**ĐẠI HỌC QUỐC GIA HÀ NỘI**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

A blue and white circle with text

AI-generated content may be incorrect.

**PHÂN TÍCH VÀ XÂY DỰNG SPLINE BẬC BA CHO DỮ LIỆU QUY MÔ LỚN (TRÊN 1000 ĐIỂM)**

**BÁO CÁO PROJECT CUỐI KỲ MÔN PHƯƠNG PHÁP TÍNH**

**Hà Nội, tháng 05/2025**

**TÓM TẮT**

Nội suy dữ liệu đóng vai trò quan trọng trong nhiều lĩnh vực khoa học và kỹ thuật, đặc biệt khi xử lý các tập dữ liệu rời rạc. Với sự gia tăng của dữ liệu quy mô lớn, việc tìm kiếm các phương pháp nội suy vừa đảm bảo độ chính xác, độ trơn của đường cong kết quả, vừa có hiệu năng tính toán tốt trở nên cấp thiết. Dự án này tập trung vào việc phân tích và xây dựng phương pháp nội suy spline bậc ba, một kỹ thuật hiệu quả để làm mịn và biểu diễn dữ liệu, cho các tập dữ liệu có quy mô lớn (trên 1000 điểm).

Mục tiêu chính của dự án là nghiên cứu sâu về cơ sở lý thuyết của spline bậc ba, bao gồm các điều kiện biên phổ biến như spline tự nhiên (natural spline). Trên cơ sở đó, nhóm đã tiến hành phân tích, lựa chọn và triển khai thuật toán xây dựng spline bậc ba tối ưu, tập trung vào việc giải hệ phương trình ba đường chéo một cách hiệu quả để xử lý lượng lớn điểm dữ liệu. Chương trình được xây dựng bằng ngôn ngữ Python với sự hỗ trợ của thư viện NumPy để tính toán các hệ số spline và thực hiện nội suy.

Kết quả thử nghiệm trên các bộ dữ liệu mô phỏng với số điểm dữ liệu đa dạng (vượt ngưỡng 1000 điểm) cho thấy phương pháp spline bậc ba đã xây dựng không chỉ cho kết quả nội suy chính xác, tạo đường cong trơn mượt qua các điểm dữ liệu mà còn đảm bảo thời gian xử lý hợp lý, phù hợp với yêu cầu của dữ liệu quy mô lớn. Dự án cung cấp một công cụ hữu ích và minh chứng cho tính khả thi và hiệu quả của spline bậc ba trong các ứng dụng phân tích dữ liệu hiện đại.

***Từ khóa****: nội suy, spline bậc ba, dữ liệu quy mô lớn, phương pháp tính, làm mịn dữ liệu.*

**CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI**

**1.1. Đặt vấn đề**

Trong kỷ nguyên số hiện nay, dữ liệu được tạo ra và thu thập với tốc độ chưa từng có ở hầu hết các lĩnh vực khoa học, kỹ thuật, kinh tế và xã hội. Việc phân tích và trích xuất thông tin hữu ích từ các tập dữ liệu rời rạc này đóng vai trò then chốt. Một trong những bài toán cơ bản và quan trọng trong quá trình xử lý dữ liệu là nội suy – ước lượng giá trị của một hàm tại các điểm chưa biết dựa trên một tập hợp các điểm dữ liệu đã cho. Mặc dù các phương pháp nội suy đa thức cổ điển như Lagrange hay Newton cung cấp lời giải trực tiếp, chúng thường gặp phải những hạn chế đáng kể khi bậc của đa thức lớn hoặc khi số lượng điểm dữ liệu tăng cao, điển hình là hiện tượng dao động mạnh (hiện tượng Runge) và tính toán không ổn định.

Để khắc phục những nhược điểm này, phương pháp nội suy spline, đặc biệt là spline bậc ba, đã được nghiên cứu và ứng dụng rộng rãi. Spline bậc ba không chỉ đảm bảo tính trơn của đường cong nội suy bằng cách duy trì sự liên tục của đạo hàm bậc nhất và bậc hai tại các điểm nút, mà còn có khả năng mô tả dữ liệu một cách chính xác và linh hoạt hơn so với một đa thức bậc cao duy nhất.

Trong bối cảnh các ứng dụng thực tế ngày càng phải xử lý những tập dữ liệu có quy mô lớn (ví dụ, trên 1000 điểm dữ liệu như trong phạm vi đề tài này), việc phân tích và xây dựng các thuật toán nội suy spline bậc ba không chỉ đòi hỏi sự chính xác mà còn cần chú trọng đến hiệu quả tính toán và sự ổn định của thuật toán. Việc này đặt ra yêu cầu về việc nghiên cứu các phương pháp triển khai hiệu quả, có khả năng mở rộng tốt khi kích thước dữ liệu tăng lên. Chính vì vậy, đề tài "Phân tích và xây dựng Spline bậc ba cho dữ liệu quy mô lớn (trên 1000 điểm)" được nhóm chúng em lựa chọn thực hiện nhằm giải quyết những thách thức này, đồng thời cung cấp một công cụ nội suy mạnh mẽ và đáng tin cậy.

**1.2. Mục tiêu của dự án**

Dự án này được thực hiện với các mục tiêu cụ thể như sau:

* Nghiên cứu và hệ thống hóa cơ sở lý thuyết về phương pháp nội suy spline, đặc biệt là spline bậc ba (cubic spline) và các điều kiện biên phổ biến.
* Phân tích các thuật toán xây dựng spline bậc ba, đánh giá ưu nhược điểm và lựa chọn thuật toán phù hợp cho việc xử lý dữ liệu quy mô lớn, đảm bảo hiệu quả về thời gian tính toán và độ chính xác.
* Thiết kế và triển khai một chương trình/module hoàn chỉnh có khả năng tính toán các hệ số của spline bậc ba và thực hiện nội suy giá trị tại các điểm bất kỳ từ một tập dữ liệu đầu vào có quy mô lớn (trên 1000 điểm).
* Kiểm thử và đánh giá hiệu năng (thời gian thực thi, mức sử dụng tài nguyên) cũng như độ chính xác của chương trình đã xây dựng thông qua các bộ dữ liệu thử nghiệm đa dạng.
* So sánh kết quả (nếu có điều kiện) với các thư viện chuẩn hoặc các công cụ đã có để có cái nhìn khách quan về chất lượng của sản phẩm.

**1.3. Đối tượng và phạm vi nghiên cứu**

* **Đối tượng nghiên cứu:**
  + Phương pháp nội suy spline bậc ba.
  + Các thuật toán giải hệ phương trình tuyến tính phát sinh trong quá trình xây dựng spline bậc ba (ví dụ: thuật toán Thomas cho hệ ba đường chéo).
  + Các tập dữ liệu một chiều (​,​) có quy mô lớn (số điểm dữ liệu *n+*1*>*1000), có thể là dữ liệu mô phỏng từ các hàm toán học hoặc dữ liệu thực tế (nếu có).
* **Phạm vi nghiên cứu:**
  + Tập trung chủ yếu vào spline bậc ba tự nhiên (natural cubic spline) do tính phổ biến và điều kiện biên đơn giản, phù hợp với nhiều loại dữ liệu.
  + Ngôn ngữ lập trình dự kiến sử dụng để triển khai là Python, sử dụng thư viện NumPy để tối ưu hóa tính toán.
  + Đánh giá độ chính xác của phương pháp nội suy (ví dụ: thông qua sai số RMSE nếu hàm gốc được biết) và hiệu năng tính toán (thời gian xây dựng spline, thời gian nội suy).
  + Dự án giới hạn ở bài toán nội suy một chiều.
  + Không đi sâu vào các loại spline phức tạp hơn như B-spline, NURBS hay nội suy đa chiều, trừ khi có mở rộng cụ thể.
  + Khía cạnh "quy mô lớn" được hiểu là số lượng điểm dữ liệu đủ lớn (trên 1000) để các yếu tố về hiệu quả thuật toán và quản lý bộ nhớ trở nên đáng kể trên một máy tính cá nhân, chứ không nhất thiết đòi hỏi các giải pháp điện toán phân tán.

**1.4. Phương pháp nghiên cứu**

Để đạt được các mục tiêu đề ra, nhóm chúng em dự kiến sử dụng kết hợp các phương pháp nghiên cứu sau:

* **Nghiên cứu lý thuyết:** Thu thập, đọc hiểu và tổng hợp kiến thức từ các giáo trình, tài liệu khoa học, bài báo chuyên ngành về phương pháp số, lý thuyết nội suy, và cụ thể là nội suy spline bậc ba.
* **Phân tích và thiết kế thuật toán:** Dựa trên nền tảng lý thuyết, tiến hành phân tích các thuật toán xây dựng spline bậc ba, đặc biệt chú trọng đến các thuật toán giải hệ phương trình hiệu quả. Thiết kế cấu trúc dữ liệu và luồng xử lý cho chương trình.
* **Triển khai thực nghiệm:** Hiện thực hóa thuật toán đã chọn bằng ngôn ngữ lập trình Python. Xây dựng các module chức năng để nhập liệu, tính toán spline và xuất kết quả.
* **Kiểm thử và đánh giá:**
  + Xây dựng các bộ dữ liệu thử nghiệm đa dạng: dữ liệu từ các hàm toán học đã biết (để kiểm tra độ chính xác), dữ liệu có nhiễu, dữ liệu với số lượng điểm khác nhau (1000, 5000, 10000 điểm, v.v.).
  + Thực hiện các kịch bản kiểm thử để đánh giá độ chính xác, tốc độ thực thi và khả năng làm mịn của đường spline.
* **So sánh và tổng hợp:** Phân tích kết quả thu được, so sánh với các mục tiêu ban đầu. Nếu có thể, đối chiếu với kết quả từ các thư viện chuẩn (ví dụ: SciPy trong Python) để có cơ sở đánh giá khách quan.

**1.5. Ý nghĩa khoa học và thực tiễn của đề tài**

* **Ý nghĩa khoa học:**
  + Góp phần củng cố và làm sâu sắc thêm kiến thức về một phương pháp nội suy quan trọng trong lĩnh vực phương pháp tính và khoa học dữ liệu.
  + Cung cấp những đánh giá thực nghiệm về hiệu quả và thách thức khi áp dụng spline bậc ba cho các tập dữ liệu quy mô lớn, làm cơ sở cho các nghiên cứu cải tiến hoặc ứng dụng sau này.
* **Ý nghĩa thực tiễn:**
  + Xây dựng được một công cụ phần mềm có khả năng thực hiện nội suy spline bậc ba hiệu quả, có thể được ứng dụng trong nhiều bài toán thực tế như: xử lý tín hiệu, phân tích chuỗi thời gian, đồ họa máy tính, mô hình hóa trong kỹ thuật, làm mịn dữ liệu trước khi phân tích sâu hơn.
  + Đối với sinh viên, đề tài giúp rèn luyện kỹ năng nghiên cứu, phân tích vấn đề, thiết kế thuật toán, lập trình và đánh giá kết quả – những kỹ năng cần thiết cho công việc sau này.
  + Kết quả của đề tài có thể làm tài liệu tham khảo cho các sinh viên khóa sau quan tâm đến lĩnh vực này.

**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT**

Chương này trình bày các khái niệm và kiến thức lý thuyết cơ bản liên quan đến bài toán nội suy, từ các phương pháp nội suy đa thức truyền thống đến nội suy spline, đặc biệt tập trung vào spline bậc ba – đối tượng nghiên cứu chính của đề tài.

**2.1. Giới thiệu về nội suy đa thức**

2.1.1. Khái niệm nội suy

Trong nhiều bài toán khoa học và kỹ thuật, chúng ta thường được cung cấp các giá trị của một hàm số  tại một tập hợp các điểm rời rạc . Mục tiêu đặt ra là tìm các giá trị xấp xỉ của hàm  cho những điểm *x* "mới" nằm giữa các điểm đã cho này, tức là những điểm mà giá trị của hàm chưa được biết. Quá trình này được gọi là **nội suy (interpolation)**.

Cụ thể hơn, chúng ta có các giá trị của hàm *f* dưới dạng:

hoặc dưới dạng các cặp giá trị có thứ tự:



Các giá trị cho trước này có thể xuất phát từ một hàm toán học (ví dụ như hàm logarit hoặc hàm Bessel) hoặc có thể được đo lường hay ghi lại tự động từ một hàm thực nghiệm (ví dụ như sức cản không khí của ô tô hoặc máy bay ở các tốc độ khác nhau, hoặc các ví dụ về các hàm "thực nghiệm" là kết quả của một quá trình hóa học ở các nhiệt độ khác nhau hoặc quy mô dân số Hoa Kỳ được lấy từ các cuộc điều tra dân số được thực hiện 10 năm một lần).

Một ý tưởng tiêu chuẩn trong nội suy là tìm một đa thức có bậc *n* (hoặc nhỏ hơn)

thỏa mãn các giá trị đã cho, nghĩa là:  Chúng ta gọi ​ là một **đa thức nội suy** và các điểm ​ là các **điểm nút (nodes)**. Nếu *f(x)* là một hàm toán học, chúng ta gọi ​ là một **xấp xỉ đa thức (polynomial approximation)** của *f*. Chúng ta sử dụng ​ để thu được các giá trị xấp xỉ cho *x* trong khoảng  ​ ("nội suy") và đôi khi cho *x* nằm ngoài khoảng này ("ngoại suy (extrapolation)").

Lý do lựa chọn đa thức cho mục đích nội suy là vì chúng thuận tiện để làm việc, dễ dàng tính toán đạo hàm và tích phân, và theo định lý xấp xỉ Weierstrass, mọi hàm liên tục trên một khoảng đóng có thể được xấp xỉ bởi một đa thức với độ chính xác mong muốn. Quan trọng hơn, với một tập hợp n+1 điểm dữ liệu (xi​,fi​) với các xi​ phân biệt, đa thức nội suy pn​(x) có bậc không quá n thỏa mãn điều kiện (\*) luôn **tồn tại và là duy nhất**.

2.1.2. Các phương pháp nội suy đa thức cơ bản và hạn chế

Một trong những cách tiếp cận phổ biến nhất cho bài toán nội suy là sử dụng đa thức. Với n+1 điểm dữ liệu, luôn tồn tại duy nhất một đa thức nội suy Pn​(x) có bậc không quá n đi qua tất cả các điểm này.

* **Đa thức nội suy Lagrange:** Đa thức nội suy Lagrange được xây dựng dựa trên các đa thức cơ sở :



trong đó các đa thức cơ sở Lagrange  được định nghĩa là:  ​​

Ưu điểm của đa thức Lagrange là dạng tường minh, dễ hình dung. Tuy nhiên, việc tính toán lại toàn bộ đa thức khi thêm một điểm dữ liệu mới là một nhược điểm.

* **Đa thức nội suy Newton:** Đa thức nội suy Newton sử dụng các tỷ sai phân (divided differences) và có dạng:  trong đó  là tỷ sai phân cấp *k*. Ưu điểm của đa thức Newton là có thể dễ dàng thêm điểm dữ liệu mới mà không cần tính toán lại từ đầu.
* **Hạn chế của nội suy đa thức bậc cao:** Mặc dù đa thức nội suy bậc *n* đi qua chính xác *n+1* điểm dữ liệu, khi *n* lớn (tức là số lượng điểm dữ liệu nhiều), đa thức nội suy có thể dao động rất mạnh giữa các điểm nút, đặc biệt là ở gần hai đầu của khoảng nội suy. Hiện tượng này được gọi là **hiện tượng Runge**. Sự dao động này làm cho đa thức nội suy bậc cao trở nên không phù hợp để xấp xỉ hàm số trên toàn khoảng, nhất là khi dữ liệu có nhiễu hoặc khi yêu cầu đường cong nội suy phải "trơn". Hơn nữa, việc tính toán các hệ số của đa thức bậc cao cũng có thể gặp vấn đề về độ ổn định số.

**2.2. Nội suy Spline**

2.2.1. Ý tưởng cơ bản

Để khắc phục những hạn chế của nội suy đa thức bậc cao, phương pháp nội suy spline được đề xuất. Ý tưởng cơ bản của nội suy spline là thay vì sử dụng một đa thức bậc cao duy nhất trên toàn bộ khoảng dữ liệu , ta chia khoảng này thành các khoảng con  bởi các điểm nút, và trên mỗi khoảng con này, ta sử dụng một đa thức bậc thấp riêng biệt. Các đa thức bậc thấp này sau đó được "nối" lại với nhau tại các điểm nút một cách trơn tru bằng cách áp đặt các điều kiện liên tục về giá trị của hàm và các đạo hàm của nó.

2.2.2. Định nghĩa Spline và bậc của Spline

Một hàm spline *S(x)* bậc *m* với các điểm nút ​ là một hàm thỏa mãn các điều kiện sau:

1. Trên mỗi khoảng con ​ (với *i=0,1,…,n−1*), *S(x)* là một đa thức bậc không quá *m*.
2. Hàm *S(x)* và các đạo hàm của nó từ cấp 1 đến cấp *m−1* là liên tục trên toàn bộ khoảng .

Bậc của spline (m) xác định độ "trơn" của đường cong nội suy. Spline bậc một (linear spline) tạo ra các đường gấp khúc nối các điểm dữ liệu. Spline bậc hai (quadratic spline) đảm bảo sự liên tục của đạo hàm bậc nhất. Spline bậc ba (cubic spline), đối tượng chính của đề tài này, đảm bảo sự liên tục của cả đạo hàm bậc nhất và bậc hai, tạo ra đường cong nội suy rất trơn và được sử dụng rộng rãi trong thực tế.

**2.3. Spline bậc ba (Cubic Spline)**

Như đã đề cập ở mục 2.2, spline bậc ba là loại spline được sử dụng rộng rãi nhất do tính trơn và khả năng mô tả dữ liệu phức tạp. Chúng ta sẽ đi sâu vào định nghĩa và cách xây dựng chúng.

2.3.1. Định nghĩa toán học và các điều kiện

Một hàm **spline bậc ba (cubic spline)** *g(x)* nội suy *n+1* điểm dữ liệu  (với ​) là một hàm thỏa mãn các điều kiện sau:

1. *g(x)* là một hàm liên tục trên toàn bộ khoảng 
2. Đạo hàm bậc nhất *g′(x)* và đạo hàm bậc hai *g′′(x)* của *g(x)* là liên tục trên 
3. **Điều kiện nội suy:** *g(x)* phải đi qua tất cả các điểm dữ liệu đã cho:

 ​với *j=0,1,…,n* (2.3.1)

1. Trên mỗi khoảng con  ( với *j=0,1,…,n−1* ), hàm *g(x)* là một đa thức bậc ba (hoặc thấp hơn), ký hiệu là . Nghĩa là: khi  (2.3.2) Từ điều kiện nội suy (2.3.1), rõ ràng mỗi đa thức *qj​(x)* phải thỏa mãn:

 ​và ​với *j=0,1,…,n−1* (2.3.3)

2.3.2. Xác định các đa thức thành phần 

Để xác định *n* đa thức , mỗi đa thức có 4 hệ số, tổng cộng chúng ta có *4n* hệ số chưa biết. Một cách để xác định  trên mỗi đoạn  là sử dụng giá trị hàm *fj*​, *fj+1*​ và giá trị đạo hàm bậc nhất ,  tại hai đầu mút của đoạn đó. Đa thức  có thể được biểu diễn dưới dạng đa thức Hermite bậc ba.

Đặt  ​ và  ​. Khi đó, với :

​

Trong công thức này,  là các giá trị đạo hàm bậc nhất của hàm spline tại các điểm nút, đây chính là các ẩn số chúng ta cần xác định.

2.3.3. Điều kiện liên tục đạo hàm cấp hai và hệ phương trình cho *kj*​

Từ biểu thức (2.3.4) của *qj​(x),* ta có thể tính đạo hàm bậc hai của nó tại các điểm  ​ và ​:

​ (2.3.5)

 ​(2.3.6)

Điều kiện quan trọng để *g(x)* là một spline bậc ba là đạo hàm bậc hai *g′′(x)* phải liên tục tại các điểm nút bên trong  (với *j=1,…,n−1*). Điều này có nghĩa là đạo hàm bậc hai của đa thức bên trái  tại  phải bằng đạo hàm bậc hai của đa thức bên phải  tại  :

 với *j=1,…,n−1*

Sử dụng biểu thức (2.3.6) cho  (thay *j* bằng *j−1* và ​ bằng  ​) và biểu thức (2.3.5) cho , sau khi thay ​ và  ​, ta thu được hệ *n−1* phương trình tuyến tính sau đây cho các ẩn  ​:

 (2.3.7)

với *j=1,…,n−1*.

2.3.4. Điều kiện biên

Hệ phương trình (2.3.7) gồm *n−1* phương trình nhưng có *n+1* ẩn số ( ​). Do đó, chúng ta cần thêm hai điều kiện nữa, gọi là các điều kiện biên, để có thể giải được hệ này.

* **Spline bậc ba ràng buộc (Clamped Cubic Spline):** Trong trường hợp này, giá trị đạo hàm bậc nhất tại hai đầu mút *x0*​ và *xn*​ được cho trước:

*g′(x0​)=k0*​ và *g′(xn​)=kn*​

Với *k0​* và *kn*​ đã biết, hệ (2.3.7) trở thành một hệ *n−1* phương trình với *n−1* ẩn *k1*​,…,*kn−1*​.

* **Spline bậc ba tự nhiên (Natural Cubic Spline):** Điều kiện biên tự nhiên yêu cầu đạo hàm bậc hai tại hai đầu mút bằng không:

*g′′(x0​)=0* và *g′′(xn​)=0* (2.3.8)

Để áp dụng các điều kiện này vào hệ (2.3.7), ta sử dụng:

* 1. : Từ (2.3.5) với *j=*0, ta có , hay:

 (2.3.9a)

* 1. : Từ (2.3.6) với *j=n−1* (điểm cuối là ​), ta có

, hay:

 (2.3.9b)

Kết hợp (2.3.9a), (2.3.9b) với *n−1* phương trình trong (2.3.7) tạo thành một hệ *n+1* phương trình tuyến tính cho *n+1* ẩn *k0​,k1​,…,kn*​.

2.3.5. Giải hệ phương trình và xác định hệ số spline

Hệ phương trình tuyến tính thu được cho các *kj​* có dạng ma trận **ba đường chéo (tridiagonal matrix)**. Ma trận này thường **chéo trội nghiêm ngặt (strictly diagonally dominant)**, đảm bảo rằng hệ có nghiệm duy nhất và có thể được giải một cách hiệu quả và ổn định về mặt số học bằng các thuật toán như thuật toán Thomas (TDMA), với chi phí tính toán *O(n).*

Sau khi đã giải được các giá trị  (với *j=0,…,n*), các hệ số ​ của mỗi đa thức thành phần  trong biểu thức (2.3.2) được xác định như sau:









với  ​.

Trong trường hợp các điểm nút cách đều,  (do đó ), hệ phương trình (2.3.7) có thể được viết đơn giản hơn thành:



với *j=1,…,n−1*

**2.4. Thuật toán giải hệ phương trình cho spline bậc ba**

Để xây dựng hàm nội suy spline bậc ba *S* cho hàm *f*, được xác định tại các điểm  ​, thỏa mãn :

**ĐẦU VÀO** n;​  ​; .

**ĐẦU RA**  với *j=0,1,…,n−1*.

(Lưu ý:  với ​)

**Bước 1** Với  đặt  ​.

**Bước 2** Với  đặt

.

**Bước 3** Đặt ;

;

.

**Bước 4** Với *i=1,2,…,n−1*

đặt ​;

 ​;

 ​.

**Bước 5** Đặt ;

;

.

**Bước 6** Với 

đặt ​;

;



**Bước 7** XUẤT KẾT QUẢ ( với ); DỪNG.