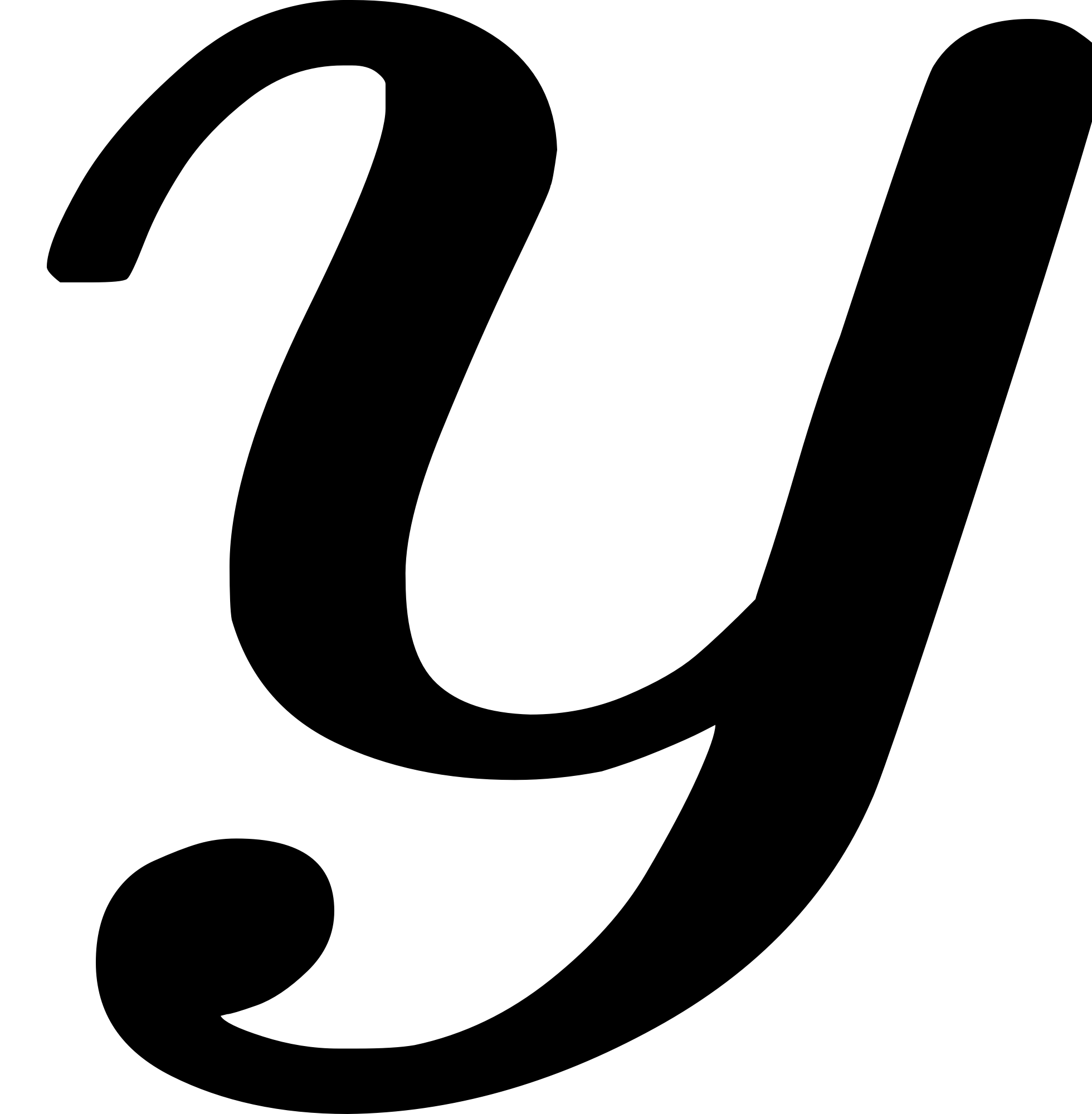
Từ một quan sát đầu vào wps, sau khi đưa vào hàm gỉa thuyết ℎ chúng ta thu

^

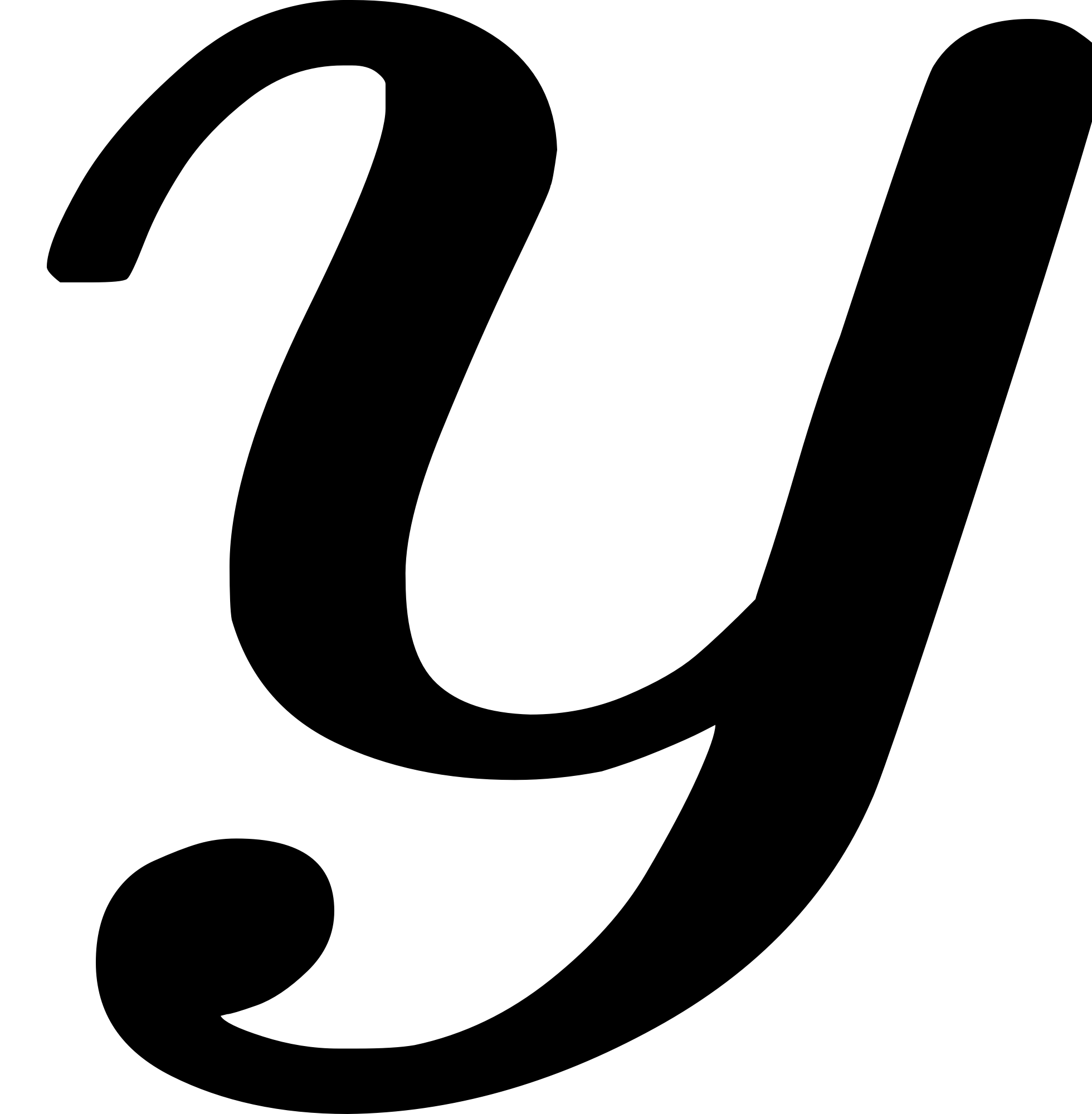
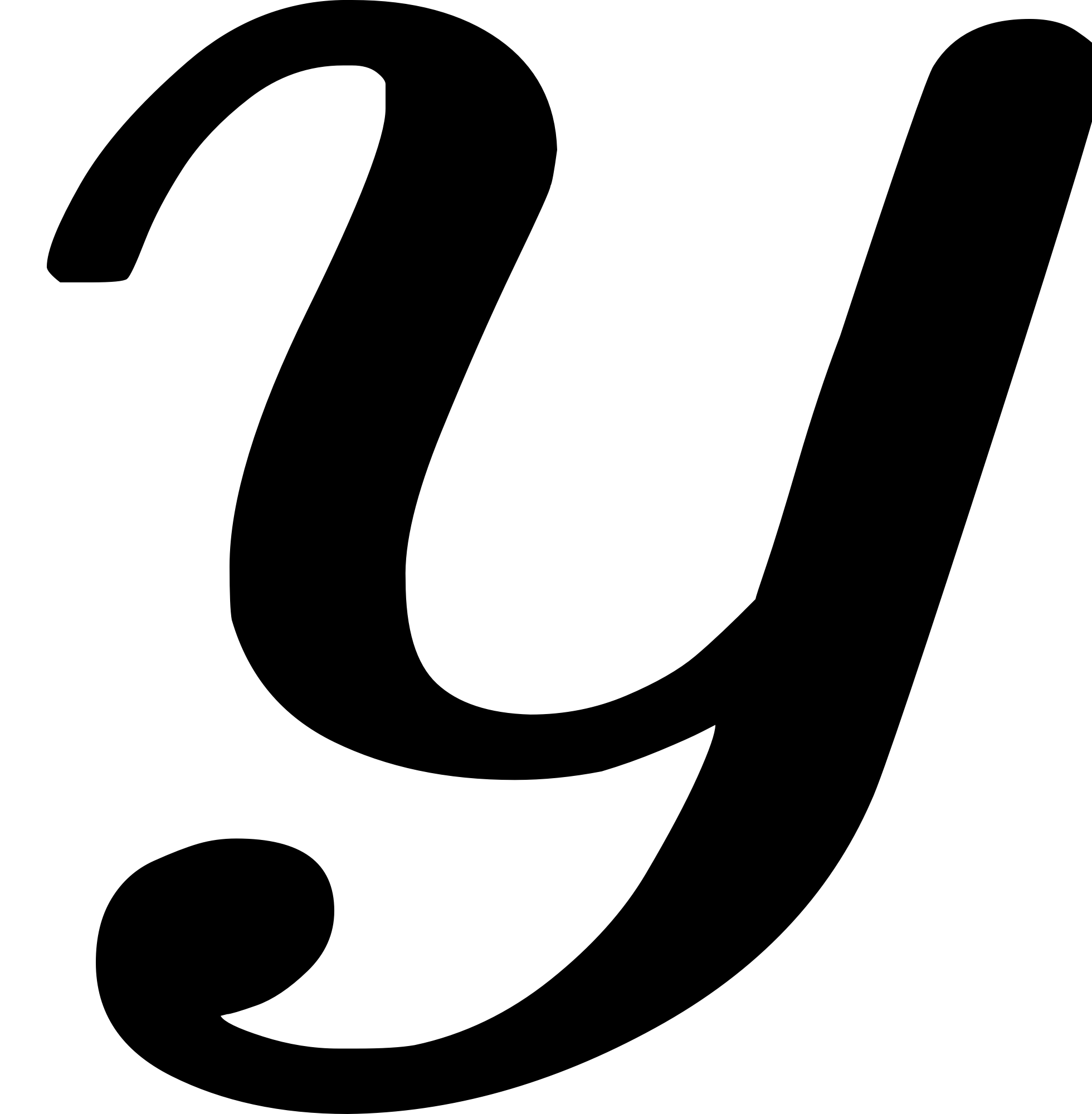
được giá trị dự báo wps ở đầu ra. Chữ wps của tên hàm thể hiện cho từ *hypothesis* có

nghĩa là *giả thuyết*, đây là một khái niệm đã tồn tại lâu năm trong thống kê. Để

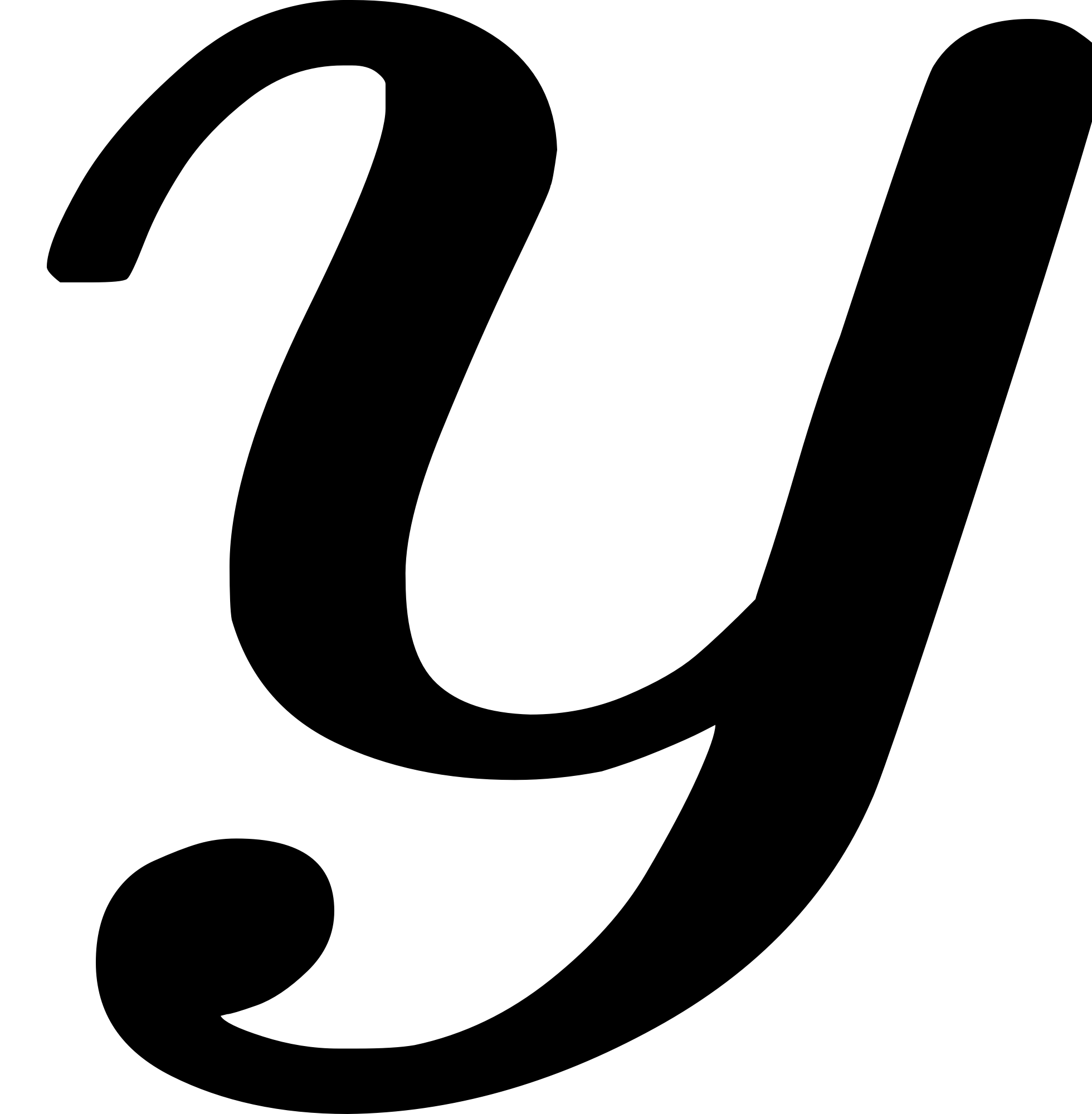
^

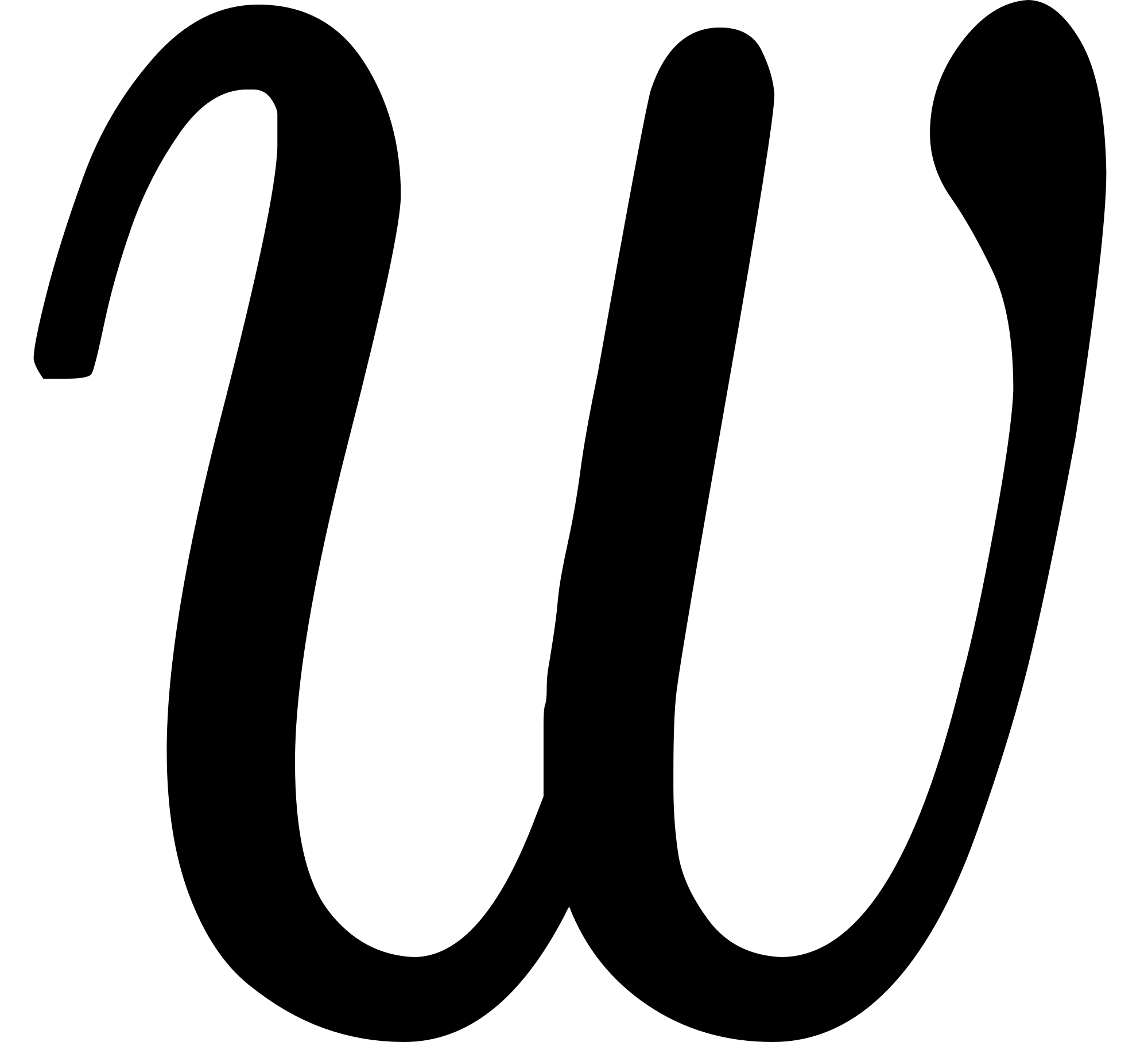
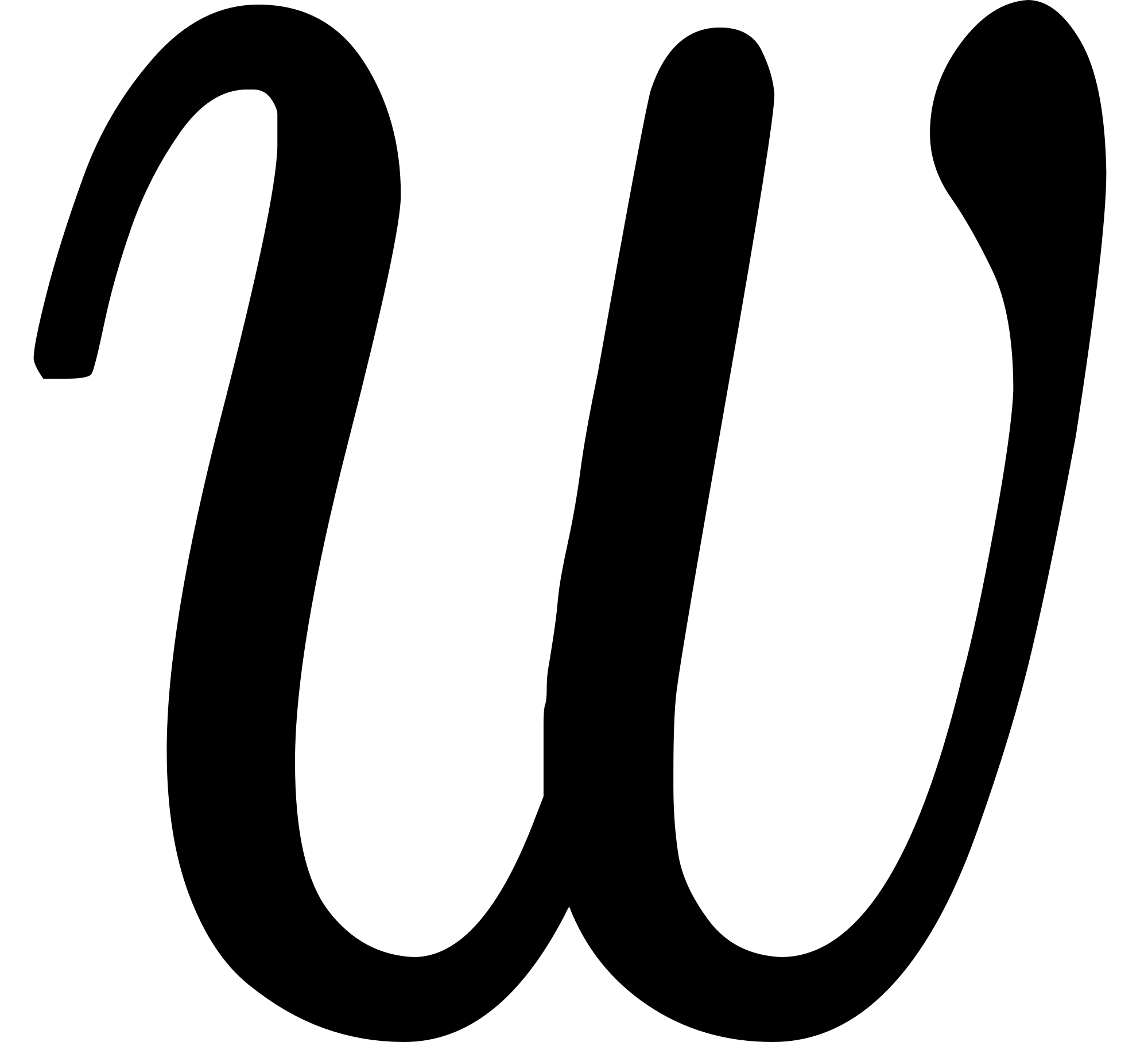
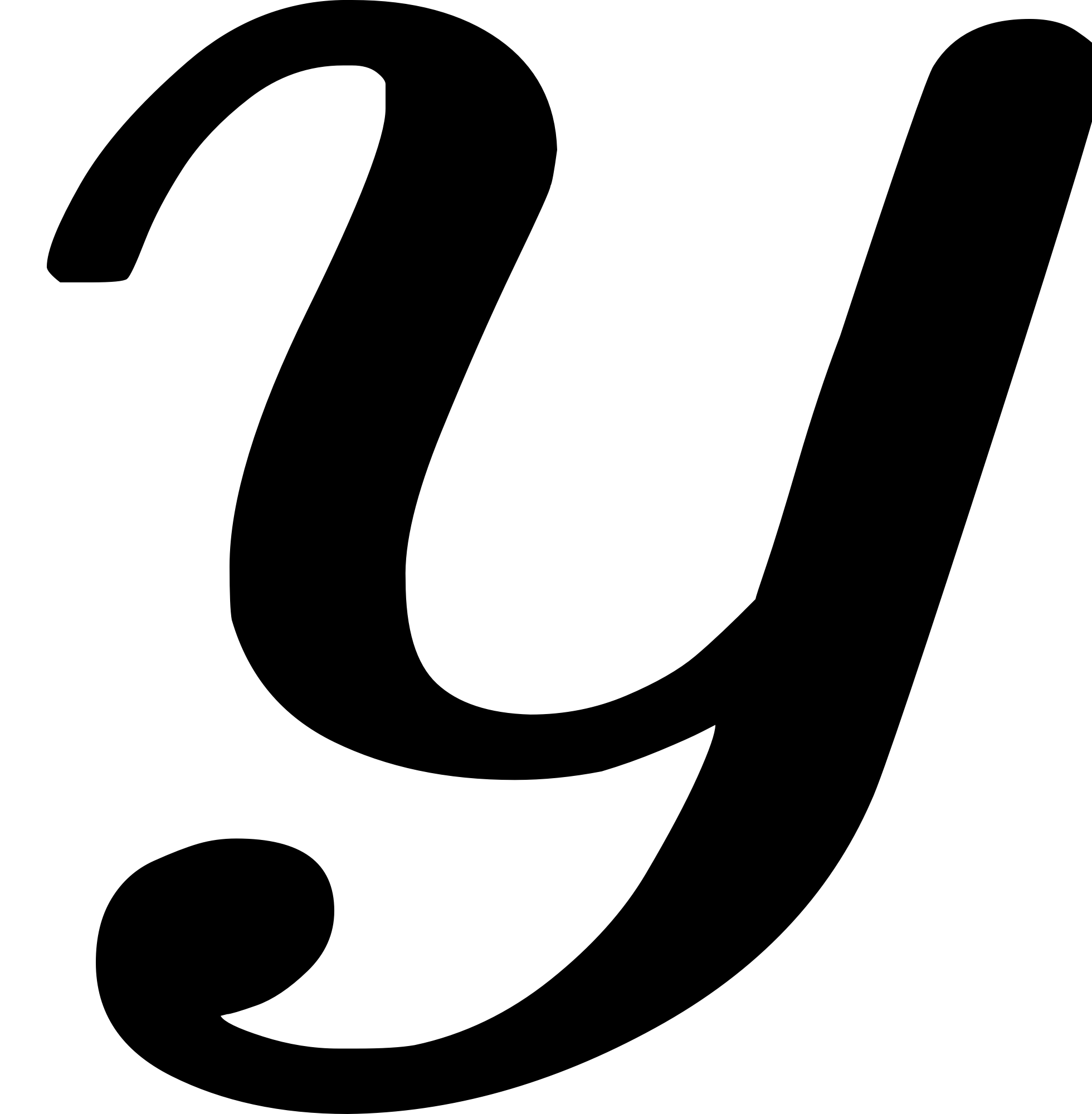
mô hình càng chuẩn xác thì sai số giữa giá trị dự báo  và ground truth  càng

^

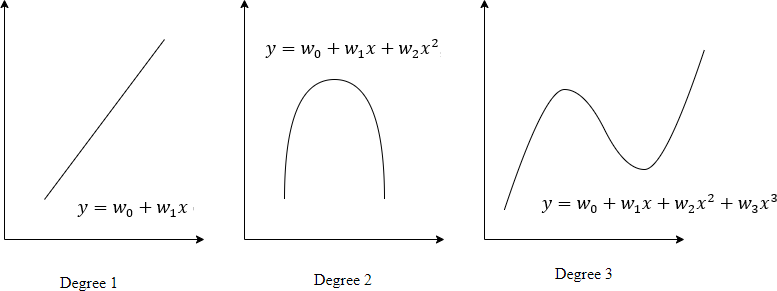
phải nhỏ. Vậy làm thế nào để đo lường được mức độ nhỏ của sai số giữa  và ?

Các thuật toán học có giám sát trong machine learning sẽ sử dụng hàm mất mát để lượng hoá sai số này.

Hàm mất mát cũng chính là mục tiêu tối ưu khi huấn luyện mô hình. Dữ liệu đầu vào wps và  được xem như là cố định và biến số của bài toán tối ưu chính là các giá trị trong véc tơ wps.

Giá trị hàm mất mát *MSE* chính là trung bình của tổng bình phương phần dư. Phần dư chính là chênh lệch giữa giá trị thực tế và giá trị dự báo. Tối thiểu hoá hàm mất mát nhằm mục đích làm cho giá trị dự báo ít chênh lệch so với giá trị thực tế, giá trị thực tế còn được gọi là ground truth. Trước khi huấn luyện mô hình chúng ta chưa thực sự biết véc tơ hệ số  là gì. Chúng ta chỉ có thể đặt ra một giả thuyết về dạng hàm dự báo (trong trường hợp này là phương trình dạng tuyến tính) và các hệ số hồi qui tương ứng. Chính vì vậy mục đích của tối thiểu hoá hàm mất mát là để tìm ra tham số  phù hợp nhất mô tả một cách khái quát quan hệ dữ liệu giữa biến đầu vào wps với biến mục tiêu  trên tập huấn luyện.

Tuy nhiên mối quan hệ này nhiều khi không mô tả được qui luật khái quát của dữ liệu nên dẫn tới hiện tượng *quá khớp*. Một trong những nguyên nhân dẫn tới sự không khái quát của mô hình đó là do mô hình quá phức tạp. Mức độ phức tạp càng cao khi độ lớn của các hệ số trong mô hình hồi qui ở những bậc cao có xu hướng lớn như phân tích trong hình bên dưới:



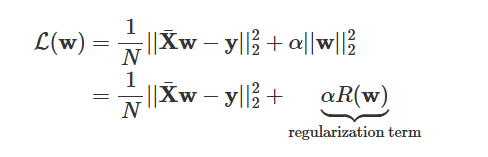
Hình trên thể hiện mức độ phức tạp của mô hình theo sự thay đổi của bậc. Phương trình có độ phức tạp lớn nhất là phương trình bậc 3: wps. Trong chương trình THPT chúng ta biết rằng phương trình bậc 3 thông thường sẽ có 2 điểm uốn và độ phức tạp lớn hơn bậc hai chỉ có 1 điểm uốn. Khi wps thì phương trình bậc 3 hội tụ về phương trình bậc 2: C:/Users/linhy/AppData/Local/Temp/wps.lXxvwxwps, lúc này phương trình là một đường cong dạng parbol và có độ phức tạp

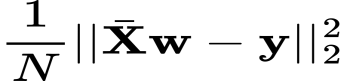
giảm. Tiếp tục kiểm soát độ lớn để C:/Users/linhy/AppData/Local/Temp/wps.RJpgXpwps trong phương trình bậc 2 ta sẽ thu được một đường thẳng tuyến tính dạng C:/Users/linhy/AppData/Local/Temp/wps.uRMpwmwps có độ phức tạp thấp nhất.

Như vậy kiểm soát độ lớn của hệ số ước lượng, đặc biệt là với bậc cao, sẽ giúp giảm bớt mức độ phức tạp của mô hình và thông qua đó khắc phục hiện tượng *quá khớp*. Vậy làm cách nào để kiểm soát chúng, cùng tìm hiểu chương bên dưới.

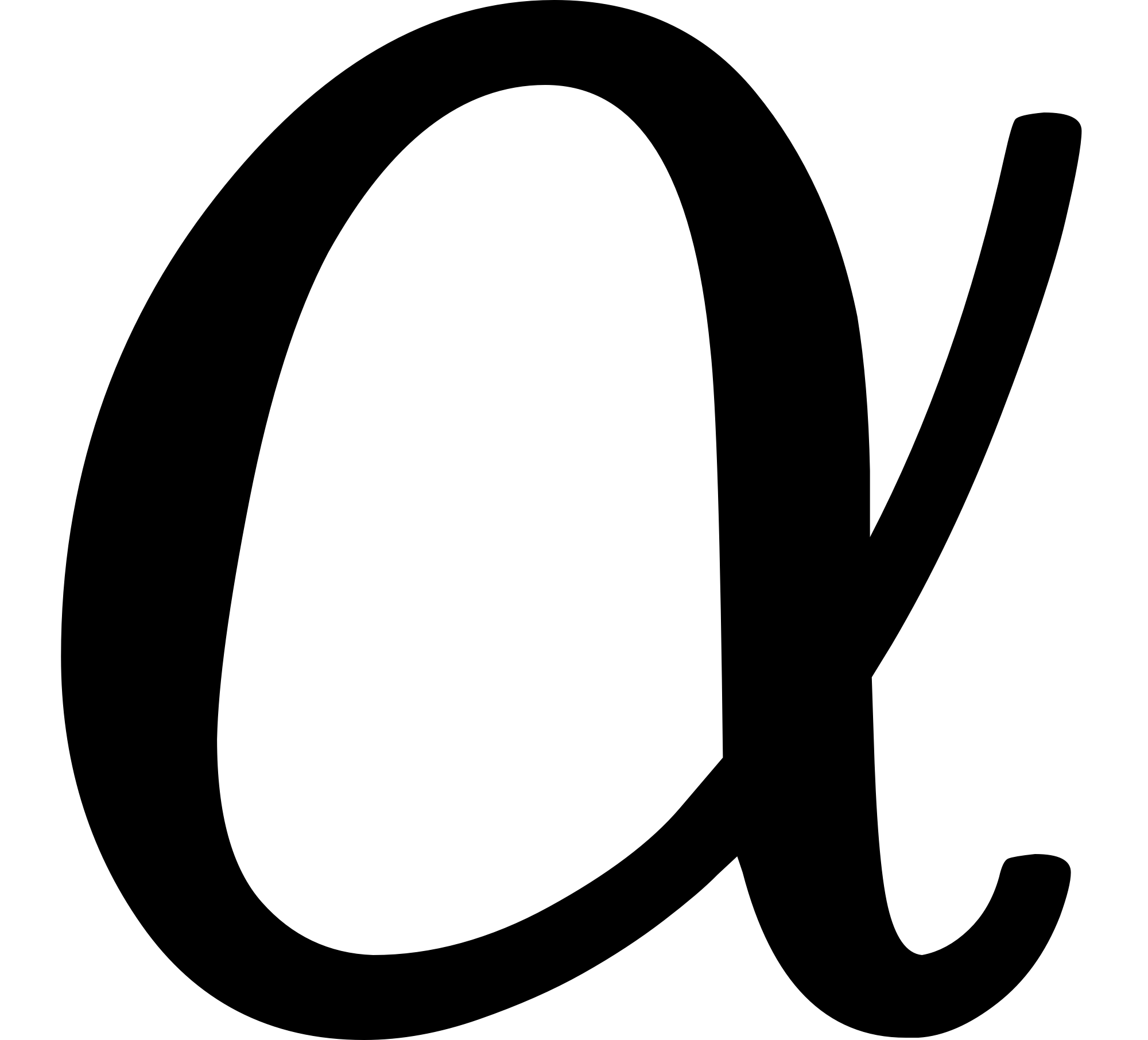
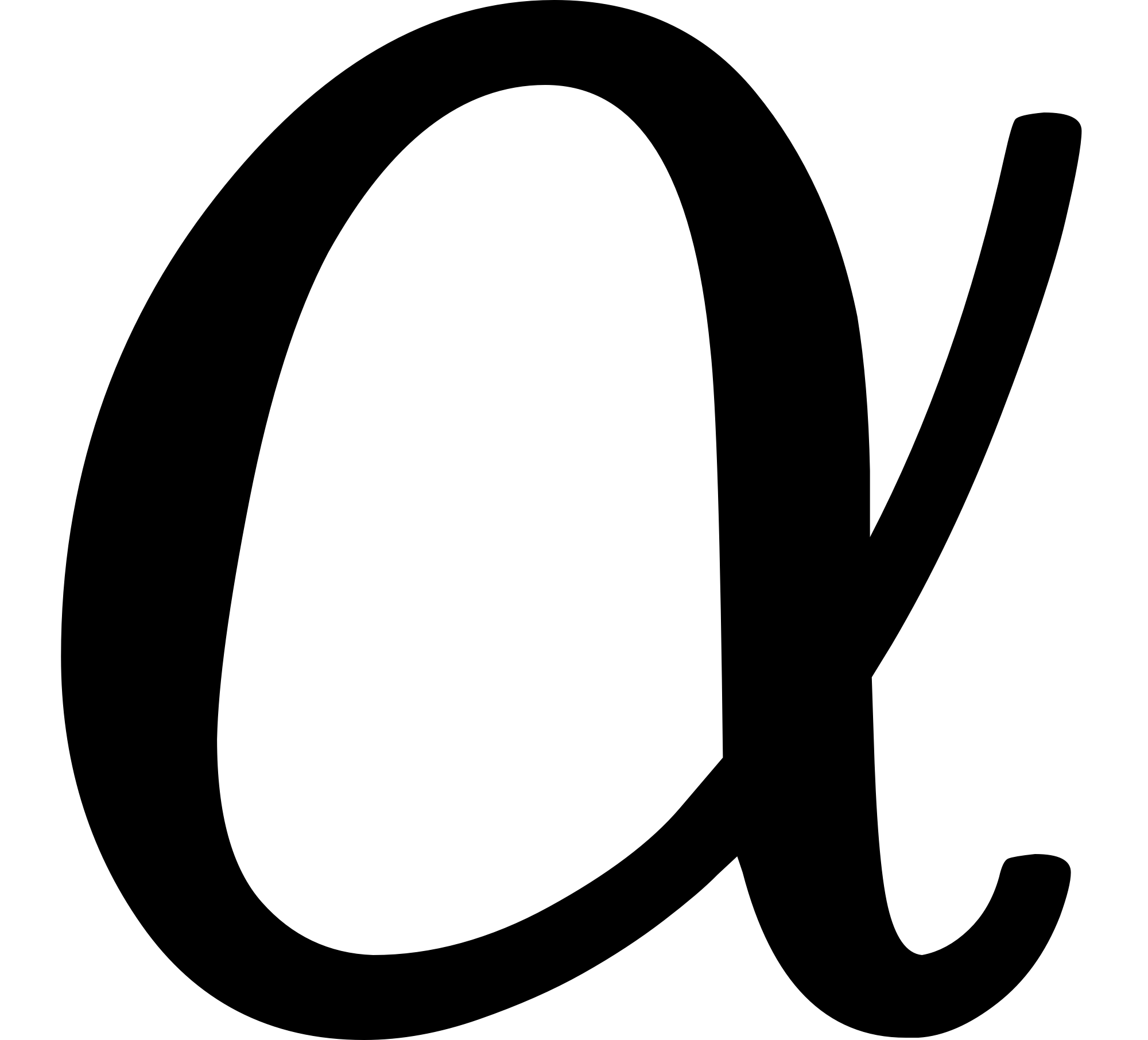
## 2.3 Sự thay đổi của hàm mất mát trong hồi qui Ridge

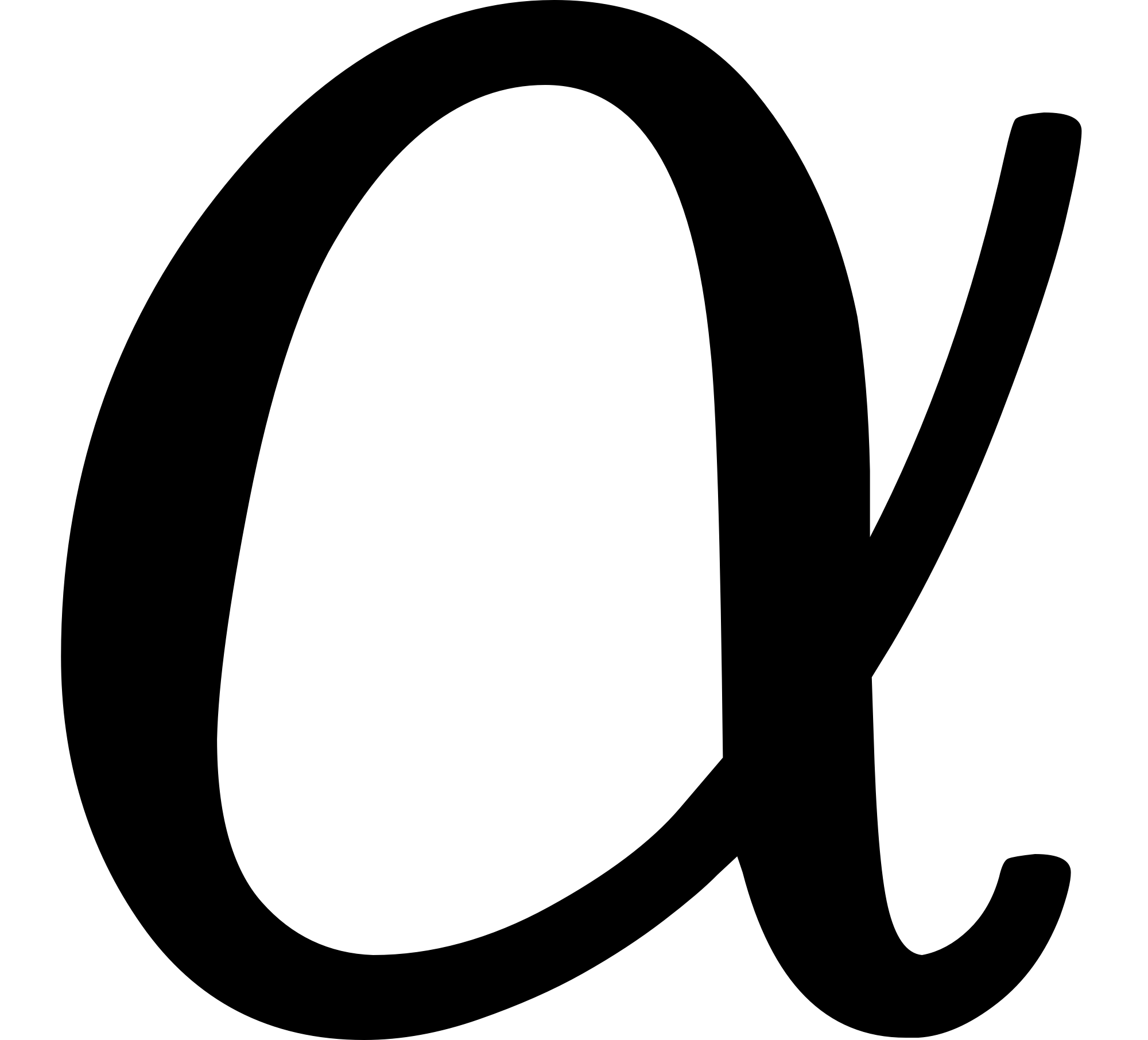
Hàm mất mát trong hồi qui Ridge sẽ có sự thay đổi so với hồi qui tuyến tính đó là *thành phần điều chuẩn* (*regularization term*) được cộng thêm vào hàm mất mát như sau:



Trong phương trình trên thì wps. chính là tổng bình phương phần dư và wps đại diện cho thành phần điều chuẩn.

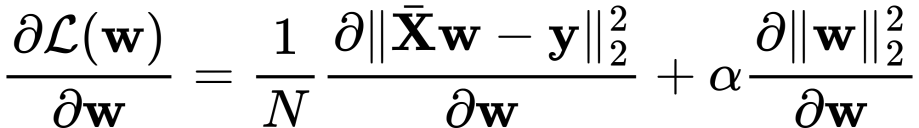
Bài toán tối ưu hàm mất mát của hồi qui *Ridge* về bản chất là tối ưu song song hai thành phần bao gồm tổng bình phương phần dư và *thành phần điều chuẩn*. Hệ số wps có tác dụng điều chỉnh độ lớn của *thành phần điều chuẩn* tác động lên hàm mất mát.

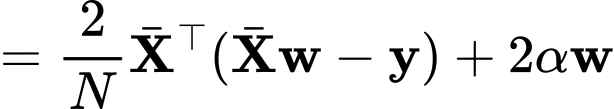
* Trường hợp wps, *thành phần điều chuẩn* bị tiêu giảm và chúng ta quay trở về bài toán hồi qui tuyến tính.
* Trường hợp  nhỏ thì vai trò của *thành phần điều chuẩn* trở nên ít quan trọng. Mức độ kiểm soát *quá khớp* của mô hình sẽ trở nên kém hơn.
* Trường hợp  lớn chúng ta muốn gia tăng mức độ kiểm soát lên độ lớn của các hệ số ước lượng và qua đó giảm bớt hiện tượng *quá khớp*.

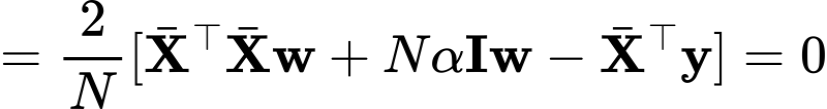
Khi tăng dần hệ số  thì *hồi qui Ridge* sẽ có xu hướng thu hẹp hệ số ước lượng từ mô hình. Chúng ta sẽ thấy rõ thông qua ví dụ mẫu bên dưới.

## 2.4 Nghiệm tối ưu của hồi qui ridge

Giải bài toán tối ưu hàm mục tiêu của hồi qui Ridge theo đạo hàm bậc nhất của véc tơ wps:





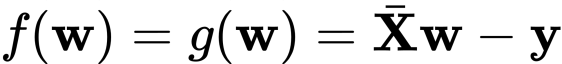


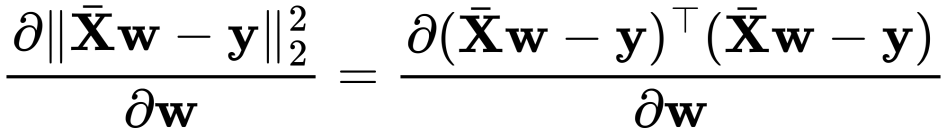
Thật vậy, từ dòng 1 suy ra dòng 2 là vì theo công thức product-rule trong matrix caculus thì:

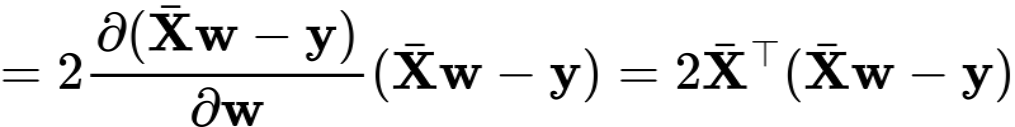
wps

Khi wps thì đạo hàm trở thành:

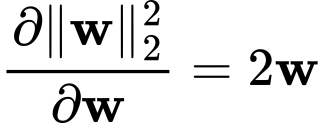
wps

Nếu thay  ta suy ra:





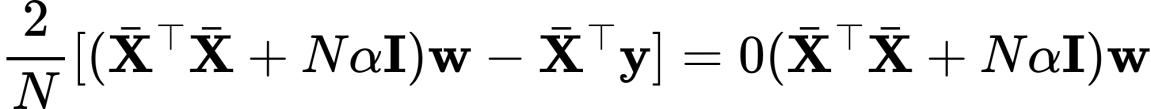
Tương tự ta cũng có:

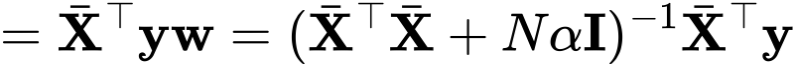


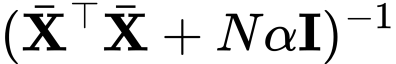
Như vậy ta nhận thấy dòng 1 suy ra dòng 2 là hoàn toàn đúng.

Ở dòng thứ 3 chúng ta áp dụng thêm một tính chất wps trong đó wps là ma trận đơn vị.

Sau cùng nghiệm của đạo hàm bậc nhất trở thành:





Thành phần wps được thêm vào trong  đóng vai trò như một thành phần kiểm soát để giá trị của wps nhỏ hơn so với ban đầu. Trên thực tế thành phần này chỉ tác động lên những phần tử thuộc đường chéo chính của ma trận và làm cho độ lớn của nghiệm giảm.

Ngoài ra ta còn chứng minh được rằng ma trận wps là một ma trận không suy biến nếu wps. Điều đó đảm bảo rằng mô hình hồi qui Ridge luôn tìm được nghiệm. Bạn đọc quan tâm tới toán có thể thấy chứng minh này ở mục bên dưới.

## 2.5 Sự đảm bảo lời giải của hồi qui ridge

Trước tiên hãy cùng ôn lại một số khái niệm liên quan tới ma trận.

Định nghĩa bán xác định dương: Ma trận số thực đối xứng wps là bán xác định dương (positive semi-definite) nếu với mọi véc tơ wpsthì wps

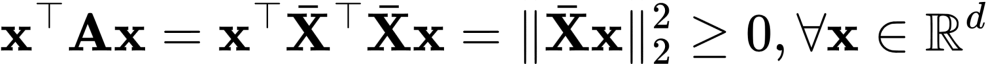
Một tính chất thú vị đó là nếu một ma trận bán xác định dương thì mọi trị riêng của chúng là những số không âm. Thật vậy, theo định nghĩa thì wpslà trị riêng (eigen-value) của ma trận wps tương ứng với một véc tơ riêng (eigen-vector) wps nếu thỏa mãn:

wps

Mặc khác vế trái không âm do \mathbf{A} là ma trận bán xác định dương. Do đó vế phải wps, từ đó suy ra wps

Để chứng minh hồi qui Ridge luôn tồn tại nghiệm chúng ta dựa vào ba tính chất lý:

1.- Ma trận wps là một ma trận thực đối xứng bán xác định dương (positive semi-definite). Thật vậy:



Từ đó suy ra wps là ma trận bán xác định dương. Như vậy các trị riêng (eigenvalues) của nó là wps không âm.

2.Nếu \mu là trị riêng của ma trận wpsvuông thì wps là trị riêng của ma trận wps

Để chứng minh ta dựa vào khai triển:

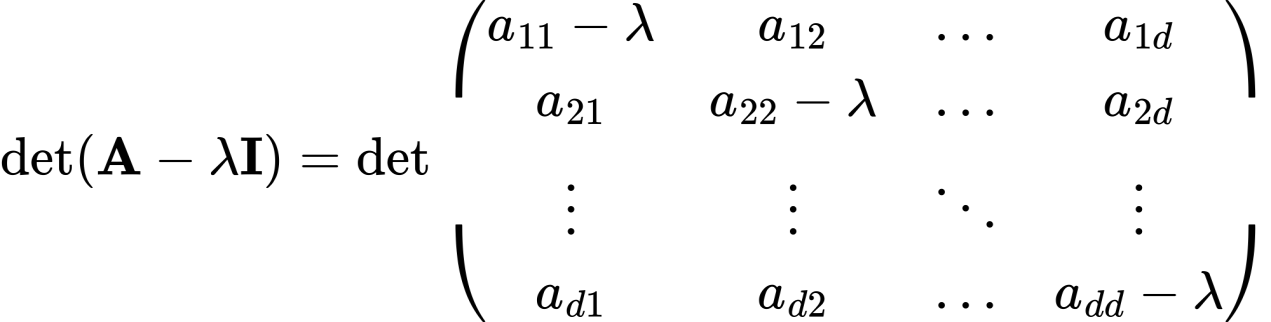
C:/Users/linhy/AppData/Local/Temp/wps.FoHTBJwps

wps

Dòng cuối cùng suy ra wpschính là trị riêng của ma trận wps

3.- Định thức của ma trận wpsbằng tích các trị riêng của wps.

Giả sử wps là các trị riêng của ma trận wps . Khi đó định thức:



wps

là một đa thức bậc d của wps

Mặc khác với mỗi trị riêng wps của ma trận wps thì tồn tại véc tơ riêng wps khác 0 thỏa mãn:

wps

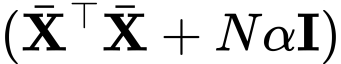
Như vậy các dòng của ma trận wps phụ thuộc tuyến tính theo véc tơ wps nên wpsTừ đó suy ra wpscó d nghiệm là các trị riêng của ma trận wps .Kết hợp với hệ số của bậc cao nhất ta suy ra:

C:/Users/linhy/AppData/Local/Temp/wps.UsTVtEwps

wps

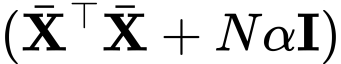
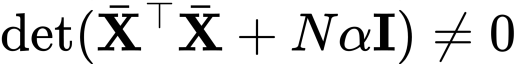
Do đó:

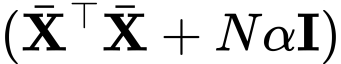
wps

Quay trở lại bài toán chứng minhlà một ma trận không suy biến.

Giả định wps là véc tơ trị riêng của ma trận wps. Như vậy từ tính chất 2 suy ra trị riêng của ma trận wps là wps.

Mặt khác theo tính chất 1 thì wps do wps bán xác định dương. Từ đó suy ra wps.

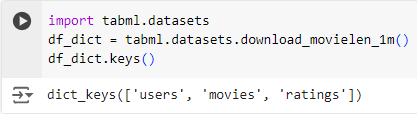
Như vậy ma trận có các trị riêng khác 0. Theo tính chất 3 ta suy ra do các trị riêng đều khác 0.

Như vậy là một ma trận không suy biến và hồi qui Ridge đảm bảo tồn tại nghiệm.

# CHƯƠNG 3: CÀI ĐẶT THỰC NGHIỆM

## 3.1 Nhập dữ liệu

### 3.1.1 Download bộ dữ liệu

****

Có ba dataframe trong bộ dữ liệu này là users, movies và ratings lần lượt chứa thông tin của người dùng, bộ phim và các đánh giá.

### 3.1.2 Khám phá bộ dữ liệu

* + - * Rating

Dưới đây là 10 dòng đầu tiên của dataframe rating. Dữ liệu rating bao gồm thông tin về mã người dùng UserID, mã phim MovieID, đánh giá trong thang điểm 5 và thời điểm đánh giá Timestamp.



A table with numbers and symbols

AI-generated content may be incorrect.

Phân phối của các điểm đánh giá cho trong biểu đồ dưới đây cho chúng ta thấy rằng điểm 4 được đánh giá nhiều nhất trong khi các điểm 1 và 2 có ít lượng đánh giá nhất . Điều này có thể được giải thích bằng sự thật là người dùng thường đánh giá khi họ rất thích một bộ phim; khi họ không thực sự thích, họ sẽ ít ra đánh giá hơn.

A graph with blue bars

AI-generated content may be incorrect.Tiếp theo, chúng ta sẽ xem số lượng bộ phim mà mỗi người dùng đánh giá cũng như số lượng đánh giá mà mỗi bộ phim nhận được:

Hình 3. 1: Biểu đồ biểu diễn điểm đánh giá

A screenshot of a computer

AI-generated content may be incorrect.

Ta thấy rằng người dùng có mã số 4169 đánh giá tới 2314 bộ phim và 20 là số lượng đánh giá ít nhất mà mỗi người dùng đưa ra. Có thể thấy rằng nhóm tác giả của bộ dữ liệu này đã lọc đi các người dùng có ít đánh giá. Sự lý tưởng này khó đạt được trong thực tế vì phần lớn người dùng không đưa ra đánh giá nào. Việc mỗi người dùng đánh giá nhiều bộ phim khiến cho độ chính xác khi gợi ý được cao hơn.

A screenshot of a computer program

AI-generated content may be incorrect.

Ở khía cạnh bộ phim, bộ phim có mã số 2858 được đánh giá nhiều nhất với 3428 lần trong khi rất nhiều bộ phim chỉ nhận được một đánh giá.

* + - * Dữ liệu người dùng

A screenshot of a computer

AI-generated content may be incorrect.

Như vậy có 6040 người dùng cùng với đầy đủ các thông tin về giới tính, tuổi, nghề nghiệp và Zip-code. Chúng ta sẽ không sử dụng thông tin về Zip-code vì số lượng các giá trị phân biệt là quá lớn. Chúng ta cùng xem nhanh phân bố của các thông tin về giới tính, tuổi và nghề nghiệp.

A screenshot of a computer code

AI-generated content may be incorrect.

Có 4331 người dùng là nam và 1709 người dùng là nữ

Hình 3. 2: Biểu đồ biểu diễn điểm đánh giáA graph with blue bars

AI-generated content may be incorrect.

Phần lớn người dùng có độ tuổi từ 18 đến 34, nhóm dưới 18 tuổi có số người dùng nhỏ nhất.

* + - * Dữ liệu bộ phim

A screenshot of a computer program

AI-generated content may be incorrect.

Có 3883 bộ phim với đầy đủ thông tin về tiêu đề (Title) và các thể loại (Genres). Cùng xem một vài dòng đầu của dataframe này:

A screenshot of a movie list

AI-generated content may be incorrect.

## 3.2 Biến đổi dữ liệu

Dữ liệu về nghề nghiệp đã được mã hóa thành các số từ 0 đến 20:

A screenshot of a computer

AI-generated content may be incorrect.

A computer screen shot of a code

AI-generated content may be incorrect.

A graph of blue bars with black text

AI-generated content may be incorrect.

Đánh số thứ tự cho các bộ phim và xây dựng một ánh xạ giữa các chỉ số này tới MovieID.



Tiếp theo, ta sẽ xây dựng vector đặc trưng nhị phân cho mỗi bộ phim như mô tả ở trên. Các vector này được gom lại trong một mảng numpy.

A screenshot of a computer program

AI-generated content may be incorrect.

## 3.3 Phân chia dữ liệu học

**A computer code with text

AI-generated content may be incorrect.**

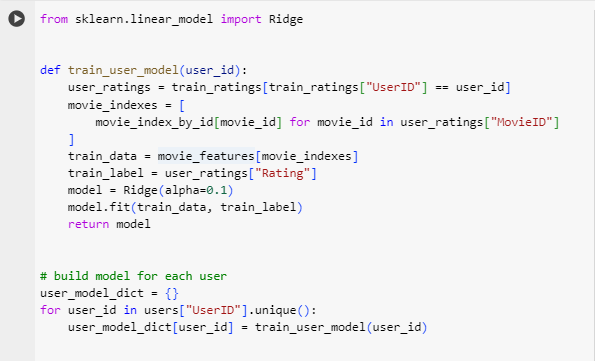
A screenshot of a computer code

AI-generated content may be incorrect.

Sau khi phân chia dữ liệu, tập kiểm thử (validation set) chỉ chứa dữ liệu của 5970/6040 người dùng.

## 3.4 Huấn luyện mô hình

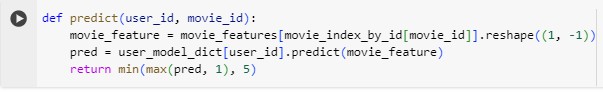
Với mỗi người dùng, ta sẽ xây dựng một bộ hồi quy Ridge dựa trên các đánh giá cho các bộ phim trong tập huấn luyện:



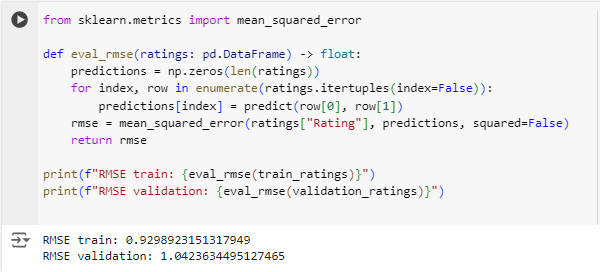
## 3.5 Tính tỉ lệ

Để dự đoán độ yêu thích của một người dùng tới một bộ phim, ta chỉ cần đưa vector đặc trưng vào mô hình ứng với người dùng đó. Vì các đánh giá nằm trong đoạn từ 1 tới 5, ta cần *cắt* những dự đoán nằm ngoài khoảng này:

18



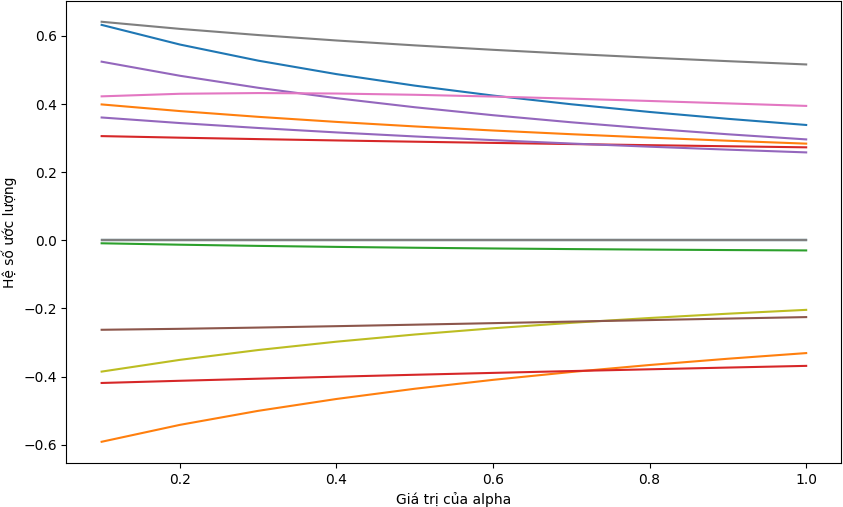
Cuối cùng, ta có thể đánh giá chất lượng hệ thống trên hai bộ dữ liệu huấn luyện và kiểm thử bằng [Root Mean Square Error - RMSE](https://en.wikipedia.org/wiki/Root-mean-square_deviation):



Trên tập huấn luyện, RMSE = 0.93, trên tập kiểm thử, RMSE = 1.04. Như vậy, trên tập kiểm thử, mỗi dự đoán bị lệch khoảng 1.04 điểm.

## 3.6 Biểu diễn

Hình 3. 3: Biểu đồ biểu diễn MSE theo các giá trị alpha trong hồi quy Ridge



Hình 3. 4: Đồ thị biểu diễn hệ số ước lượng theo các giá trị alpha

# KẾT LUẬN

Kết quả thu được từ mô hình hồi quy Ridge cho thấy một mức độ chính xác đáng kể trong việc dự đoán thể loại phim yêu thích của người dùng, điều này chứng minh tiềm năng của mô hình trong việc gợi ý nội dung phù hợp với sở thích cá nhân. Điều này không chỉ cải thiện trải nghiệm người dùng mà còn có thể hỗ trợ các nền tảng phát trực tuyến trong việc tối ưu hóa hệ thống đề xuất của họ. Trong tương lai, chúng em kỳ vọng sẽ mở rộng nghiên cứu bằng cách thử nghiệm với các mô hình học máy khác và so sánh hiệu suất của chúng, cũng như tích hợp thêm các biến đầu vào để cải thiện khả năng dự đoán của mô hình.

Ngoài ra, chúng ta cũng có thể nhấn mạnh về tầm quan trọng của việc thu thập và tiền xử lý dữ liệu đúng cách để đảm bảo rằng mô hình hồi quy Ridge hoạt động hiệu quả. Sự hiểu biết sâu sắc về dữ liệu là chìa khóa để xây dựng một mô hình chính xác và tin cậy.

Cuối cùng, trong tương lai, nếu có cơ hội, nghiên cứu có thể mở rộng bằng cách thử nghiệm các phương pháp hồi quy khác nhau và kết hợp chúng với các kỹ thuật máy học khác để tăng cường hiệu suất dự đoán. Điều này có thể giúp chúng ta hiểu sâu hơn về đặc điểm của dữ liệu và cách tối ưu hóa mô hình để đạt được kết quả tốt nhất.

# DANH MỤC TÀI LIỆU THAM KHẢO

1. Tom Mitchell. *Machine Learning.* McGraw-Hill, 1997.
2. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep Learning. MIT press, 2016.
3. Jiawei Han, Micheline Kamber, Jian Pei. Data Mining: Concepts and Techniques (3rd Edition). Morgan Kaufmann, 2011.
4. Bài giảng môn Nhập môn học máy – Th.S Nguyễn Thị Thanh Tân