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

Phạm Thị Thanh Hà

20216823

Khoa Toán Tin, Đại học Bách Khoa Hà Nội

Giảng viên hướng dẫn: TS. Phạm Thị Hoài

Ngày 16 tháng 01 năm 2025

Outline

1. Giới thiệu
2. Thuật toán SGD với cỡ bước Polyak ngẫu nhiên
3. Phân tích hội tụ
4. Lập trình thử nghiệm
5. Tổng kết

Outline

Bài toán tối ưu tổng hữu hạn

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

Thuật toán hướng giảm gradient ngẫu nhiên (SGD)

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

trong đó,

- $i \in [n]$: được chọn ngẫu nhiên.
- γ_k: cỡ bước của lần lặp k.

Giới thiệu

1. Một số cỡ bước đã được chọn cho thuật toán hướng giảm gradient ngẫu nhiên (SGD) như cỡ bước hằng, cỡ bước giảm dần, cỡ bước thích nghi: AdaGrad, RMSProp...
2. Cỡ bước Polyak có khả năng tự điều chỉnh dựa trên thông tin hàm mất mát, giúp hội tụ nhanh và hiệu quả. Tuy nhiên, nó yêu cầu biết trước giá trị tối ưu.
3. Với dữ liệu huấn luyện lớn, cỡ bước Polyak ngẫu nhiên (SPS) được đề xuất [1], mang lại hiệu quả cao hơn, giảm chi phí tính toán và phù hợp với thuật toán SGD trong các mô hình học máy hiện đại.

[1] "*Stochastic Polyak Step-size for SGD: An Adaptive Learning Rate for Fast Convergence*" - Nicolas Loizou, Sharan Vaswani, Issam Laradji, Simon Lacoste-Julien, 2021

Outline

Thuật toán SGD với cỡ bước Polyak ngẫu nhiên

Polyak tất định[2]: $\boxed{\gamma_k = \frac{f(x^k) - f^*}{\|\nabla f(x^k)\|^2}}$

trong đó γ_k là cỡ bước của thuật toán hướng giảm (GD): $x^{k+1} = x^k - \gamma_k \nabla f(x^k)$.

Thuật toán SGD với SPS

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

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

- Yêu cầu thông tin về $f_i^* := \inf_x f_i(x)$.
- $c > 0$: phụ thuộc đặc tính của hàm f (ví dụ: $c = 1/2$ hoặc $c = 1$).
- $\gamma_b > 0$: một giới hạn ngăn SPS trở nên quá lớn.

[2] B. Poliak, "Introduction to Optimization", 1987

Outline

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

Hàm mục tiêu	Đại lượng	Hội tụ
Hàm lồi mạnh	$\mathbb{E}[\|x^k - x^*\|^2]$	Tuyến tính
Hàm lồi	$\mathbb{E}\left[f\left(\bar{x}^k\right) - f\left(x^*\right)\right]$	Dưới tuyến tính: $\mathcal{O}(1/k)$
Polyak-Lojasiewicz (PL)	$\mathbb{E}\left[f\left(x^k\right) - f\left(x^*\right)\right]$	Tuyến tính
Hàm không lồi	$\mathbb{E}[\|\nabla f(x^k)\|^2]$	Dưới tuyến tính: $\mathcal{O}(1/k)$

Bảng 1: Tóm tắt các kết quả phân tích hội tụ [3] .

[3]Nicolas Loizou và cộng sự - "*Stochastic Polyak Step-size for SGD: An Adaptive Learning Rate for Fast Convergence*"

Định lí 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 $\text{SPS}_{\max}$ với $c \geq 1/2$ hội tụ như sau:

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

với $\alpha := \min\left\{\frac{1}{2cL_{\max}}, \gamma_b\right\}$ và $L_{\max} = \max\{L_i\}_{i=1}^n$ ($c = 1/2$ tốc độ tốt nhất).

- **Hệ quả 1:** Giả sử nội suy ($\sigma = 0$). SGD với SPS và $c = 1/2$ 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:** Nếu $\gamma_b \leq \frac{1}{L_{\max}}$, thì SGD với cỡ bước hằng $\gamma_b \leq \frac{1}{L_{\max}}$ thỏa mãn:

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

Định lí 2

Giả sử rằng f_i là các hàm lồi, L_i-trơn. Thuật toán SGD với cỡ bước SPS_{max}, $c = 1$, hội tụ như sau:

$$\mathbb{E}[f(\bar{x}^k) - f(x^*)] \leq \frac{\|x^0 - x^*\|^2}{\alpha K} + \frac{2\sigma^2\gamma_b}{\alpha}, \tag{3}$$

trong đó:

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

Định lí 3

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

$$\mathbb{E}[f(x^k) - f(x^*)] \leq \nu^k [f(x^0) - f(x^*)] + \frac{L\sigma^2\gamma_b}{2(1-\nu)c}, \tag{4}$$

trong đó:

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

Định lí 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 $\mathbb{E}[||\nabla f_i(x)||^2] \leq \rho||\nabla f(x)||^2 + \delta$. SGD cỡ bước $\text{SPS}_{\max}$ với $c > \frac{\rho L}{4L_{\max}}$ và $\gamma_b < \max\left\{\frac{2}{L\rho}, \bar{\gamma_b}\right\}$ hội tụ như sau:

$$\min_{k\in[K]} \|\nabla f(x^k)\|^2 \leq \frac{2}{\zeta K}\left(f(x^0) - f(x^*)\right) + \frac{\left(\gamma_b - \alpha + L\gamma_b^2\right)\delta}{\zeta}, \tag{5}$$

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

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

Outline

So sánh SPS với cỡ bước hằng trong cùng thuật toán SGD

Hàm mất mát hồi quy logistic với L2-Regularization

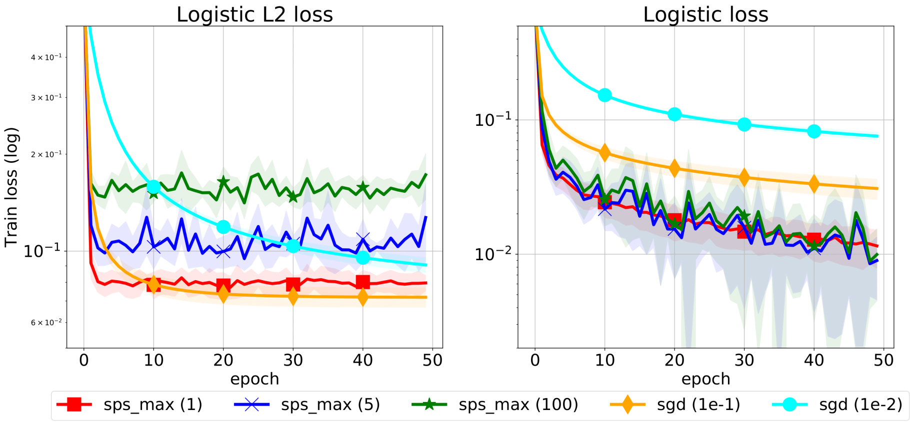
$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-b_i\langle A_i, x\rangle}\right) + \frac{\lambda}{2}\|x\|^2, \tag{6}$$

trong đó:

- A_i là vector đặc trưng đầu vào của điểm dữ liệu thứ i.
- $b_i \in \{-1, 1\}$ là nhãn tương ứng.
- λ là tham số điều chỉnh (regularization parameter).

So sánh SPS với cỡ bước hằng trong cùng thuật toán SGD

Kết quả



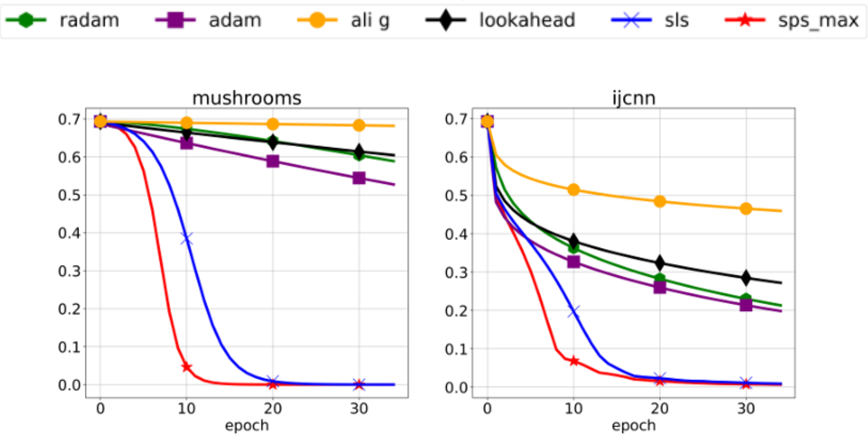
Hình 1: Kết quả so sánh SGD với cỡ bước SPS_{max} và cỡ bước hằng cho bài toán phân loại nhị phân có và không sử dụng hàm mất mát hồi quy logistic L2.

Phân loại nhị phân sử dụng kernel

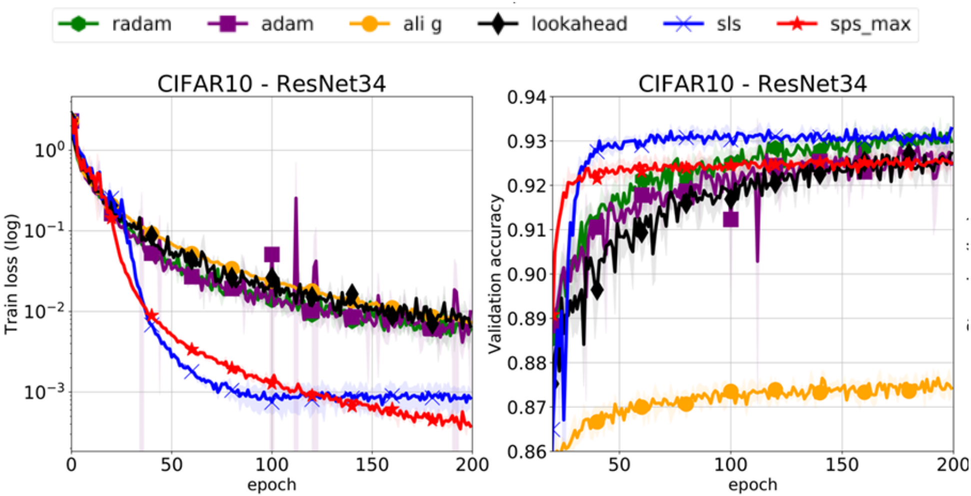
Thiết lập thí nghiệm:

- Sử dụng kernel RBF và hàm mất mát logistic.
- Bộ dữ liệu: mushrooms, ijcnn từ bộ dữ liệu LIBSVM.

Phân loại nhị phân sử dụng kernel



Hình 2: Thử nghiệm phân loại nhị phân trên bộ dữ liệu mushrooms và ijcnn với kernel RBF (train loss).



Hình 3: Kết quả chạy trên bộ dữ liệu CIFAR10 với mô hình ResNet34.

Outline

Tổng kết

1. Tìm hiểu về cỡ bước SPS, SPSmax và thuật toán SGD với SPS.
2. Hiểu được phân tích sự hội tụ của thuật toán trong các trường hợp hàm lồi mạnh, hàm lồi, và hàm không lồi.
3. **SPS là một lựa chọn hấp dẫn cho SGD:** Yêu cầu thông tin về f_i^*.
4. Hiệu suất mạnh mẽ của SGD với SPS so với các phương pháp tối ưu hóa khác trong một vài ví dụ thử nghiệm.

Hướng phát triển

- SPS trong phương pháp momentum.
- SPS trong thiết lập phân tán.

Cảm ơn thầy cô và các bạn đã chú ý lắng nghe!