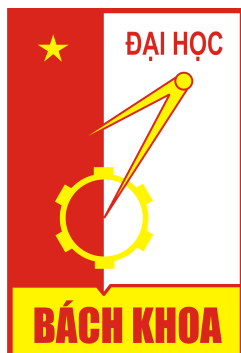


ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN TIN



ĐỒ ÁN II

**Thuật toán hướng giảm gradient ngẫu nhiên
với cỡ bước Polyak và ứng dụng**

Giảng viên hướng dẫn: TS. Phạm Thị Hoài

Sinh viên thực hiện: Phạm Thị Thanh Hà

Mã số sinh viên: 20216823

Lớp: Toán-Tin 01 K66

Hà Nội, Ngày 3 tháng 1 năm 2025

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

1. Mục tiêu và nội dung của đề án

- (a) Mục tiêu: Tìm hiểu về thuật toán hướng giảm gradient ngẫu nhiên với cỡ bước Polyak.
- (b) Nội dung: Trình bày phương pháp hướng giảm gradient ngẫu nhiên với cỡ bước Polyak giải bài toán tối ưu tổng hữu hạn trong các trường hợp lồi mạnh, lồi và không lồi. Tính toán thử nghiệm một số bài toán tương ứng với các trường hợp.

2. Kết quả đạt được

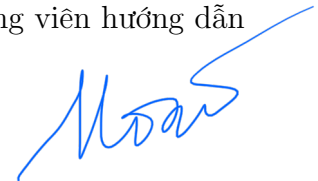
- (a) Nắm được các kiến thức chuẩn bị liên quan đến thuật toán hướng giảm gradient ngẫu nhiên, hàm L -trơn, hàm lồi, mô hình quá tham số, điều kiện PL.
- (b) Hiểu và trình bày được phương pháp SGD với cỡ bước Polyak ngẫu nhiên trong các trường hợp lồi mạnh, lồi và không lồi.
- (c) Lập trình thử nghiệm một số bài toán tương ứng.

3. Ý thức làm việc của sinh viên

Tốt

Hà Nội, ngày 03 tháng 01 năm 2025

Giảng viên hướng dẫn



TS. Phạm Thị Hoài

Lời cảm ơn

Để có thể hoàn thành đồ án này, em xin được gửi lời cảm ơn chân thành và sâu sắc nhất đến giảng viên hướng dẫn - TS. Phạm Thị Hoài. Cô đã tận tình hướng dẫn, góp ý, và hỗ trợ em trong suốt quá trình thực hiện đồ án. Nhờ đó, em học hỏi được nhiều điều cũng như phát triển kỹ năng một cách đáng kể. Đồ án này sẽ không thể hoàn thiện được nếu thiếu đi những sự giúp đỡ vô cùng to lớn và kịp thời của cô.

Em cũng xin gửi lời cảm ơn đến Khoa Toán - Tin đã thiết kế các học phần cần thiết trong chương trình giảng dạy của ngành Toán Tin để em được trang bị những kiến thức nền tảng sử dụng trong nội dung đồ án. Những kiến thức, kỹ năng và cách tư duy ấy sẽ luôn là hành trang quan trọng nhất của em trên con đường học tập hiện tại và sau này.

Em rất hy vọng đồ án sẽ cung cấp được những thông tin hữu ích, góp phần vào việc nâng cao hiểu biết về việc tối ưu hóa các thuật toán máy học hiện có. Song, bởi kiến thức và kinh nghiệm của bản thân còn nhiều hạn chế, em hiểu rằng mình không thể tránh khỏi những sai lầm và thiếu sót. Vì vậy, em rất mong nhận được những lời nhận xét, những ý kiến đóng góp, phê bình từ phía Thầy/Cô để đồ án được hoàn thiện hơn.

Cuối cùng, em xin kính chúc quý thầy cô thật nhiều sức khỏe, thành công và hạnh phúc.

Em xin chân thành cảm ơn!

Hà Nội, ngày 03 tháng 01 năm 2025

Tác giả đồ án

Phạm Thị Thanh Hà

Mục lục

Danh mục kí hiệu và chữ viết tắt	6
Danh sách hình vẽ	7
Mở đầu	8
1 Kiến thức chuẩn bị	9
1.1 Hàm lồi và một số tính chất	9
1.1.1 Hàm lồi và một số tính chất	9
1.1.2 Hàm lồi mạnh	10
1.2 Hàm L -trơn	11
1.3 Thuật toán hướng giảm	13
1.4 Điều kiện Polyak-Lojasiewicz (PL)	15
1.5 Mô hình quá tham số trong học sâu	17
2 Thuật toán hướng giảm Gradient ngẫu nhiên với cỡ bước Polyak	19
2.1 Thuật toán hướng giảm gradient ngẫu nhiên (SGD)	20
2.2 Cỡ bước Polyak ngẫu nhiên	20
2.2.1 Cỡ bước Polyak tắt định	20
2.2.2 Cỡ bước Polyak ngẫu nhiên	21
2.2.3 Sự chênh lệch mục tiêu tối ưu	22
2.3 Phân tích hội tụ của thuật toán gradient ngẫu nhiên với cỡ bước Polyak	23
2.3.1 Giới hạn trên và dưới của cỡ bước Polyak ngẫu nhiên	23
2.3.2 Hàm mục tiêu f là tổng của các hàm lồi và f lồi mạnh	23
2.3.3 Hàm f là tổng của các hàm lồi	25
2.3.4 Hàm f là tổng các hàm không lồi - f thỏa mãn điều kiện PL . .	28
2.3.5 Hàm f là hàm không lồi tổng quát	30

3	Lập trình thử nghiệm	34
3.1	Thử nghiệm so sánh thuật toán SGD với cỡ bước hằng và SPS_{\max} . . .	34
3.2	Thử nghiệm cho mô hình quá tham số	35
3.2.1	Phân loại nhị phân sử dụng kernel (Bài toán lỗi)	36
3.2.2	Nhân tử hóa ma trận - Deep matrix factorization (Bài toán không lỗi thỏa mãn PL)	37
3.2.3	Phân loại đa lớp sử dụng mạng học sâu (Bài toán không lỗi tổng quát)	38
	Kết luận chung	40
	Tài liệu tham khảo	41

Danh mục kí hiệu và chữ viết tắt

v.đ.k	Với điều kiện.
\mathbb{R}	Tập các số thực.
\mathbb{R}^n	Không gian Euclide n chiều.
$\nabla f(x)$	Véc tơ gradient của hàm f tại điểm x .
$\ \cdot\ $	Chuẩn Euclide trong không gian \mathbb{R}^n .
$\mathbb{R}^{m \times n}$	Tập các ma trận m hàng n cột.
SGD	Stochastic Gradient Descent - Thuật toán hướng giảm gradient ngẫu nhiên.
SPS	Stochastic Polyak Step-size - Cỡ bước Polyak ngẫu nhiên.
f^*	Giá trị tối ưu của hàm f .
f_i^*	Giá trị tối ưu của hàm f_i .

Danh sách hình vẽ

1.1	Minh họa cho bất đẳng thức $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$. . .	9
1.2	Một số hàm loss cơ bản.	10
3.1	Kết quả so sánh SGD với cỡ bước SPS_{\max} và cỡ bước hằng cho bài toán phân loại nhị phân có và không sử dụng hàm mất mát hồi quy logistic L2.	35
3.2	Kí hiệu các phương pháp thích nghi.	36
3.3	Thử nghiệm phân loại nhị phân trên bộ dữ liệu mushrooms và ijcnn với kernel RBF (train loss).	36
3.4	Nhân tử hóa ma trận với $k = 4$ và $k = 10$	38
3.5	Kết quả chạy trên bộ dữ liệu CIFAR10 với mô hình ResNet34.	38

Mở đầu

Thuật toán hướng giảm gradient ngẫu nhiên (SGD) là một công cụ quan trọng trong việc giải quyết các bài toán tối ưu lớn trong học máy và trí tuệ nhân tạo. Tuy nhiên, một thách thức lớn trong việc triển khai SGD là lựa chọn cỡ bước phù hợp để đảm bảo tốc độ hội tụ và sự chính xác của thuật toán. Nhiều phương pháp đã được áp dụng để cải thiện cỡ bước trong SGD như cỡ bước hằng, cỡ bước giảm dần hay các phương pháp Adam và RMSProp... Dù các phương pháp thích nghi này mang lại hiệu quả cao nhưng lại đòi hỏi tính chỉnh siêu tham số phức tạp.

Từ nhược điểm trên chúng ta nhớ tới một trong những cỡ bước thích nghi đầu tiên với đặc điểm là tự điều chỉnh dựa trên thông tin hàm mất mát, hội tụ nhanh chóng và không yêu cầu nhiều tham số phức tạp – cỡ bước Polyak (1987) [1]. Cỡ bước này chỉ yêu cầu biết trước giá trị tối ưu, song đây không phải vấn đề vì giá trị tối ưu của hàm mất mát trong lĩnh vực học máy và trí tuệ nhân tạo đều được kì vọng bằng 0.

Với dữ liệu đầu vào ngày một lớn của mô hình học máy hiện đại, trong khi cỡ bước Polyak mỗi lần cập nhật cần tính toán gradient trên toàn bộ tập dữ liệu sẽ khiến gia tăng đáng kể thời gian xử lý và nhu cầu về tài nguyên. Để khắc phục hạn chế này và nâng cao hiệu quả khi kết hợp với thuật toán SGD, một biến thể hiệu quả hơn của cỡ bước Polyak - cỡ bước Polyak ngẫu nhiên - đã được đề xuất bởi nhóm tác giả Nicolas Loizou, Sharan Vaswani, Issam Laradji và Simon Lacoste-Julien [8]. Và đây cũng chính là nội dung em sẽ tìm hiểu và phân tích trong đề án này.

Nội dung chính của đề án bao gồm ba chương:

- **Chương 1:** Trình bày các kiến thức chuẩn bị như lý thuyết hàm lồi, điều kiện PL và thuật toán hướng giảm gradient...
- **Chương 2:** Tập trung vào thuật toán SGD với cỡ bước Polyak ngẫu nhiên và phân tích sự hội tụ của nó trong các trường hợp hàm mục tiêu khác nhau.
- **Chương 3:** Thử nghiệm và đánh giá hiệu quả của thuật toán trên các mô hình cụ thể.

Chương 1

Kiến thức chuẩn bị

Mở đầu đề án ta tìm hiểu về kiến thức nền tảng cần sử dụng để thực hiện việc nghiên cứu bài toán hướng giảm gradient ngẫu nhiên với cỡ bước Polyak. Nguồn tham khảo trong chương này tới từ các tài liệu [2], [3], [4], [7].

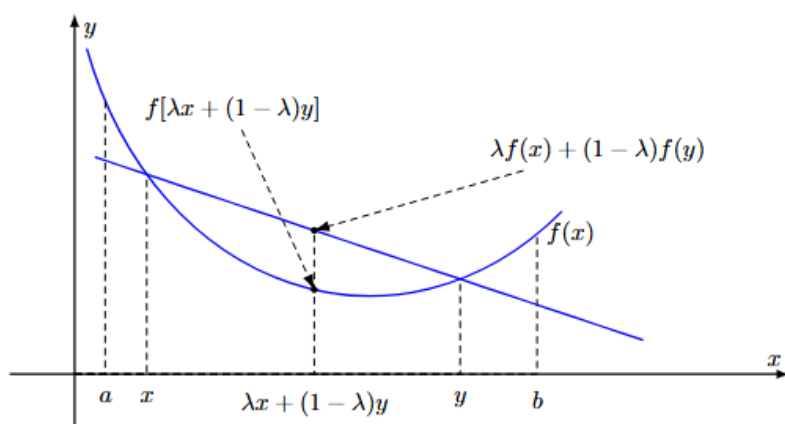
1.1 Hàm lồi và một số tính chất

1.1.1 Hàm lồi và một số tính chất

Định nghĩa 1. (Hàm lồi) Một hàm $f : \mathbb{R}^n \rightarrow \mathbb{R}$ được gọi là lồi nếu

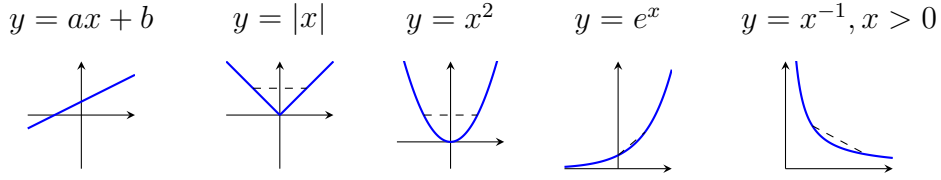
$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad \forall x, y \in \mathbb{R}^n, \lambda \in [0, 1]. \quad (1)$$

Bất đẳng thức (1) được minh họa trong Hình 1.1.



Hình 1.1: Minh họa cho bất đẳng thức $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$.

Ví dụ: Một số hàm lồi cơ bản:



Hình 1.2: Một số hàm lồi cơ bản.

Định nghĩa 2. (Hàm lồi chặt) Một hàm $f : \mathbb{R}^n \rightarrow \mathbb{R}$ được gọi là lồi chặt nếu

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y) \quad \forall x, y \in \mathbb{R}^n, x \neq y, \lambda \in (0, 1).$$

Một số tính chất

Tính chất 1. Nếu $f(x)$ là hàm lồi thì $af(x)$ cũng là hàm lồi với $a > 0$.

Tính chất 2. Tổng của hai hàm lồi là một hàm lồi, với tập xác định là giao của hai tập xác định kia.

Tính chất 3. Cho f là hàm khả vi trên tập lồi mở $X \subseteq \mathbb{R}^n$. Hàm f là hàm lồi trên X khi và chỉ khi:

$$f(y) - f(x) \geq \langle \nabla f(x), y - x \rangle \quad \forall x, y \in X. \quad (2)$$

1.1.2 Hàm lồi mạnh

Định nghĩa 3. Hàm $f : \mathbb{R}^n \rightarrow \mathbb{R}$ được gọi là lồi mạnh μ , nếu tồn tại một hằng số $\mu > 0$ sao cho với mọi $x, y \in \mathbb{R}^n$:

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2. \quad (3)$$

Nếu bất đẳng thức (3) thỏa mãn với $\mu = 0$, thì hàm f là hàm lồi.

Từ (3) ta thu được tính chất:

$$f(x^*) \geq f(x) - \frac{1}{2\mu} \|\nabla f(x)\|^2.$$

Với x^* là nghiệm tối ưu của $f(x)$. Chứng minh. Vì f là hàm lồi mạnh μ , ta có

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2.$$

Tiếp theo, thực hiện tối thiểu hóa bất đẳng thức theo y . Việc tối thiểu hóa sẽ bảo toàn quan hệ bất đẳng thức, do đó chúng ta sẽ tập trung vào từng vế riêng biệt. Với vế trái, ta có:

$$\min_{y \in \mathbb{R}^n} \{f(y)\} = f(x^*).$$

Để giải vế phải, chúng ta lấy đạo hàm theo y thu được

$$\nabla f(x) + \mu(y - x) = 0 \iff y = x - \frac{1}{\mu} \nabla f(x).$$

Thay giá trị tối ưu y mà chúng ta tìm được vào vế phải, ta được

$$f(x) - \frac{1}{\mu} \|\nabla f(x)\|^2 + \frac{1}{2\mu} \|\nabla f(x)\|^2.$$

Và cuối cùng, thu được:

$$f(x^*) \geq f(x) - \frac{1}{2\mu} \|\nabla f(x)\|^2.$$

Như vậy, ta có điều phải chứng minh.

1.2 Hàm L -trơn

Định nghĩa 4. Cho $f : \mathbb{R}^n \rightarrow \mathbb{R}$, và $L > 0$. Ta nói rằng f là L -trơn (L -smooth) nếu f khả vi và ∇f là L -Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{với mọi } x, y \in \mathbb{R}^n. \quad (5)$$

Hằng số L được gọi là độ trơn của f .

Đối với hàm f là L -trơn tổng quát ta có hai bổ đề sau:

Bổ đề 1. Nếu $f : \mathbb{R}^n \rightarrow \mathbb{R}$ là L -trơn thì:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|^2 \quad \text{với mọi } x, y \in \mathbb{R}^n. \quad (6)$$

Chứng minh. Theo định lý cơ bản của giải tích,

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x + t(y - x)), y - x \rangle dt.$$

Do đó,

$$f(y) - f(x) = \langle \nabla f(x), y - x \rangle + \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt.$$

Vậy nên,

$$\begin{aligned}
|f(y) - f(x) - \langle \nabla f(x), y - x \rangle| &= \left| \int_0^1 \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle dt \right| \\
&\leq \int_0^1 |\langle \nabla f(x + t(y - x)) - \nabla f(x), y - x \rangle| dt \\
&\leq \int_0^1 \|\nabla f(x + t(y - x)) - \nabla f(x)\| \cdot \|y - x\| dt \quad (*) \\
&\leq \int_0^1 tL\|y - x\|^2 dt \\
&= \frac{L}{2}\|y - x\|^2,
\end{aligned}$$

trong đó ở (*) chúng ta đã sử dụng bất đẳng thức Cauchy–Schwarz tổng quát.

Chú ý: Từ (3) và (6), ta nhận thấy rằng nếu một hàm là L -trơn và μ -lồi mạnh thì $\mu \leq L$.

Bổ đề 2. Nếu f là L -trơn và $\lambda > 0$, thì:

$$f(x - \lambda \nabla f(x)) - f(x) \leq -\lambda \left(1 - \frac{\lambda L}{2}\right) \|\nabla f(x)\|^2. \quad (7)$$

Nếu thêm vào f bị chặn dưới thì:

$$\frac{1}{2L} \|\nabla f(x)\|^2 \leq f(x) - \inf_{\mathbb{R}^n} f. \quad (8)$$

Chứng minh.

Bất đẳng thức đầu tiên (7) được suy ra bằng cách thế $y = x - \lambda \nabla f(x)$ vào (6),

vì:

$$\begin{aligned}
f(x - \lambda \nabla f(x)) &\leq f(x) - \lambda \langle \nabla f(x), \nabla f(x) \rangle + \frac{L}{2} \|\lambda \nabla f(x)\|^2, \\
&= f(x) - \lambda \left(1 - \frac{\lambda L}{2}\right) \|\nabla f(x)\|^2.
\end{aligned}$$

Giả sử thêm rằng f bị chặn dưới. Bằng cách sử dụng (7) với $\lambda = 1/L$, ta được :

$$\inf_{\mathbb{R}^n} f - f(x) \leq f\left(x - \frac{1}{L} \nabla f(x)\right) - f(x) \leq -\frac{1}{2L} \|\nabla f(x)\|^2.$$

Nhân hai vế của bất đẳng thức trên (bất đẳng thức sau) với -1 ta được (8).

Đối với hàm f lồi và L -trơn ta có hai bổ đề sau:

Bổ đề 3. Nếu $f : \mathbb{R}^n \rightarrow \mathbb{R}$ là hàm lồi và L -trơn, thì với mọi $x, y \in \mathbb{R}^n$, ta có:

$$\frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle, \quad (9)$$

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|^2 \leq \langle \nabla f(y) - \nabla f(x), y - x \rangle. \quad (Co-coercivity) \quad (10)$$

Chứng minh. Để chứng minh (9), cho $x, y \in \mathbb{R}^n$, áp dụng tính lồi và L -trơn của f , với mọi $z \in \mathbb{R}^n$:

$$\begin{aligned} f(x) - f(y) &= f(x) - f(z) + f(z) - f(y) \\ &\stackrel{(2)+(6)}{\leq} \langle \nabla f(x), x - z \rangle + \langle \nabla f(y), z - y \rangle + \frac{L}{2} \|z - y\|^2. \end{aligned}$$

Để có được chặn trên chặt nhất ở vế phải, ta có thể tối thiểu hóa vế phải theo z , ta được:

$$z = y - \frac{1}{L}(\nabla f(y) - \nabla f(x)).$$

Thay thế z và sắp xếp lại các số hạng:

$$\begin{aligned} f(x) - f(y) &\leq \langle \nabla f(x), x - z \rangle + \langle \nabla f(y), z - y \rangle + \frac{L}{2} \|z - y\|^2 \\ &= \langle \nabla f(x), x - y \rangle - \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2 + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \\ &= \langle \nabla f(x), x - y \rangle - \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2. \end{aligned}$$

Như vậy ta chứng minh được (9). Áp dụng (9) hai lần bằng cách hoán đổi vai trò của x và y :

$$\begin{aligned} \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 &\leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle, \\ \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2 &\leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle, \end{aligned}$$

sau đó cộng hai bất đẳng thức này ta thu được (10).

1.3 Thuật toán hướng giảm

Xét bài toán quy hoạch không ràng buộc

$$\min f(x) \quad \text{v.đ.k. } x \in \mathbb{R}^n, \quad (P^{krb})$$

trong đó $f : \mathbb{R}^n \rightarrow \mathbb{R}$ là hàm phi tuyến, khả vi trên \mathbb{R}^n .

Theo Định lý về Điều kiện bậc nhất (Định lý 6.1, trang 221) [4], nếu f là hàm khả vi trên \mathbb{R}^n thì điều kiện cần của nghiệm cực tiểu địa phương x^* là $\nabla f(x^*) = 0$. Hệ phương trình $\nabla f(x) = 0$ có n ẩn, n phương trình. Tuy nhiên, ngoài trừ một số trường hợp đơn giản, nói chung việc giải trực tiếp hệ này là khó thực hiện được. Ví

dự xét hàm một biến $f(x) = e^x + x^2 + \cos x$ và ta không có công thức tính nghiệm của phương trình $f'(x) = e^x + 2x - \sin x = 0$.

Ý tưởng cơ bản của *phương pháp hướng giảm* (descent method) để giải bài toán (P^{krb}) với hàm mục tiêu f khả vi là: Xuất phát một điểm bất kỳ $x^0 \in \mathbb{R}^n$, ta xây dựng một dãy điểm $x^1, x^2, \dots, x^k, \dots$ sao cho

$$f(x^0) \geq f(x^1) \geq f(x^2) \geq \dots$$

và dãy $\{x^k\}$ hội tụ đến điểm dừng $x^* \in \mathbb{R}^n$ của hàm f , tức $\nabla f(x^*) = 0$. Trường hợp f là hàm lồi, thì điểm x^* cũng chính là nghiệm cực tiểu toàn cục của bài toán (P^{krb}) .

Các thuật toán lặp mà chúng ta sẽ xem xét có dạng:

$$x^{k+1} = x^k + t_k d^k, \quad k = 0, 1, 2, \dots,$$

trong đó d^k được gọi là hướng giảm của f tại x^k và t_k là độ dài bước. Chúng ta sẽ định nghĩa các hướng giảm như sau:

Định nghĩa 5. (*Hướng giảm*). Giả sử f là một hàm khả vi liên tục trên \mathbb{R}^n , và $x \in \mathbb{R}^n$; d là một hướng giảm của f tại x . Khi đó tồn tại $\varepsilon > 0$ sao cho

$$f(x + td) < f(x)$$

với mọi $t \in (0, \varepsilon]$.

Chứng minh. Vì $f'(x; d) < 0$, từ định nghĩa của đạo hàm theo hướng, ta có

$$f'(x^0, d) = \lim_{t \rightarrow 0^+} \frac{f(x^0 + td) - f(x^0)}{t} = \langle \nabla f(x^0), d \rangle < 0.$$

Do đó, $f(x^0 + td) - f(x^0) < 0$ với t đủ nhỏ. Mệnh đề đã được chứng minh.

Dưới đây là lược đồ chung cho thuật toán hướng giảm:

Lược đồ chung thuật toán hướng giảm

Khởi tạo: Chọn $x^0 \in \mathbb{R}^n$ tùy ý. Gán $k := 0$.

Bước lặp k: Với bất kỳ $k = 0, 1, 2, \dots$, thực hiện

- (k_1) **If** x^k thỏa mãn điều kiện dừng **Then** Dừng thuật toán
 Else xác định $x^{k+1} := x^k + t_k d^k$ sao cho $f(x^{k+1}) < f(x^k)$,
 trong đó $t_k > 0$ là cỡ bước tại lần lặp k .
- (k_2) Gán $k := k + 1$; Quay lại Bước lặp k .

Trong lược đồ trên, điều kiện dừng của thuật toán tại Bước (k_1) thường là

$$\nabla f(x^k) \approx 0 \quad \text{hoặc} \quad \|x^k - x^{k-1}\| \text{ đủ nhỏ.}$$

1.4 Điều kiện Polyak-Lojasiewicz (PL)

Nội dung phần này được tham khảo tại [5].

Ban đầu, người ta chứng minh được khi hàm f lồi mạnh, hướng giảm gradient trong bài toán tối ưu đạt tốc độ hội tụ tuyến tính toàn cục (Nesterov, 2004). Tuy nhiên, nhiều mô hình cơ bản như least square và logistic regression có hàm mục tiêu là lồi nhưng không lồi mạnh. Thêm nữa, nếu f chỉ lồi thì hướng giảm gradient chỉ đạt được tốc độ dưới tuyến tính. Chính điều này đã thúc đẩy việc nghiên cứu một số điều kiện để khi thỏa mãn những điều kiện này bài toán vẫn có thể đạt được tốc độ hội tụ tuyến tính. Trong những điều kiện đã được công bố, có một điều kiện khá yếu (cần ít tiêu chuẩn) sẽ được sử dụng ở trong đề cáo này là điều kiện Polyak-Lojasiewicz (PL). Điều kiện này yếu đến mức chỉ cần thỏa mãn L -trơn, không cần lồi mạnh và thậm chí là không lồi. Dưới đây là nội dung điều kiện:

Định nghĩa 6. Hàm $f : \mathbb{R}^n \rightarrow \mathbb{R}$ thỏa mãn điều kiện Polyak-Lojasiewicz (PL) nếu tồn tại một hằng số $\mu > 0$ sao cho với mọi $x \in \mathbb{R}^n$:

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f^*). \quad (4)$$

Bất đẳng thức này đơn giản yêu cầu rằng gradient của hàm tăng nhanh hơn

một hàm bậc hai khi chúng ta di chuyển ra xa giá trị hàm tối ưu. Lưu ý, bất đẳng thức này ngụ ý rằng mọi điểm dừng là cực tiểu toàn cục.

Tốc độ hội tụ tuyến tính của phương pháp hướng giảm gradient dưới các giả định này đã được chứng minh đầu tiên bởi Polyak (1963). Dưới đây là một chứng minh đơn giản của kết quả này khi sử dụng cỡ bước hằng $\frac{1}{L}$.

Chứng minh. Xét bài toán tối ưu hóa trơn không ràng buộc cơ bản:

$$\min_{x \in \mathbb{R}^n} f(x),$$

với f thỏa mãn bất đẳng thức PL và ∇f là L -Lipschitz liên tục,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2.$$

Áp dụng hướng giảm gradient (GD) với cỡ bước hằng $\frac{1}{L}$,

$$x^{k+1} = x^k - \frac{1}{L} \nabla f(x^k),$$

ta có

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2.$$

Thay $x^{k+1} - x^k = -\frac{1}{L} \nabla f(x^k)$ vào biểu thức, ta được

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2L} \|\nabla f(x^k)\|^2.$$

Áp dụng bất đẳng thức PL, ta có

$$f(x^{k+1}) \leq f(x^k) - \frac{\mu}{L} [f(x^k) - f^*].$$

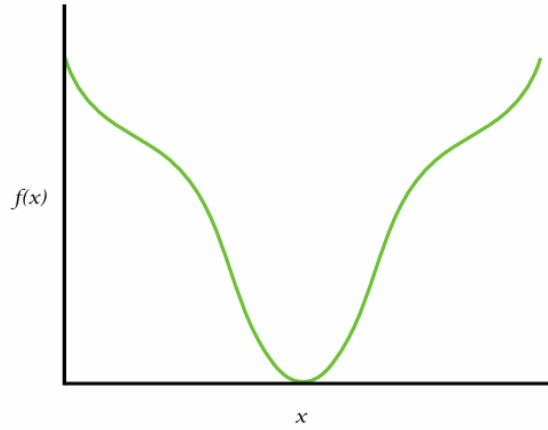
Trừ f^* và áp dụng đệ quy cho thấy tốc độ hội tụ tuyến tính toàn cục:

$$f(x^k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k [f(x^0) - f^*].$$

Như vậy, ta thu được điều phải chứng minh.

Ví dụ: Một hàm không lồi nhưng thỏa mãn bất đẳng thức PL:

$$f(x) = x^2 + 3 \sin^2(x)$$



Chứng minh hàm $f(x) = x^2 + 3 \sin^2(x)$ thỏa mãn điều kiện Polyak-Łojasiewicz (PL):

Gradient của $f(x)$ là

$$\nabla f(x) = 2x + 3 \sin(2x).$$

Giá trị tối ưu toàn cục f^* đạt được khi $x = 0$, do đó $f^* = 0$.

Xét bất đẳng thức PL:

$$\frac{1}{2} \|\nabla f(x)\|^2 = \frac{1}{2} (2x + 3 \sin(2x))^2.$$

Ta cần chứng minh rằng tồn tại $\mu > 0$ sao cho:

$$\frac{1}{2} (2x + 3 \sin(2x))^2 \geq \mu (x^2 + 3 \sin^2(x)).$$

Do $\sin^2(x)$ và $\sin(2x)$ bị chặn (≤ 1) nên điều kiện trên có thể được thỏa mãn với $\mu > 0$ đủ nhỏ. Bằng cách tìm chặn dưới cho tỷ lệ giữa hai vế (vế trái chia vế phải), ta có thể khẳng định hàm $f(x)$ thỏa mãn điều kiện PL.

1.5 Mô hình quá tham số trong học sâu

Định nghĩa:

Mô hình quá tham số là mô hình có số lượng tham số lớn hơn đáng kể so với lượng dữ liệu huấn luyện hoặc độ phức tạp của bài toán. Trong trường hợp này, mô hình có khả năng học chính xác dữ liệu huấn luyện, bao gồm cả nhiễu, nhưng thường

gặp khó khăn trong việc tổng quát hóa trên dữ liệu kiểm tra. Hiện tượng này thường xảy ra trong các mạng học sâu với số tầng lớn hoặc số lượng neuron cao.

Điều kiện xác định tính quá tham số:

1. **Số lượng tham số lớn hơn số lượng dữ liệu:** Nếu số lượng tham số $|\theta|$ trong mô hình vượt quá số lượng mẫu dữ liệu N :

$$|\theta| > N$$

Ví dụ: Một mạng neuron với hàng triệu tham số được huấn luyện trên vài nghìn mẫu dữ liệu.

2. **Hàm mất mát có nhiều nghiệm:** Nếu hàm mất mát $L(f(x; \theta), y)$ có nhiều nghiệm tối ưu hoặc vô số giá trị tham số θ thỏa mãn, điều này cho thấy mô hình bị quá tham số.
3. **Dấu hiệu từ kết quả huấn luyện:** Hiện tượng mất mát rất thấp trên tập huấn luyện nhưng lỗi kiểm tra cao (quá khớp - overfitting). Mô hình có khả năng ghi nhớ toàn bộ dữ liệu huấn luyện, kể cả nhiễu, mà không thể tổng quát hóa.

Mặc dù mô hình quá tham số có nguy cơ dẫn đến quá khớp, tuy nhiên một số mô hình học sâu quá tham số như GPT, ResNet, hoặc BERT vẫn có thể đạt hiệu suất cao nhờ các kỹ thuật tối ưu hóa và regularization.

Kết luận

Như vậy, chương này đã cung cấp nền tảng lý thuyết cần thiết như hàm lỗi, điều kiện Polyak-Lojasiewicz (PL), thuật toán hướng giảm, hàm L -trơn ... Những khái niệm này làm cơ sở cho việc hiểu và áp dụng cơ bước Polyak ngẫu nhiên trong các thuật toán tối ưu.

Chương 2

Thuật toán hướng giảm Gradient ngẫu nhiên với cỡ bước Polyak

Trong chương này, ý tưởng về các phương pháp được tham khảo từ N. Loizou, S. Vaswani, I. Laradji, and S. Lacoste-Julien [8]. Trong đề án này, chúng ta giải quyết bài toán tối ưu tổng hữu hạn:

$$\min_{x \in \mathbb{R}^d} \left[f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x) \right]. \quad (2.1)$$

Trong đó,

- n : số điểm huấn luyện.
- d : số chiều của tham số mô hình.
- x : tham số của mô hình.
- $f_i(x)$: hàm mất mát trên điểm dữ liệu huấn luyện i .
- $f(x)$: hàm mất mát trung bình.

Mục tiêu của bài toán là tối ưu hàm mất mát trung bình $f(x)$.

Kí hiệu:

- x^* : nghiệm tối ưu.
- $X^* \subset \mathbb{R}^d$: tập hợp các nghiệm tối ưu x^* .
- $f_i^* := \inf_x f_i(x)$: giá trị tối ưu tại $f_i(x)$.

2.1 Thuật toán hướng giảm gradient ngẫu nhiên (SGD)

Ta đi giải bài toán tối ưu (2.1) bằng phương pháp SGD:

Lược đồ thuật toán hướng giảm gradient ngẫu nhiên

Khởi tạo: Chọn $x^0 \in \mathbb{R}^n$ tùy ý. Gán $k := 0$.

Bước lặp k : Với $k = 0, 1, 2, \dots$, thực hiện:

(k_1) Nếu x^k thỏa mãn điều kiện dừng, **Kết thúc thuật toán**.

(k_2) Ngẫu nhiên chọn $i \in \{1, 2, \dots, n\}$. Tính gradient của f tại điểm i và cập nhật:

$$x^{k+1} = x^k - \gamma_k \nabla f_i(x^k), \quad (\text{SGD})$$

trong đó $\gamma_k > 0$ là cỡ bước tại lần lặp k .

(k_3) Đặt $k := k + 1$. Quay lại Bước lặp k .

2.2 Cỡ bước Polyak ngẫu nhiên

2.2.1 Cỡ bước Polyak tất định

Đối với các hàm lồi, cỡ bước Polyak xác định tại lần lặp k là giá trị tối thiểu hóa một cận trên $Q(\gamma)$ đối với khoảng cách của điểm lặp x_{k+1} tới nghiệm tối ưu:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - \gamma_k \nabla f(x^k) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f(x^k) \rangle + \gamma_k^2 \|\nabla f(x^k)\|^2 \\ &\leq \|x^k - x^*\|^2 - 2\gamma_k [f(x^k) - f(x^*)] + \gamma_k^2 \|\nabla f(x^k)\|^2 = Q(\gamma), \end{aligned}$$

trong đó dòng cuối cùng có được từ tính chất của hàm lồi:

$$f(x^*) \geq f(x^k) + \langle x^k - x^*, \nabla f(x^k) \rangle.$$

Tối ưu $Q(\gamma)$ ta thu được cỡ bước được đề xuất bởi Polyak (1987),

$$\gamma_k = \arg \min_{\gamma} Q(\gamma) = \frac{f(x^k) - f^*}{\|\nabla f(x^k)\|^2}. \quad (2.2)$$

Lưu ý rằng cỡ bước trên chỉ có thể sử dụng khi giá trị tối ưu f^* là đã biết.

Phân tích hội tụ.

Bằng cách tiếp tục phân tích khoảng cách của x^{k+1} tới nghiệm tối ưu:

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - 2\gamma_k [f(x^k) - f(x^*)] + \gamma_k^2 \|\nabla f(x^k)\|^2 \\ &= \|x^k - x^*\|^2 - \frac{[f(x^k) - f(x^*)]^2}{\|\nabla f(x^k)\|^2}.\end{aligned}\quad (\text{Thay (2.2)})$$

Từ trên, lưu ý rằng $\|x^k - x^*\|^2$ là một dãy đơn điệu. Bây giờ, sử dụng đệ quy và giả sử $\|\nabla f(x^k)\|^2 < G^2$, G là hằng số, ta có:

$$\|x^{k+1} - x^*\|^2 \leq \|x^0 - x^*\|^2 - \frac{1}{G^2} \sum_{i=0}^k [f(x^i) - f(x^*)]^2.$$

Do đó,

$$\frac{1}{G^2} \sum_{i=0}^k [f(x^i) - f(x^*)]^2 \leq \|x^0 - x^*\|^2 - \|x^{k+1} - x^*\|^2 \leq \|x^0 - x^*\|^2.$$

Gọi $f_*^k = \min\{f(x^i) : i = 0, 1, \dots, k\}$. Khi đó: $[f_*^k - f(x^*)]^2 \leq \frac{G^2 \|x^0 - x^*\|^2}{k+1}$ và

$$f_*^k - f(x^*) \leq \frac{G \|x^0 - x^*\|}{\sqrt{k+1}} = O\left(\frac{1}{\sqrt{k}}\right).$$

2.2.2 Cỡ bước Polyak ngẫu nhiên

Dễ thấy rằng việc sử dụng cỡ bước Polyak tấp định vào bài toán SGD là không thực tế. Nó yêu cầu tính toán giá trị hàm f và gradient đầy đủ của nó trong mỗi lần lặp.

Để tránh điều này, nhóm tác giả [8] đề xuất cỡ bước Polyak ngẫu nhiên (SPS) cho SGD:

$$\text{SPS: } \gamma_k = \frac{f_i(x^k) - f_i^*}{c \|\nabla f_i(x^k)\|^2}. \quad (2.3)$$

Điều đặc biệt của cỡ bước này là không yêu cầu 2 tham số L -trơn hay μ -lồi mạnh.

SPS chỉ yêu cầu tính toán gradient ngẫu nhiên $\nabla f_i(x^k)$ và giá trị của hàm $f_i(x^k)$ tại điểm lặp hiện tại (các đại lượng này có thể được tính toán trong quy tắc cập nhật của SGD mà không tốn thêm chi phí). Tuy nhiên, nó yêu cầu biết giá trị f_i^* .

Một đại lượng quan trọng trong cỡ bước này là tham số $0 < c \in \mathbb{R}$, có thể được đặt lý thuyết dựa trên các thuộc tính của hàm đang nghiên cứu. Ví dụ, đối với các hàm lồi mạnh, nên chọn $c = 1/2$ để hội tụ tối ưu; đối với các hàm lồi và không lồi, chọn $c = 1$.

Ngoài SPS, trong một số kết quả hội tụ, nhóm tác giả đã đưa ra một cỡ bước có cận trên:

$$\text{SPS}_{\max} : \quad \gamma_k = \min \left\{ \frac{f_i(x^k) - f_i^*}{c \|\nabla f_i(x^k)\|^2}, \gamma_b \right\}, \quad (2.4)$$

trong đó $\gamma_b > 0$ là một giới hạn để ngăn SPS trở nên quá lớn và là điều cần thiết để đảm bảo hội tụ đến một vùng nhỏ xung quanh nghiệm. Nếu $\gamma_b = \infty$ thì SPS_{\max} tương đương với SPS.

2.2.3 Sự chênh lệch mục tiêu tối ưu

Nội dung ở mục này được tham khảo trong [6] và [7].

Nhận xét 1 (Ước lượng không chệch của gradient). *Một đặc điểm quan trọng của thuật toán **SGD** là tại mỗi bước lặp, ta sử dụng hướng $-\nabla f_i(x^k)$, đây là một ước lượng không chệch của $-\nabla f(x^k)$. Thật vậy, vì:*

$$\mathbb{E}[\nabla f_i(x^k) \mid x^k] = \sum_{i=1}^n \frac{1}{n} \nabla f_i(x^k) = \nabla f(x^k).$$

Xét giả định một chênh lệch mục tiêu tối ưu hữu hạn:

Giả định 2.1 (Sự chênh lệch mục tiêu tối ưu hữu hạn).

$$\sigma^2 := \mathbb{E}_i[f_i(x^*) - f_i^*] = f(x^*) - \mathbb{E}_i[f_i^*] < \infty. \quad (2.5)$$

Đây là một giả định rất yếu. Cụ thể:

1. **Không yêu cầu hàm đạt cực tiểu tại x^* :** Giả định chỉ yêu cầu giá trị $f_i(x^*) - f_i^*$ hữu hạn trung bình, thay vì yêu cầu mỗi hàm f_i đạt cực tiểu tại cùng một điểm x^* . Điều này làm cho giả định này linh hoạt hơn, vì không phải mọi hàm f_i đều có thể đạt cực tiểu tại cùng một điểm x^* .
2. **Không yêu cầu nhiễu gradient hữu hạn:** Trong nhiều phân tích SGD khác, người ta thường giả định rằng nhiễu gradient $z^2 := \mathbb{E}[\|\nabla f_i(x^*)\|^2]$ phải hữu hạn,

điều này có thể không thực tế cho một số hàm mất mát hoặc cấu trúc dữ liệu. Ở đây, giả định chỉ yêu cầu chênh lệch mục tiêu tối ưu là hữu hạn, một điều kiện ít hạn chế hơn so với việc yêu cầu gradient phải hữu hạn.

3. Phù hợp với các mô hình quá tham số: Trong các mô hình quá tham số như mạng nơ-ron sâu hoặc trong các bài toán mà dữ liệu có thể phân tách tuyến tính, giá trị $f_i(x^*) - f_i^*$ có thể bằng 0, làm cho $\sigma = 0$, nghĩa là giả định này tự động thỏa mãn mà không cần thêm điều kiện nào.

Vì những lý do trên, giả định này dễ thỏa mãn và áp dụng rộng rãi hơn, đặc biệt trong các trường hợp mà các giả định khắt khe khác có thể không phù hợp. Và một chú ý rằng hàm nội suy khi $\sigma = 0$.

2.3 Phân tích hội tụ của thuật toán gradient ngẫu nhiên với cỡ bước Polyak

2.3.1 Giới hạn trên và dưới của cỡ bước Polyak ngẫu nhiên

Nếu một hàm g là lồi mạnh với hệ số μ và L -trơn (L-smooth), các giới hạn sau sẽ được thỏa mãn:

$$\frac{1}{2L} \|\nabla g(x)\|^2 \leq g(x) - \inf_x g(x) \leq \frac{1}{2\mu} \|\nabla g(x)\|^2.$$

Sử dụng các giới hạn này và giả sử rằng các hàm f_i trong bài toán (2.1) là lồi mạnh với hệ số μ_i và L_i -trơn, ta dễ dàng thấy rằng SPS có thể bị giới hạn trên và dưới như sau:

$$\frac{1}{2cL_{\max}} \leq \frac{1}{2cL_i} \leq \gamma_k = \frac{f_i(x^k) - f_i^*}{c\|\nabla f_i(x^k)\|^2} \leq \frac{1}{2c\mu_i}, \quad (2.6)$$

trong đó $L_{\max} = \max\{L_i\}_{i=1}^n$.

Nếu $\gamma_b < \frac{1}{2cL_{\max}} \implies$ SGD với cỡ bước hằng γ_b .

2.3.2 Hàm mục tiêu f là tổng của các hàm lồi và f lồi mạnh

Trong phần này, ta giả định rằng tất cả các thành phần f_i đều là các hàm lồi với ít nhất một trong số chúng là lồi mạnh và hàm mục tiêu f là lồi mạnh với hệ số μ .

Định lý 1. Giả sử f_i là các hàm lồi, L_i -trơn và giả định rằng hàm mục tiêu f là hàm lồi mạnh với hệ số μ . Khi đó, SGD với SPS_{\max} với $c \geq 1/2$ hội tụ như sau:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \mu\alpha)^k \|x^0 - x^*\|^2 + \frac{2\gamma_b\sigma^2}{\mu\alpha}, \quad (2.7)$$

trong đó $\alpha := \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}$ và $L_{\max} = \max\{L_i\}_{i=1}^n$ là hằng số trơn tối đa. Tốc độ hội tụ tốt nhất và vùng lân cận chặt nhất đạt được khi $c = 1/2$.

Chứng minh.

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - \gamma_k \nabla f_i(x^k) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_i(x^k) \rangle + \gamma_k^2 \|\nabla f_i(x^k)\|^2 \\ &\stackrel{(2.4)}{\leq} \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_i(x^k) \rangle + \frac{\gamma_k}{c} [f_i(x^k) - f_i^*] \\ &\stackrel{c \geq 1/2}{\leq} \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_i(x^k) \rangle + 2\gamma_k [f_i(x^k) - f_i^*] \\ &= \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_i(x^k) \rangle + 2\gamma_k [f_i(x^k) - f_i(x^*) + f_i(x^*) - f_i^*] \\ &= \|x^k - x^*\|^2 + 2\gamma_k [-\langle x^k - x^*, \nabla f_i(x^k) \rangle + f_i(x^k) - f_i(x^*)] + 2\gamma_k [f_i(x^*) - f_i^*]. \end{aligned}$$

Từ tính lồi của hàm f_i ta có $-\langle x^k - x^*, \nabla f_i(x^k) \rangle + f_i(x^k) - f_i(x^*) \leq 0, \forall i \in [n]$. Do đó,

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 + 2\gamma_k \underbrace{[-\langle x^k - x^*, \nabla f_i(x^k) \rangle + f_i(x^k) - f_i(x^*)]}_{\leq 0} + 2\gamma_k \underbrace{[f_i(x^*) - f_i^*]}_{\geq 0} \\ &\stackrel{(2.6), (2.4)}{\leq} \|x^k - x^*\|^2 + 2 \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\} [-\langle x^k - x^*, \nabla f_i(x^k) \rangle + f_i(x^k) - f_i(x^*)] \\ &\quad + 2\gamma_b [f_i(x^*) - f_i^*]. \end{aligned}$$

Lấy kỳ vọng có điều kiện dựa trên x^k

$$\begin{aligned} \mathbb{E}_i \|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 + 2 \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\} [-\langle x^k - x^*, \nabla f(x^k) \rangle + f(x^k) - f(x^*)] \\ &\quad + 2\gamma_b \mathbb{E}_i [f_i(x^*) - f_i^*] \\ &\stackrel{(2.5)}{=} \|x^k - x^*\|^2 + 2 \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\} [-\langle x^k - x^*, \nabla f(x^k) \rangle + f(x^k) - f(x^*)] \\ &\quad + 2\gamma_b \sigma^2. \end{aligned}$$

Từ tính lồi mạnh của hàm mục tiêu f ta có $f(x^k) - f(x^*) - \langle x^k - x^*, \nabla f(x^k) \rangle \leq -\frac{\mu}{2} \|x^k - x^*\|^2$. Do đó, ta thu được:

$$\mathbb{E}_i \|x^{k+1} - x^*\|^2 \leq \left(1 - \mu \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}\right) \|x^k - x^*\|^2 + 2\gamma_b \sigma^2.$$

Lấy kì vọng 1 lần nữa:

$$\mathbb{E}\|x^{k+1} - x^*\|^2 \leq \left(1 - \mu \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}\right) \mathbb{E}\|x^k - x^*\|^2 + 2\gamma_b\sigma^2.$$

Áp dụng đệ quy thu được:

$$\begin{aligned} \mathbb{E}\|x^k - x^*\|^2 &\leq \left(1 - \mu \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}\right)^k \|x^0 - x^*\|^2 \\ &\quad + 2\gamma_b\sigma^2 \sum_{j=0}^{k-1} \left(1 - \mu \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}\right)^j \\ &\leq \left(1 - \mu \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}\right)^k \|x^0 - x^*\|^2 + 2\gamma_b\sigma^2 \frac{1}{\mu \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}}. \end{aligned}$$

Đặt $\alpha = \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}$,

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \mu\alpha)^k \|x^0 - x^*\|^2 + \frac{2\gamma_b\sigma^2}{\mu\alpha}.$$

Từ định nghĩa của α rõ ràng rằng việc có tham số c nhỏ sẽ cải thiện tốc độ hội tụ $1 - \mu\alpha$ và lân cận $\frac{2\gamma_b\sigma^2}{\mu\alpha}$. Vì chúng ta có ràng buộc $c \geq \frac{1}{2}$ nên lựa chọn tối ưu sẽ là $c = \frac{1}{2}$. \square

Hệ quả 1. Giả sử tính chất nội suy ($\sigma = 0$) và tất cả các giả thiết của **Định lý 1** được thỏa mãn. Thuật toán SGD với SPS với $c = 1/2$ hội tụ như sau:

$$\mathbb{E}\|x^k - x^*\|^2 \leq \left(1 - \frac{\mu}{L_{\max}}\right)^k \|x^0 - x^*\|^2.$$

Hệ quả 2. Giả sử tất cả các giả thiết của **Định lý 1** được thỏa mãn. Thuật toán SGD với SPS_{max} với $c = 1/2$ và $\gamma \leq \frac{1}{L_{\max}}$ trở thành SGD với cỡ bước hằng $\gamma \leq \frac{1}{L_{\max}}$ và hội tụ như sau:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \eta\mu)^k \|x^0 - x^*\|^2 + \frac{2\sigma^2}{\mu}.$$

Nếu ta giả sử thêm tính chất nội suy ($\sigma = 0$), các lần lặp của SGD với cỡ bước hằng $\gamma = \frac{1}{L_{\max}}$ thỏa mãn:

$$\mathbb{E}\|x^k - x^*\|^2 \leq (1 - \eta\mu)^k \|x^0 - x^*\|^2.$$

2.3.3 Hàm f là tổng của các hàm lồi

Ở đây, chúng ta xem xét tốc độ hội tụ khi tất cả các hàm thành phần f_i đều lồi nhưng không có tính chất lồi mạnh và thu được định lý sau:

Định lý 2. Giả sử rằng f_i là các hàm lồi, L_i -trơn. Thuật toán SGD với cỡ bước SPS, $c = 1$, hội tụ như sau:

$$\mathbb{E}[f(\bar{x}^k) - f(x^*)] \leq \frac{\|x^0 - x^*\|^2}{\alpha K} + \frac{2\sigma^2\gamma_b}{\alpha}, \quad (2.8)$$

trong đó:

$$\alpha = \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}, \quad \text{và} \quad \bar{x}^k = \frac{1}{K} \sum_{k=0}^{K-1} x^k.$$

Chứng minh.

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - \gamma_k \nabla f_i(x^k) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma_k \langle x^k - x^*, \nabla f_i(x^k) \rangle + \gamma_k^2 \|\nabla f_i(x^k)\|^2 \\ &\stackrel{\text{convexity}}{\leq} \|x^k - x^*\|^2 - 2\gamma_k [f_i(x^k) - f_i(x^*)] + \gamma_k^2 \|\nabla f_i(x^k)\|^2 \\ &\stackrel{(2.4)}{\leq} \|x^k - x^*\|^2 - 2\gamma_k [f_i(x^k) - f_i(x^*)] + \frac{\gamma_k}{c} [f_i(x^k) - f_i^*] \\ &= \|x^k - x^*\|^2 - 2\gamma_k [f_i(x^k) - f_i^* + f_i^* - f_i(x^*)] + \frac{\gamma_k}{c} [f_i(x^k) - f_i^*] \\ &= \|x^k - x^*\|^2 - \gamma_k \left(2 - \frac{1}{c}\right) \underbrace{[f_i(x^k) - f_i^*]}_{>0} + 2\gamma_k \underbrace{[f_i(x^*) - f_i^*]}_{>0}. \end{aligned} \quad (2.3.1)$$

Đặt $\alpha = \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}$ và nhắc lại từ định nghĩa của SPS_{\max} (2.4) ta có:

$$\alpha \stackrel{(2.6), (2.4)}{\leq} \gamma_k \leq \gamma_b. \quad (2.3.2)$$

Từ trên, nếu $\alpha = \frac{1}{2cL_{\max}}$ thì cỡ bước nằm trong khoảng của cỡ bước Polyak ngẫu nhiên (2.6). Trong trường hợp $\alpha = \gamma_b$ thì bài toán trở thành SGD với cỡ bước hằng $\gamma_k = \gamma_b$.

Vì $c > \frac{1}{2}$ nên $(2 - \frac{1}{c}) > 0$. Sử dụng (2.3.2) vào (2.3.1) ta được:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &\leq \|x^k - x^*\|^2 - \gamma_k \left(2 - \frac{1}{c}\right) [f_i(x^k) - f_i^*] + 2\gamma_k [f_i(x^*) - f_i^*] \\ &\stackrel{(2.3.2)}{\leq} \|x^k - x^*\|^2 - \alpha \left(2 - \frac{1}{c}\right) [f_i(x^k) - f_i^*] + 2\gamma_b [f_i(x^*) - f_i^*] \\ &= \|x^k - x^*\|^2 - \alpha \left(2 - \frac{1}{c}\right) [f_i(x^k) - f_i(x^*) + f_i(x^*) - f_i^*] \\ &\quad + 2\gamma_b [f_i(x^*) - f_i^*] \\ &= \|x^k - x^*\|^2 - \alpha \left(2 - \frac{1}{c}\right) [f_i(x^k) - f_i(x^*)] - \alpha \left(2 - \frac{1}{c}\right) [f_i(x^*) - f_i^*] \\ &\quad + 2\gamma_b [f_i(x^*) - f_i^*] \\ &\leq \|x^k - x^*\|^2 - \alpha \left(2 - \frac{1}{c}\right) [f_i(x^k) - f_i(x^*)] + 2\gamma_b [f_i(x^*) - f_i^*]. \end{aligned} \quad (2.3.3)$$

trong đó bất đẳng thức cuối sử dụng $\alpha \left(2 - \frac{1}{c}\right) [f_i(x^*) - f_i^*] > 0$. Sắp xếp lại:

$$\alpha \left(2 - \frac{1}{c}\right) [f_i(x^k) - f_i(x^*)] \leq \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 + 2\gamma_b [f_i(x^*) - f_i^*] \quad (2.3.4)$$

Lấy kì vọng có điều kiện dựa trên x^k và chia 2 vế cho $\alpha \left(2 - \frac{1}{c}\right)$:

$$f(x^k) - f(x^*) \leq \frac{c}{\alpha(2c-1)} (\|x^k - x^*\|^2 - \mathbb{E}_i \|x^{k+1} - x^*\|^2) + 2\gamma_b \frac{c}{\alpha(2c-1)} \sigma^2.$$

Lấy kì vọng 1 lần nữa:

$$\mathbb{E}[f(x^k) - f(x^*)] \leq \frac{c}{\alpha(2c-1)} (\mathbb{E}\|x^k - x^*\|^2 - \mathbb{E}\|x^{k+1} - x^*\|^2) + 2\gamma_b \frac{c}{\alpha(2c-1)} \sigma^2.$$

Tổng từ $k = 0$ to $K - 1$ và chia cho K :

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} [f(x^k) - f(x^*)] &= \frac{c}{\alpha(2c-1)} \frac{1}{K} \sum_{k=0}^{K-1} (\mathbb{E}\|x^k - x^*\|^2 - \mathbb{E}\|x^{k+1} - x^*\|^2) \\ &\quad + \frac{1}{K} \sum_{k=0}^{K-1} \frac{2c\gamma_b\sigma^2}{\alpha(2c-1)} \\ &= \frac{2c}{\alpha(2c-1)} \frac{1}{K} [\|x^0 - x^*\|^2 - \mathbb{E}\|x^K - x^*\|^2] + \frac{2c\gamma_b\sigma^2}{\alpha(2c-1)} \\ &\leq \frac{c}{\alpha(2c-1)} \frac{1}{K} \|x^0 - x^*\|^2 + \frac{2c\gamma_b\sigma^2}{\alpha(2c-1)}. \end{aligned} \quad (2.3.5)$$

Đặt $\bar{x}^K = \frac{1}{K} \sum_{k=0}^{K-1} x^k$ được:

$$\mathbb{E} [f(\bar{x}^K) - f(x^*)] \stackrel{Jensen}{\leq} \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} [f(x^k) - f(x^*)] \leq \frac{c}{\alpha(2c-1)} \frac{1}{K} \|x^0 - x^*\|^2 + \frac{2c\gamma_b\sigma^2}{\alpha(2c-1)}.$$

Với $c = 1$:

$$\mathbb{E} [f(\bar{x}^K) - f(x^*)] \leq \frac{\|x^0 - x^*\|^2}{\alpha K} + \frac{2\gamma_b\sigma^2}{\alpha}, \quad (2.3.6)$$

việc chứng minh hoàn thành. \square

Tương tự như trường hợp lỗi mạnh, kích thước của vùng lân cận tỷ lệ thuận với γ_b . Khi điều kiện nội suy được thỏa mãn và $\sigma = 0$, chúng ta quan sát rằng biến thể của SPS với $\gamma_b = \infty$ hội tụ đến nghiệm tối ưu với tốc độ $O(1/K)$. Trong trường hợp lỗi mạnh, bằng cách đặt $\gamma_b \leq \frac{1}{2L_{\max}}$, chúng ta thu được tốc độ hội tụ đạt được bởi SGD với cỡ bước hằng.

2.3.4 Hàm f là tổng các hàm không lồi - f thỏa mãn điều kiện PL

Trước tiên, chúng ta tập trung vào một lớp đặc biệt của các hàm không lồi thỏa mãn điều kiện Polyak-Lojasiewicz (PL) (được giới thiệu đầu tiên bởi Polyak, 1987) mà em đã trình bày ở phần kiến thức chuẩn bị. Giả định rằng hàm f thỏa mãn điều kiện PL nhưng không giả định tính lồi của các hàm thành phần f_i .

Định lý 3. *Giả sử rằng hàm f thỏa mãn điều kiện PL (2.5) và các hàm f_i là các hàm L -trơn. Thuật toán SGD với cỡ bước SPS_{max} với $c > \frac{L_{max}}{4\mu}$ và $\gamma_b \geq \frac{1}{2cL_{max}}$ hội tụ như sau:*

$$\mathbb{E}[f(x^k) - f(x^*)] \leq \nu^k [f(x^0) - f(x^*)] + \frac{L\sigma^2\gamma_b}{2(1-\nu)c},$$

trong đó:

$$\nu = \gamma_b \left(\frac{1}{\alpha} - 2\mu + \frac{L_{max}}{2c} \right) \in [0, 1], \quad \alpha = \min \left\{ \frac{1}{2cL_{max}}, \gamma_b \right\}.$$

Chứng minh. Vì hàm f là L -trơn nên ta có:

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2.$$

Kết hợp với quy tắc cập nhật của SGD ta thu được:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &= f(x^k) - \gamma_k \langle \nabla f(x^k), \nabla f_i(x^k) \rangle + \frac{L\gamma_k^2}{2} \|\nabla f_i(x^k)\|^2. \end{aligned} \quad (2.3.7)$$

Sắp xếp lại:

$$\begin{aligned} \frac{f(x^{k+1}) - f(x^k)}{\gamma_k} &\leq -\langle \nabla f(x^k), \nabla f_i(x^k) \rangle + \frac{L\gamma_k}{2} \|\nabla f_i(x^k)\|^2 \\ &\stackrel{(2.4)}{\leq} -\langle \nabla f(x^k), \nabla f_i(x^k) \rangle + \frac{L}{2c} [f_i(x^k) - f_i^*] \\ &= -\langle \nabla f(x^k), \nabla f_i(x^k) \rangle + \frac{L}{2c} [f_i(x^k) - f_i(x^*)] + \frac{L}{2c} [f_i(x^*) - f_i^*]. \end{aligned}$$

và lấy kì vọng có điều kiện dựa trên x^k :

$$\begin{aligned} \mathbb{E}_i \left[\frac{f(x^{k+1}) - f(x^k)}{\gamma_k} \right] &\leq -\langle \nabla f(x^k), \nabla f(x^k) \rangle + \frac{L}{2c} [f(x^k) - f(x^*)] + \frac{L}{2c} \mathbb{E}_i [f_i(x^*) - f_i^*] \\ &\stackrel{(2.5)}{\leq} -\|\nabla f(x^k)\|^2 + \frac{L}{2c} [f(x^k) - f(x^*)] + \frac{L}{2c} \sigma^2 \\ &\stackrel{(4)}{\leq} -2\mu [f(x^k) - f(x^*)] + \frac{L}{2c} [f(x^k) - f(x^*)] + \frac{L}{2c} \sigma^2. \end{aligned}$$

Đặt $\alpha = \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}$. Ta được,

$$\begin{aligned}
\mathbb{E}_i \left[\frac{f(x^{k+1}) - f(x^*)}{\gamma_k} \right] &\leq \mathbb{E}_i \left[\frac{f(x^k) - f(x^*)}{\gamma_k} \right] - 2\mu [f(x^k) - f(x^*)] \\
&\quad + \frac{L}{2c} [f(x^k) - f(x^*)] + \frac{L}{2c} \sigma^2 \\
&\stackrel{(2.6), (2.4)}{\leq} \frac{1}{\alpha} [f(x^k) - f(x^*)] - 2\mu [f(x^k) - f(x^*)] \\
&\quad + \frac{L}{2c} [f(x^k) - f(x^*)] + \frac{L}{2c} \sigma^2 \\
&= \left(\frac{1}{\alpha} - 2\mu + \frac{L}{2c} \right) [f(x^k) - f(x^*)] + \frac{L}{2c} \sigma^2 \\
&\stackrel{L \leq L_{\max}}{=} \left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c} \right) [f(x^k) - f(x^*)] + \frac{L}{2c} \sigma^2. \quad (2.3.8)
\end{aligned}$$

Sử dụng $\gamma_k \leq \gamma_b$ và lấy kì vọng 1 lần nữa:

$$\mathbb{E} [f(x^{k+1}) - f(x^*)] \leq \underbrace{\gamma_b \left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c} \right)}_{\nu} \mathbb{E} [f(x^k) - f(x^*)] + \frac{L\sigma^2\gamma_b}{2c}. \quad (2.3.9)$$

Có $\nu \in (0, 1]$, sử dụng đệ quy và tính tổng ta thu được:

$$\begin{aligned}
\mathbb{E} [f(x^k) - f(x^*)] &\leq \nu^k [f(x^0) - f(x^*)] + \frac{L\sigma^2\gamma_b}{2c} \sum_{j=0}^{k-1} \nu^j \\
&\leq \nu^k [f(x^0) - f(x^*)] + \frac{L\sigma^2\gamma_b}{2(1-\nu)c}. \quad (2.3.10)
\end{aligned}$$

Trong kết quả bên trên chúng ta yêu cầu $0 < \nu = \gamma_b \left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c} \right) \leq 1$. Để điều này đúng, chúng ta cần đưa ra các giả định bổ sung về giá trị của γ_b và tham số c . Đây là điều chúng ta sẽ làm tiếp theo.

Phân tích thành hai trường hợp dựa trên giá trị của tham số α . Cụ thể:

- (i) Nếu $\frac{1}{2cL_{\max}} \leq \gamma_b$ thì,

$$\alpha = \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\} = \frac{1}{2cL_{\max}} \quad \text{và} \quad \nu = \gamma_b \left(\left(2c + \frac{1}{2c} \right) L_{\max} - 2\mu \right).$$

Bằng tính toán cơ bản, dễ dàng chỉ ra rằng $\nu > 0$ với mọi $c \geq 0$. Tuy nhiên với $\nu \leq 1$ ta cần yêu cầu rằng $\gamma_b \leq \frac{1}{\left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c} \right)}$ và vì chúng ta đã giả định $\frac{1}{2cL_{\max}} \leq \gamma_b$ chúng ta cần bắt buộc

$$\frac{1}{2cL_{\max}} \leq \frac{1}{\left(\frac{1}{\alpha} - 2\mu + \frac{L_{\max}}{2c} \right)}$$

để tránh mâu thuẫn. Điều này chỉ đúng nếu $c > \frac{L_{\max}}{4\mu}$, đây là giả định của Định lý 3.

- (ii) Nếu $\gamma_b \leq \frac{1}{2cL_{\max}}$ thì,

$$\alpha = \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\} = \gamma_b \quad \text{và} \quad \nu = \gamma_b \left(\frac{1}{\gamma_b} - 2\mu + \frac{L_{\max}}{2c} \right) = 1 - 2\mu\gamma_b + \frac{L_{\max}}{2c}\gamma_b.$$

Lưu ý rằng nếu $c > \frac{L_{\max}}{4\mu}$ (một giả định của Định lý 3) thì $\nu \leq 1$. Thêm nữa, có thể chỉ ra rằng $\nu > 0$ nếu $\gamma_b < \frac{2c}{4\mu c - L_{\max}}$. Cuối cùng, với $c > \frac{L_{\max}}{4\mu}$ có $\frac{1}{2cL_{\max}} \leq \frac{2c}{4\mu c - L_{\max}}$, và kết quả là $\nu > 0$ với mọi $\gamma_b \leq \frac{1}{2cL_{\max}}$.

Như vậy, ta có điều phải chứng minh. \square

Trong trường hợp nội suy, khi $\sigma = 0$, thuật toán SPS_{\max} hội tụ tới nghiệm tối ưu với tốc độ tuyến tính. Nếu $\gamma_b \leq \min \left\{ \frac{1}{2cL_{\max}}, \frac{2c}{4\mu - L_{\max}} \right\}$ (sử dụng cận dưới trong (2.6)), phương pháp phân tích trở thành SGD với cỡ bước hằng và chúng ta có hệ quả sau.

Hệ quả 3. *Giả sử rằng f thỏa mãn điều kiện PL (2.5) và các hàm f_i là các hàm L -trơn. Thuật toán SGD với cỡ bước hằng $\gamma_k = \gamma \leq \frac{\mu}{L_{\max}^2}$ hội tụ như sau:*

$$\mathbb{E}[f(x^k) - f(x^*)] \leq \nu^k [f(x^0) - f(x^*)] + \frac{L\sigma^2\gamma}{2(1-\nu)c}.$$

Hệ quả trên đạt được bằng cách đơn giản sử dụng $c = \frac{L_{\max}}{2\mu}$ trong trường hợp (ii) của chứng minh bên trên. Trong trường hợp này ta có $\gamma \leq \frac{\mu}{L_{\max}^2}$ và $\nu = 1 - \mu\gamma$.

2.3.5 Hàm f là hàm không lồi tổng quát

Trong phần này, chúng tôi giả định một điều kiện phổ biến được sử dụng để chứng minh sự hội tụ của SGD trong thiết lập không lồi (Bottou, 2018) [9].

$$\mathbb{E}[\|\nabla f_i(x)\|^2] \leq \rho \|\nabla f(x)\|^2 + \delta, \quad (2.9)$$

trong đó $\rho, \delta > 0$ là các hằng số.

Định lý 4. *Cho f và f_i là những hàm L -trơn và giả sử tồn tại $\rho, \delta > 0$ sao cho thỏa mãn điều kiện (2.9). SGD cỡ bước SPS_{\max} với $c > \frac{\rho L}{4L_{\max}}$ và $\gamma_b < \max \left\{ \frac{2}{L\rho}, \bar{\gamma}_b \right\}$ hội tụ như sau:*

$$\min_{k \in [K]} \mathbb{E} \|\nabla f(x^k)\|^2 \leq \frac{2}{\zeta K} (f(x^0) - f(x^*)) + \frac{(\gamma_b - \alpha + L\gamma_b^2) \delta}{\zeta},$$

trong đó $\alpha = \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\}$, $\zeta = (\gamma_b + \alpha) - \rho (\gamma_b - \alpha + L\gamma_b^2)$ và

$$\bar{\gamma}_b := \frac{-(\rho - 1) + \sqrt{(\rho - 1)^2 + \frac{4L\rho(\rho + 1)}{2cL_{\max}}}}{2L\rho}.$$

Từ định lý trên, chúng ta quan sát thấy rằng SGD với SPS đạt được hội tụ $O(1/K)$ đến một vùng lân cận được kiểm soát bởi δ . Trong trường hợp $\delta = 0$, điều kiện (2.9) thu gọn thành điều kiện tăng trưởng mạnh (SGC). Dễ dàng chứng minh rằng các hàm thỏa mãn điều kiện SGC cũng thỏa mãn tính chất nội suy [10]. Trong trường hợp đặc biệt của nội suy, SGD với SPS có thể tìm được điểm dừng bậc nhất hiệu quả như phương pháp hướng giảm gradient. Hơn nữa, với $c \in \left(\frac{\rho L}{4L_{\max}}, \frac{\rho L}{2L_{\max}} \right]$, cận dưới $\frac{1}{2cL_{\max}}$ của SPS nằm trong khoảng $\left[\frac{1}{\rho L}, \frac{2}{\rho L} \right)$, do đó, cỡ bước lớn hơn $\frac{1}{\rho L}$, là cỡ bước hằng tốt nhất được phân tích trong trường hợp này [10].

Chứng minh. Trước tiên ta có:

$$\begin{aligned} -\gamma_k \langle \nabla f(x^k), \nabla f_i(x^k) \rangle &= \frac{\gamma_k}{2} \|\nabla f_i(x^k) - \nabla f(x^k)\|^2 - \frac{\gamma_k}{2} \|\nabla f_i(x^k)\|^2 - \frac{\gamma_k}{2} \|\nabla f(x^k)\|^2 \\ &\stackrel{(2.3.2)}{\leq} \frac{\gamma_b}{2} \|\nabla f_i(x^k) - \nabla f(x^k)\|^2 - \frac{\alpha}{2} \|\nabla f_i(x^k)\|^2 - \frac{\alpha}{2} \|\nabla f(x^k)\|^2 \\ &= \frac{\gamma_b}{2} \|\nabla f_i(x^k)\|^2 + \frac{\gamma_b}{2} \|\nabla f(x^k)\|^2 - \gamma_b \langle \nabla f(x^k), \nabla f_i(x^k) \rangle \\ &\quad - \frac{\alpha}{2} \|\nabla f_i(x^k)\|^2 - \frac{\alpha}{2} \|\nabla f(x^k)\|^2 \\ &= \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} \right) \|\nabla f_i(x^k)\|^2 + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} \right) \|\nabla f(x^k)\|^2 \\ &\quad - \gamma_b \langle \nabla f(x^k), \nabla f_i(x^k) \rangle. \end{aligned} \tag{2.3.11}$$

Với tính L -trơn của hàm f ta có:

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2.$$

Kết hợp với quy tắc cập nhật của SGD:

$$\begin{aligned} f(x^{k+1}) &\leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2 \\ &\leq f(x^k) - \gamma_k \langle \nabla f(x^k), \nabla f_i(x^k) \rangle + \frac{L\gamma_k^2}{2} \|\nabla f_i(x^k)\|^2 \\ &\stackrel{(2.4)}{\leq} f(x^k) - \gamma_k \langle \nabla f(x^k), \nabla f_i(x^k) \rangle + \frac{L\gamma_b^2}{2} \|\nabla f_i(x^k)\|^2 \\ &\stackrel{(2.3.11)}{\leq} f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2} \right) \|\nabla f_i(x^k)\|^2 + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} \right) \|\nabla f(x^k)\|^2 \\ &\quad - \gamma_b \langle \nabla f(x^k), \nabla f_i(x^k) \rangle. \end{aligned} \tag{2.3.12}$$

Lấy kì vọng có điều kiện dựa trên x^k :

$$\begin{aligned}
\mathbb{E}_i f(x^{k+1}) &\stackrel{(2.3.12)}{\leq} f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2} \right) \mathbb{E}_i \|\nabla f_i(x^k)\|^2 + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} \right) \|\nabla f(x^k)\|^2 \\
&\quad - \gamma_b \langle \nabla f(x^k), \mathbb{E}_i \nabla f_i(x^k) \rangle \\
&= f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2} \right) \mathbb{E}_i \|\nabla f_i(x^k)\|^2 + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} \right) \|\nabla f(x^k)\|^2 \\
&\quad - \gamma_b \|\nabla f(x^k)\|^2 \\
&= f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2} \right) \mathbb{E}_i \|\nabla f_i(x^k)\|^2 - \left(\frac{\gamma_b}{2} + \frac{\alpha}{2} \right) \|\nabla f(x^k)\|^2
\end{aligned} \tag{2.3.13}$$

Vì $0 < \alpha \leq \gamma_b$ ta có $\left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2} \right) > 0$. Do đó, chúng ta có thể sử dụng (2.9):

$$\begin{aligned}
\mathbb{E}_i f(x^{k+1}) &\leq f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2} \right) \mathbb{E}_i \|\nabla f_i(x^k)\|^2 - \left(\frac{\gamma_b}{2} + \frac{\alpha}{2} \right) \|\nabla f(x^k)\|^2 \\
&\stackrel{(2.9)}{\leq} f(x^k) + \left(\frac{\gamma_b}{2} - \frac{\alpha}{2} + \frac{L\gamma_b^2}{2} \right) [\rho \|\nabla f(x)\|^2 + \delta] - \left(\frac{\gamma_b}{2} + \frac{\alpha}{2} \right) \|\nabla f(x^k)\|^2 \\
&= f(x^k) + \frac{1}{2} [(\gamma_b - \alpha + L\gamma_b^2) \rho - (\gamma_b + \alpha)] \|\nabla f(x)\|^2 + \frac{1}{2} (\gamma_b - \alpha + L\gamma_b^2) \delta.
\end{aligned}$$

Sắp xếp và lấy tiếp kì vọng:

$$\underbrace{[(\gamma_b + \alpha) - (\gamma_b - \alpha + L\gamma_b^2) \rho]}_{\zeta} \mathbb{E} [\|\nabla f(x^k)\|^2] \leq 2 (\mathbb{E}[f(x^k)] - \mathbb{E}[f(x^{k+1})]) + (\gamma_b - \alpha + L\gamma_b^2) \delta.$$

Nếu $\zeta > 0$ thì:

$$\mathbb{E} [\|\nabla f(x^k)\|^2] \leq \frac{2}{\zeta} (\mathbb{E}[f(x^k)] - \mathbb{E}[f(x^{k+1})]) + \frac{(\gamma_b - \alpha + L\gamma_b^2) \delta}{\zeta}. \tag{2.3.14}$$

Lấy tổng từ $k = 0$ đến $K - 1$ và chia cho K :

$$\begin{aligned}
\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E} [\|\nabla f(x^k)\|^2] &\leq \frac{2}{\zeta} \frac{1}{K} \sum_{k=0}^{K-1} (\mathbb{E}[f(x^k)] - \mathbb{E}[f(x^{k+1})]) + \frac{1}{K} \sum_{k=0}^{K-1} \frac{(\gamma_b - \alpha + L\gamma_b^2) \delta}{\zeta} \\
&\leq \frac{2}{\zeta} \frac{1}{K} (f(x^0) - \mathbb{E}[f(x^K)]) + \frac{(\gamma_b - \alpha + L\gamma_b^2) \delta}{\zeta} \\
&\leq \frac{2}{\zeta K} (f(x^0) - f(x^*)) + \frac{(\gamma_b - \alpha + L\gamma_b^2) \delta}{\zeta}.
\end{aligned} \tag{2.3.15}$$

Trong kết quả trên, chúng ta yêu cầu rằng $\zeta = (\gamma_b + \alpha) - (\gamma_b - \alpha + L\gamma_b^2) \rho > 0$.

Để điều này đúng, chúng ta cần đưa ra các giả định bổ sung về giá trị của γ_b và tham số c . Đây là điều chúng ta sẽ làm tiếp theo.

Phân tích thành hai trường hợp như sau:

- (i) Nếu $\frac{1}{2cL_{\max}} \leq \gamma_b$ thì $\alpha = \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\} = \frac{1}{2cL_{\max}}$ và

$$\zeta = (\gamma_b + \alpha) - \left(\gamma_b - \alpha + L\gamma_b^2 \right) \rho = \left(\gamma_b + \frac{1}{2cL_{\max}} \right) - \left(\gamma_b - \frac{1}{2cL_{\max}} + L\gamma_b^2 \right) \rho.$$

Bằng cách giải biểu thức bậc hai của ζ theo γ_b , có thể dễ dàng chứng minh rằng $\zeta > 0$ nếu

$$0 < \gamma_b < \bar{\gamma}_b := \frac{-(\rho - 1) + \sqrt{(\rho - 1)^2 + \frac{4L\rho(\rho + 1)}{2cL_{\max}}}}{2L\rho}.$$

Để tránh mâu thuẫn, bất đẳng thức $\frac{1}{2cL_{\max}} < \bar{\gamma}_b$ cần phải đúng, trong đó $\bar{\gamma}_b$ là cận trên của γ_b đã được đề cập ở trên. Điều này xảy ra khi $c > \frac{L\rho}{4L_{\max}}$, đây là giả định của Định lý 4.

- (ii) Nếu $\gamma_b \leq \frac{1}{2cL_{\max}}$ thì $\alpha = \min \left\{ \frac{1}{2cL_{\max}}, \gamma_b \right\} = \gamma_b$ và

$$\zeta = (\gamma_b + \alpha) - \left(\gamma_b - \alpha + L\gamma_b^2 \right) \rho = (\gamma_b + \gamma_b) - \left(\gamma_b - \gamma_b + L\gamma_b^2 \right) \rho = 2\gamma_b - L\gamma_b^2\rho.$$

Trong trường hợp này, qua các phép toán sơ bộ, có thể chứng minh rằng $\zeta > 0$ nếu $\gamma_b < \frac{2}{L\rho}$. Đối với $c > \frac{L\rho}{4L_{\max}}$, cũng có điều kiện $\frac{1}{2cL_{\max}} < \frac{2}{L\rho}$.

□

Kết luận

Như vậy ta đã hoàn thành việc tìm hiểu và phân tích thuật toán hướng giảm Gradient ngẫu nhiên (SGD) với cỡ bước Polyak ngẫu nhiên (SPS), một phương pháp tự động điều chỉnh cỡ bước dựa trên giá trị hàm mục tiêu. Thuật toán này được chứng minh là hiệu quả trong cả ba trường hợp: hàm lồi mạnh, hàm lồi, và hàm không lồi. Cỡ bước SPS giúp giảm thiểu rủi ro từ việc thiết lập sai tốc độ học, tăng tính ổn định và đơn giản trong triển khai so với các phương pháp truyền thống. Tuy nhiên, nó có thể gặp khó khăn trong các bài toán phức tạp khi giá trị mục tiêu không được xác định chính xác. Chương này đã đặt nền tảng lý thuyết quan trọng cho việc thực nghiệm trong chương tiếp theo.

Chương 3

Lập trình thử nghiệm

Trong chương này, ta sẽ tiến hành thực nghiệm để kiểm chứng các kết quả lý thuyết đã trình bày ở chương trước.

3.1 Thử nghiệm so sánh thuật toán SGD với cỡ bước hằng và SPS_{\max}

Thiết lập thử nghiệm theo quy trình được thực hiện trong [11]:

- Bài toán: Phân loại nhị phân.
- Dữ liệu: $n = 1000$, $d = 100$.
- Hàm mất mát: hồi quy logistic, hồi quy logistic L2-regularization.
- Hàm mục tiêu: dữ liệu được tạo ra để đảm bảo lời mạnh.

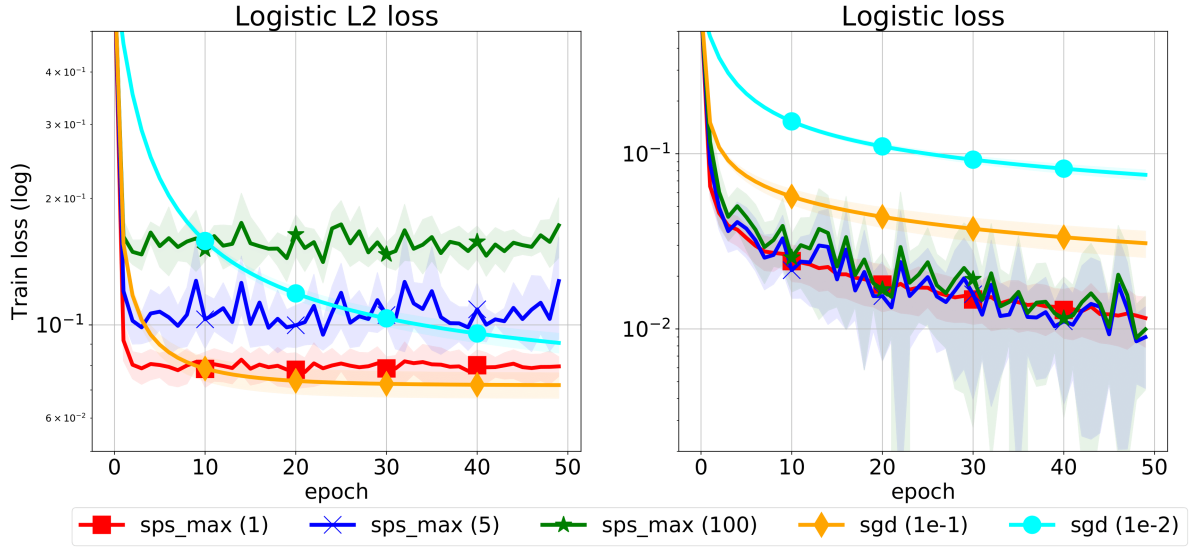
Chúng ta đánh giá hiệu suất của SPS_{\max} và đặt $c = 1/2$ như được đề xuất bởi Định lý 1, thử nghiệm với ba giá trị $\gamma_b = \{1, 5, 100\}$.

Trong trường hợp sử dụng L2-regularization, f_i^* được tính trước dưới dạng tường minh cho mỗi i bằng cách sử dụng hàm Lambert W [12]; trong khi f_i^* bằng 0 trong trường hợp còn lại.

Trong cả hai trường hợp, chúng ta so sánh hiệu suất của SGD với cỡ bước SPS_{\max} và cỡ bước hằng lần lượt là $\gamma = \{0.1, 0.01\}$.

Kết quả: Từ Hình 3.1, trong cả 2 trường hợp ta nhận thấy rằng tuy SGD hội tụ ổn định nhưng các biến thể SPS cho ta kết quả hội tụ nhanh hơn. SGD hội tụ tốt

với cỡ bước 0.1, hội tụ chậm khi sử dụng cỡ bước 0.01. Ngược lại, tất cả các biến thể của SPS đều hội tụ đến một vùng lân cận của điểm tối ưu và kích thước của vùng lân cận tăng lên khi γ_b tăng như được dự đoán bởi lý thuyết.



Hình 3.1: Kết quả so sánh SGD với cỡ bước SPS_{\max} và cỡ bước hằng cho bài toán phân loại nhị phân có và không sử dụng hàm mất mát hồi quy logistic L2.

3.2 Thử nghiệm cho mô hình quá tham số

Trong phần này, chúng ta huấn luyện các mô hình quá tham số (over-parameterized) xấp xỉ thỏa mãn điều kiện nội suy (interpolation condition). Các phương pháp thích nghi được sử dụng ở phần này để so sánh với cỡ bước SPS_{\max} là:

- Adam (2015) [13]
- Stochastic Line-Search (SLS) (2019) [14]
- ALI-G (2020) [15]
- Rectified Adam (RADAM) (2019) [16]
- Look-Ahead Optimizer (2019) [17].

Và được kí hiệu như sau:



Hình 3.2: Kí hiệu các phương pháp thích nghi.

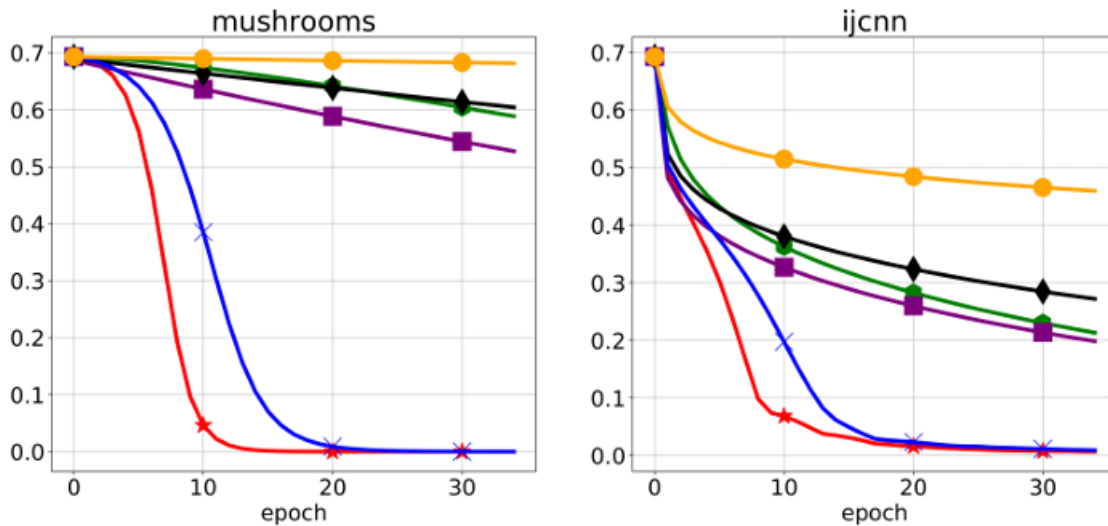
3.2.1 Phân loại nhị phân sử dụng kernel (Bài toán lỗi)

Trong phần này, chúng ta so sánh hiệu suất của các bộ tối ưu hóa trong điều kiện lỗi và nội suy (interpolation regime) thông qua bài toán phân loại nhị phân.

Thiết lập thí nghiệm:

- Sử dụng kernel RBF và hàm mất mát logistic *chính quy hóa* (*regularization*).
- Bộ dữ liệu: mushrooms, ijcnn từ bộ dữ liệu LIBSVM.

Kết quả:



Hình 3.3: Thử nghiệm phân loại nhị phân trên bộ dữ liệu mushrooms và ijcnn với kernel RBF (train loss).

- SPS_{\max} cho thấy hội tụ nhanh chóng và ổn định trên các bộ dữ liệu khác nhau.
- Trên cả hai bộ dữ liệu **mushrooms** và **ijcnn**, SPS_{\max} đạt hiệu suất giảm mất mát huấn luyện tốt hơn so với các phương pháp khác.

Kết luận: Phương pháp SPS_{\max} không chỉ đơn giản để triển khai mà còn thể hiện hiệu suất cao trong việc giảm mất mát huấn luyện trong điều kiện lỗi và nội suy. Điều này làm cho SPS_{\max} trở thành một lựa chọn tối ưu cho các bài toán phân loại nhị phân sử dụng kernel.

3.2.2 Nhân tử hóa ma trận - Deep matrix factorization (Bài toán không lỗi thỏa mãn PL)

Mục tiêu của thí nghiệm là nghiên cứu tác động của việc quá tham số hóa (over-parameterization) đối với hiệu suất của các phương pháp tối ưu khác nhau. Bài toán được đặt ra là một bài toán hồi quy không lỗi, với hàm mục tiêu:

$$\min_{W_1, W_2} \mathbb{E}_{x \sim \mathcal{N}(0, I)} \|W_2 W_1 x - Ax\|^2,$$

trong đó:

- $A \in \mathbb{R}^{10 \times 6}$ có số điều kiện $\kappa(A) = 10^{10}$, trong đó số điều kiện ($\kappa(A)$) là tỷ lệ giữa giá trị kỳ dị lớn nhất và nhỏ nhất của ma trận A :

$$\kappa(A) = \frac{\sigma_{\max}}{\sigma_{\min}}.$$

Nó đo độ nhạy cảm của nghiệm bài toán $Ax = b$ đối với nhiễu trong A hoặc b .

Số điều kiện nhỏ ($\kappa(A) \approx 1$) cho thấy ma trận ổn định, còn số điều kiện lớn cho thấy ma trận dễ bị sai lệch.

- Tập dữ liệu cố định bao gồm 1.000 mẫu.

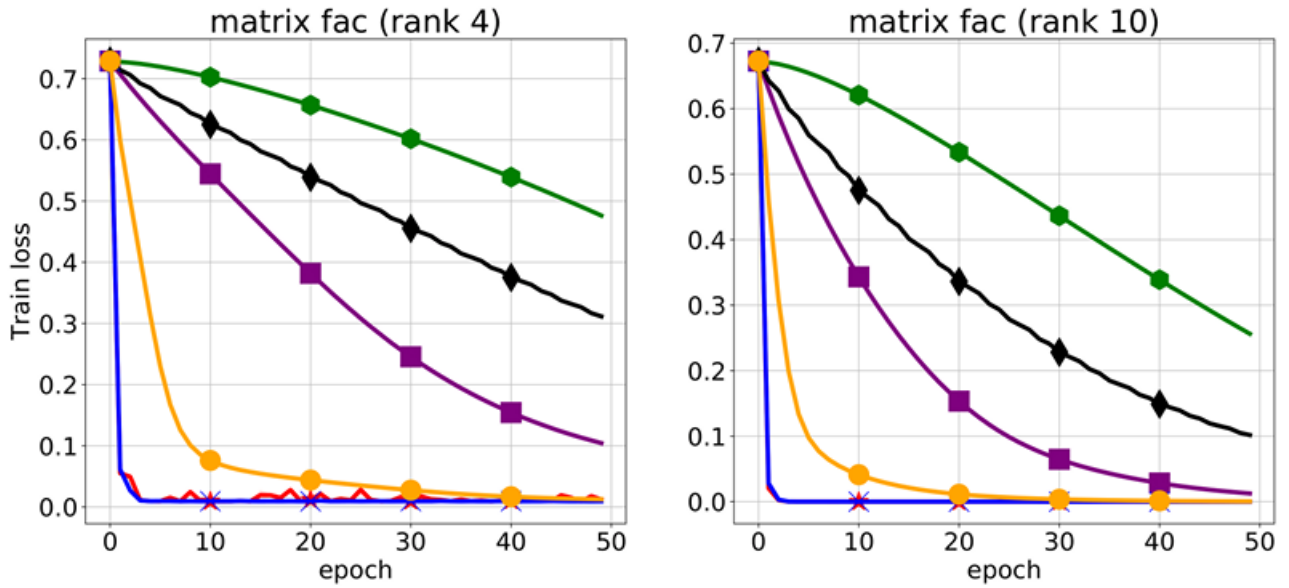
Mức độ quá tham số hóa được kiểm soát thông qua hạng k của các ma trận:

$$W_1 \in \mathbb{R}^{k \times 6}, \quad W_2 \in \mathbb{R}^{10 \times k}.$$

Hai giá trị của k được thử nghiệm:

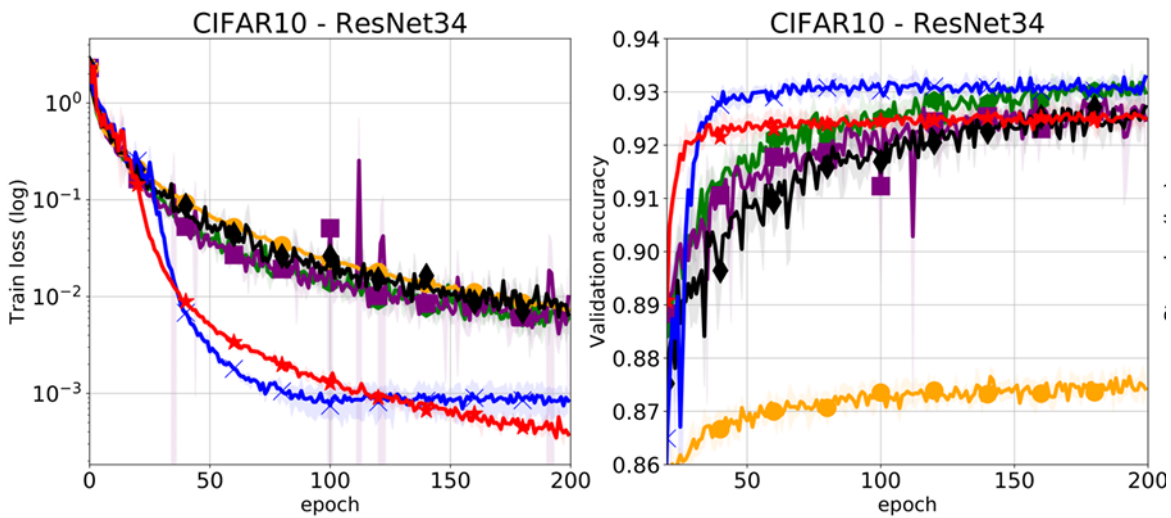
- $k = 4$: Điều kiện nội suy không được thỏa mãn, cả SPS_{\max} và SLS đều hội tụ nhanh chóng.
- $k = 10$: Điều kiện nội suy được thỏa mãn chính xác, SPS_{\max} vẫn duy trì hiệu suất tốt mà không cần tối ưu tham số phức tạp.

Trên thí nghiệm này, ta thấy được sự ổn định của SPS.



Hình 3.4: Nhân tử hóa ma trận với $k = 4$ và $k = 10$.

3.2.3 Phân loại đa lớp sử dụng mạng học sâu (Bài toán không lỗi tổng quát)



Hình 3.5: Kết quả chạy trên bộ dữ liệu CIFAR10 với mô hình ResNet34.

Kết quả:

- SPS_{\max} thể hiện ấn tượng khi đạt mức mất mát huấn luyện thấp nhất trên mô hình ResNet-34 sử dụng bộ dữ liệu CIFAR-10.
- SPS_{\max} về cơ bản là tốt hơn so với các phương pháp tối ưu khác đồng thời có độ ổn định tốt.

Kết luận

Kết quả thực nghiệm cho thấy SPS_{\max} đạt hiệu suất về cơ bản là tốt hơn các phương pháp hiện đại khác như Adam và SLS, đặc biệt trong việc giảm mất mát huấn luyện. SPS_{\max} cũng thể hiện sự ổn định và khả năng thích nghi cao, ngay cả trong các mô hình quá tham số. Những kết quả này đã minh họa rõ ràng tính ứng dụng và đầy tiềm năng của SPS_{\max} trong các bài toán tối ưu hiện đại.

Kết luận chung

1. Tổng kết

Đồ án đã đạt được các mục tiêu đề ra

Tìm hiểu và phân tích thuật toán hướng giảm Gradient ngẫu nhiên (SGD) với cỡ bước Polyak ngẫu nhiên (SPS) nhằm giải quyết các bài toán tối ưu có nhiều ứng dụng trong học máy, học sâu.

Kết quả của đồ án

- Trình bày các kiến thức nền tảng như hàm lồi, điều kiện Polyak-Lojasiewicz (PL), hàm L -trơn, thuật toán hướng giảm ...
- Phân tích lý thuyết về cỡ bước SPS_{\max} , chứng minh sự hội tụ của thuật toán trong các trường hợp hàm lồi mạnh, hàm lồi, và hàm không lồi.
- Thử nghiệm so sánh giữa cỡ bước Polyak ngẫu nhiên và cỡ bước hằng, mô hình quá tham số như phân loại nhị phân, phân loại đa lớp trong học sâu, và nhân tử hóa ma trận; chứng minh SPS_{\max} đạt tốc độ hội tụ ổn định và cơ bản là tốt hơn so với các phương pháp hiện đại khác.

2. Hướng phát triển đề tài

- Mở rộng nghiên cứu cỡ bước SPS sang các lĩnh vực khác như học tăng cường, xử lý ngôn ngữ tự nhiên, và các bài toán tối ưu phi lồi phức tạp.
- Cải tiến thuật toán hướng giảm gradient ngẫu nhiên với cỡ bước SPS kết hợp với momentum để tăng tốc tối nghiệm tối ưu, xử lý một số bài toán phức tạp hơn.

Tài liệu tham khảo

- [1] B. Poliak, *Introduction to Optimization*, ser. Translations series in mathematics and engineering. Optimization Software, Publications Division, 1987. [Online]. Available: <https://books.google.com.vn/books?id=gUXvAAAAMAAJ>
- [2] A. Beck, *Introduction to Nonlinear Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2014. [Online]. Available: <https://epubs.siam.org/doi/abs/10.1137/1.9781611973655>
- [3] —, *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017. [Online]. Available: <https://epubs.siam.org/doi/abs/10.1137/1.9781611974997>
- [4] *Giáo trình các phương pháp tối ưu: Lý thuyết và thuật toán, author=Nguyễn, T.B.K., url=https://books.google.com.vn/books?id=cYAqnQAACAAJ, year=2014, publisher=Nxb Bách Khoa.*
- [5] H. Karimi, J. Nutini, and M. Schmidt, “Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition,” 2020. [Online]. Available: <https://arxiv.org/abs/1608.04636>
- [6] R. Mansel Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtarik, “Sgd: General analysis and improved rates,” *arXiv e-prints*, pp. arXiv–1901, 2019.
- [7] G. Garrigos and R. M. Gower, “Handbook of convergence theorems for (stochastic) gradient methods,” 2024. [Online]. Available: <https://arxiv.org/abs/2301.11235>

- [8] N. Loizou, S. Vaswani, I. Laradji, and S. Lacoste-Julien, “Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence,” 2021. [Online]. Available: <https://arxiv.org/abs/2002.10542>
- [9] L. Bottou, F. E. Curtis, and J. Nocedal, “Optimization methods for large-scale machine learning,” 2018. [Online]. Available: <https://arxiv.org/abs/1606.04838>
- [10] S. Vaswani, F. Bach, and M. Schmidt, “Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron,” 2019. [Online]. Available: <https://arxiv.org/abs/1810.07288>
- [11] J. Nutini, I. Laradji, and M. Schmidt, “Let’s make block coordinate descent converge faster: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence,” 2017. [Online]. Available: <https://arxiv.org/abs/1712.08859>
- [12] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, “On the lambertw function,” *Advances in Computational Mathematics*, vol. 5, pp. 329–359, 1996. [Online]. Available: <https://api.semanticscholar.org/CorpusID:29028411>
- [13] D. P. Kingma, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [14] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel, and S. Lacoste-Julien, “Painless stochastic gradient: Interpolation, line-search, and convergence rates,” *Advances in neural information processing systems*, vol. 32, 2019.
- [15] L. Berrada, A. Zisserman, and M. P. Kumar, “Training neural networks for and by interpolation,” in *International conference on machine learning*. PMLR, 2020, pp. 799–809.
- [16] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” *arXiv preprint arXiv:1908.03265*, 2019.

- [17] M. Zhang, J. Lucas, J. Ba, and G. E. Hinton, “Lookahead optimizer: k steps forward, 1 step back,” *Advances in neural information processing systems*, vol. 32, 2019.