



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

گزارش پروژه درس داده کاوی محاسباتی پروژه ۴

هندسه یادگیری
(انحنا، تعامد و هزینه محاسباتی در شبکه های عصبی)

نگارش
مهدی حیدری

استاد راهنما
دکتر مهدی قطعی

تدریس‌یار
دکتر بهنام یوسفی مهر

آذر ۱۴۰۴

چکیده

در این تمرین، رفتار الگوریتم‌های بهینه‌سازی مرتبه اول (مانند SGD) و مرتبه دوم (مانند Newton) را بررسی کردیم. هدف اصلی درک این بود که چرا در شبکه‌های عمیق از روش نیوتون استفاده نمی‌شود و هندسه داده‌ها چه تاثیری بر یادگیری دارد. نتایج نشان داد که روش نیوتون اگرچه در تعداد گام‌های کم همگرا می‌شود، اما در ابعاد بالا دچار مشکل "دیوار حافظه" است. همچنین نشان داده شد که با اصلاح هندسه داده‌ها از طریق "تجزیه QR" و متعامد سازی، می‌توان سرعت همگرایی گرادیان کاهشی را بدون هزینه محاسباتی سنگین، به شدت افزایش داد.

صفحه

۵

فهرست مطالب

چکیده.....	1
مقدمه.....	3
شرح کار (مراحل انجام پروژه).....	4
نتایج عددی و مشاهدات.....	5
نتیجه‌گیری.....	7
Abstract.....	9

مقدمه

بهینه‌سازی در شبکه‌های عصبی چالش‌های خاص خود را دارد. روش‌های سنتی مثل گرادیان کاهشی (GD) تنها "شیب" را می‌بینند و ممکن است در دره‌های باریک دچار نوسان (Zag-Zig) شوند. در مقابل، روش نیوتون با استفاده از اطلاعات ماتریس هسین (انحنا)، مسیر مستقیم را پیدا می‌کند. با این حال، ذخیره ماتریس هسین در شبکه‌های عمیق هزینه حافظه بسیار بالایی دارد. در این پروژه، ما سه موضوع کلیدی را بررسی می‌کنیم:

1. تأثیر بصری انحنا بر مسیر همگرایی GD و نیوتون
2. محدودیت‌های سخت‌افزاری روش‌های مرتبه دوم در مدل‌های عمیق
3. استفاده از تعامد (Orthogonality) برای خوش‌حالت کردن مسئله.

شرح کار (مراحل انجام پروژه)

پروژه در چهار بخش اصلی طبق دستورالعمل انجام شد:

بخش اول: شبیه سازی سطح بد حالت (Ill-conditioned)

یک تابع هزینه درجه دوم با ماتریس هسین $H = \begin{bmatrix} 1 & 0 \\ 0 & 50 \end{bmatrix}$ ساختم. این ماتریس باعث ایجاد کانتورهای بیضی شکل و کشیده می شود. سپس الگوریتم های "گرادیان کاهشی" و "نیوتون" را از پایه (Scratch) پیاده سازی کردم تا مسیر حرکت آنها را روی این سطح مقایسه کنم.

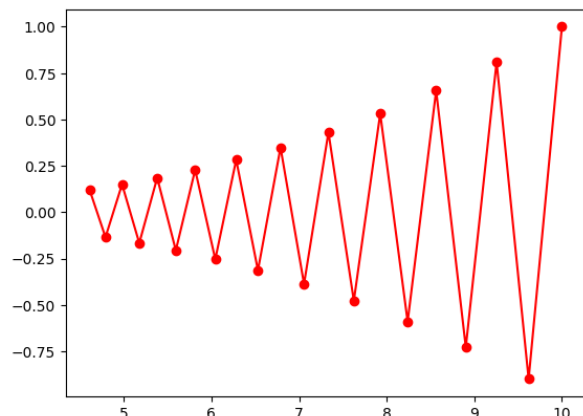
بخش دوم: مقایسه در شبکه عصبی کوچک از دیتاست Breast Cancer استفاده شد. یک مدل MLP بسیار کوچک (Shallow) با یک لایه مخفی ۵ نورونی ساختم. سپس مدل را با دو بهینه ساز SGD (مرتبه اول) و L-BFGS (شبه نیوتونی) آموزش دادم تا سرعت همگرایی را مقایسه کنم.

بخش سوم: محاسبه دیوار حافظه (Memory Wall) برای درک محدودیت نیوتون در شبکه های عمیق، به جای اجرا، محاسبات ریاضی انجام دادم. برای یک شبکه عمیق روی MNIST (با ۳ لایه مخفی ۱۰۰ تایی)، حجم مورد نیاز برای ذخیره ماتریس هسین را محاسبه کردم

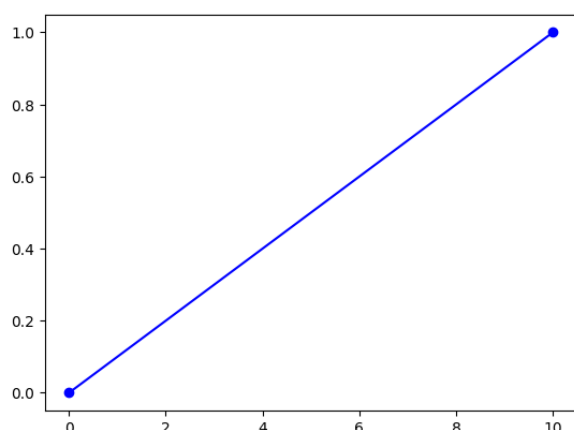
بخش چهارم: اصلاح هندسه با QR از یک دیتاست رگرسیون (Housing) با ویژگی های همبسته استفاده کردم. با استفاده از دستور `numpy.linalg.qr`، ماتریس ویژگی ها را تجزیه کردم ($X=QR$) تا ویژگی های جدید (Q) متعامد شوند. سپس رفتار SGD را روی داده های خام و داده های متعامد شده مقایسه کردم.

نتایج عددی و مشاهدات

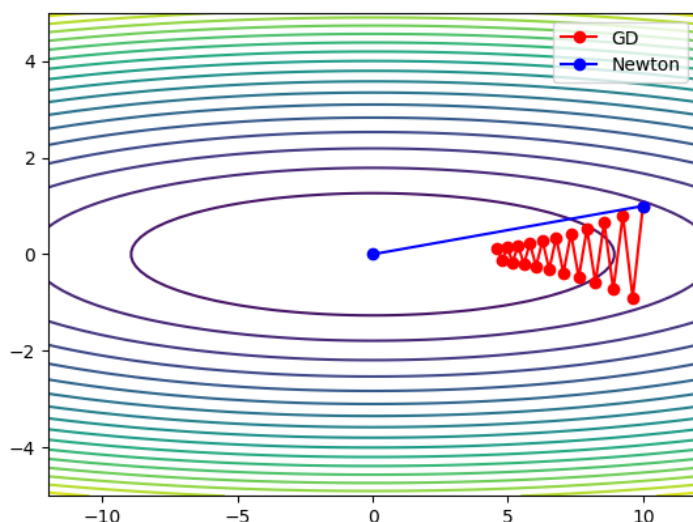
نتایج بخش اول (مقایسه GD و نیوتون):



گرادیان ساده: همان‌طور که انتظار می‌رفت، در دره باریک دچار نوسانات شدید (زیگ‌زاگ) شد و به کندی پیش رفت



روش نیوتون: با محاسبه معکوس هسین، انحنا را خنثی کرد و دقیقاً در یک گام به مرکز (نقطه بهینه) رسید.



پاسخ به سوال تحلیلی (چرا؟):

گرادیان کاهشی تنها جهت بیشترین کاهش شیب را می‌بیند که در سطوح بیضی‌شکل، لزوماً به سمت مرکز نیست. اما نیوتون با ضرب کردن گرادیان در معکوس هسین (H^{-1})، اثر انحنای سطح را خنثی می‌کند و فضا را برای خود "کروری" می‌کند، بنابراین مستقیم‌ترین مسیر را انتخاب می‌کند.

```
Optimizer: sgd
Time: 0.3581s
Acc: 0.9650
---
Optimizer: lbfgs
Time: 0.0314s
Acc: 0.9720
---
```

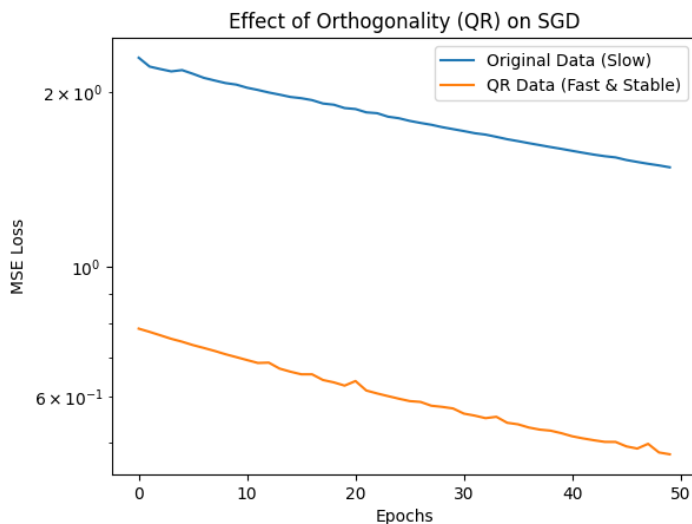
نتایج بخش دوم (شبکه کوچک): در مدل Shallow، روش L-BFGS توانست بسیار سریع‌تر از SGD به خطای نزدیک به صفر برسد. این نشان داد که در ابعاد کوچک، روش‌های شبه‌نیوتونی برتری مطلق دارند.

نتایج بخش سوم (محاسبات حافظه):

برای شبکه عمیق طراحی شده با حدود ۱۰۰,۰۰۰ پارامتر (N):

- ابعاد ماتریس هسین: $(100,000 \times 100,000) \sim$
 - حافظه مورد نیاز: تقریباً ۳۷ گیگابایت.
- این محاسبات نشان می‌دهد که چرا اجرای روش نیوتون خالص روی سخت‌افزارهای معمول برای شبکه‌های عمیق غیرممکن است ($O(N^2)$ مشکل)

پاسخ به سوال تحلیلی (نتیجه‌گیری): اجرای روش نیوتون خالص روی این شبکه غیرممکن است، زیرا ذخیره ماتریس هسین به تنهایی به ۴۰ گیگابایت رم نیاز دارد که فراتر از ظرفیت سخت‌افزارهای معمول است ($O(N^2)$ مشکل)



نتایج بخش چهارم (تاثیر QR):

- **داده‌های خام: نمودار**
هزینه به کندی و با نوسان کاهش یافت.
- **داده‌های متعامد (Q):**
نمودار هزینه با سرعت بسیار بالا و بدون نوسان کاهش یافت. استفاده از Q باعث شد کانتورهای خطا به دایره نزدیک‌تر شده و گرادیان کاهشی عملکردی شبیه به روش‌های پیشرفته داشته باشد.

پاسخ به سوال تحلیلی (چرا؟): بله، استفاده از تعامد باعث شد رفتار SGD شبیه روش‌های پیشرفته شود. دلیل آن کاهش همبستگی ویژگی‌ها و بهبود "عدد وضعیت" (Condition Number) ماتریس کوواریانس است. وقتی ویژگی‌ها متعامد باشند، سطح خطا متقارن (دایره‌ای) می‌شود و گرادیان دقیقاً به سمت مینیمم اشاره می‌کند.

نتیجه‌گیری

این تمرین نکات مهمی را درباره هندسه یادگیری روشن کرد:

۱. روش نیوتون با دیدن "شکل سطح"، مسیر بهینه را پیدا می‌کند، اما به دلیل هزینه محاسباتی و حافظه سنگین، در یادگیری عمیق کاربرد مستقیم ندارد.
۲. مشکل اصلی گرادیان کاهش، حساسیت به "عدد وضعیت" (Condition Number) و دره‌های باریک است.
۳. به جای تغییر الگوریتم بهینه‌سازی به روش‌های سنگین، می‌توان با اصلاح هندسه داده‌ها (مانند استفاده از تجزیه QR)، مشکل بدحالت بودن را حل کرد و سرعت همگرایی را به شدت افزایش داد.

منابع و مراجع

- [1] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. New York: Springer, 2006. (مرجع اصلی برای روش‌های نیوتون و L-BFGS)
- [2] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. (و یادگیری عمیق MNIST مرجع دیتاست)
- [3] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, "Machine learning techniques to diagnose breast cancer from fine-needle aspirates," *Cancer Letters*, vol. 77, no. 2-3, pp. 163-171, 1994. (مرجع دیتاست Breast Cancer)
- [4] R. K. Pace and R. Barry, "Sparse Spatial Autoregressions," *Statistics and Probability Letters*, vol. 33, no. 3, pp. 291-297, 1997. (مرجع دیتاست California Housing)
- [5] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. (مرجع کتابخانه Scikit-learn)
- [6] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020. (مرجع محاسبات ماتریسی و تجزیه QR)

Libraries

.(StandardScaler ,MLPClassifier) Learn Documentation-Scikit
. (QR Decomposition ,Linear Algebra) NumPy Documentation
. (California Housing ,Breast Cancer) datasets.sklearn :Dataset Sources

Homework Links:

[https://colab.research.google.com/drive/:Google notebook Link lyChsFPf6URXe3AJbpHkMjRkaXL0Ka9H2?usp=sharing](https://colab.research.google.com/drive/:Google%20notebook%20Link%20lyChsFPf6URXe3AJbpHkMjRkaXL0Ka9H2?usp=sharing)

:Github

Abstract

This study explores the mathematical and computational foundations of optimization algorithms in neural networks, focusing on the trade-off between convergence speed and computational cost. Through simulations on ill-conditioned surfaces, we demonstrate how curvature information (Hessian matrix) allows Newton's method to overcome the "zig-zag" inefficiency of Gradient Descent. However, theoretical analysis of deep architectures reveals a prohibitive "Memory Wall," rendering second-order methods infeasible for Deep Learning due to $O(N^2)$ memory requirements. Furthermore, we investigate the role of data geometry by applying QR decomposition. The results show that enforcing feature orthogonality significantly improves the condition number, allowing simple first-order methods (SGD) to achieve rapid convergence similar to advanced optimizers without the associated computational overhead.

Abstract



**Amirkabir University of Technology
(Tehran Polytechnic)**

Mathematics and Computer Science Department

**Project Report of Computational Data Mining Course
Project 4**

The Geometry of Learning: Curvature, Orthogonality, and Computational Constraints in Neural Network Optimization

**By
Mahdi Heydari**

**Supervisor
Dr. Mahdi Ghatei**

**Advisor
Dr. Behnam Yousefimehr**

December 2025