Търсене и извличане на информация. Приложение на дълбоко машинно обучение

Стоян Михов

Лекция 9: Сходимост на спускането по градиента. Стохастично спускане по градиента.

План на лекцията

1. **Формалности за курса (5 мин)**
2. Спускане по градиент, развиване в ред на Тейлър и операторна норма (15 мин)
3. Сходимост на спускането по градиента (15 мин)
4. Спускане по стохастичен градиент и условно очакване (15 мин)
5. Сходимост на спускането по стохастичен градиент (20 мин)
6. Партиден стохастичен градиент и спускане по стохастичен градиент с адаптация и инерционен момент (20 мин)

Формалности

- Решенията на първото домашно задание следва да бъдат качени в Moodle до края на деня на 30.11.2025
- Второто домашно задание ще бъде публикувано в Moodle около средата на декември.
- Деветата лекция се базира на глави 4 и 5 от втория учебник.

План на лекцията

1. Формалности за курса (5 мин)

2. **Спускане по градиент, развиване в ред на Тейлър и операторна норма (15 мин)**

3. Сходимост на спускането по градиента (15 мин)

4. Спускане по стохастичен градиент и условно очакване (15 мин)

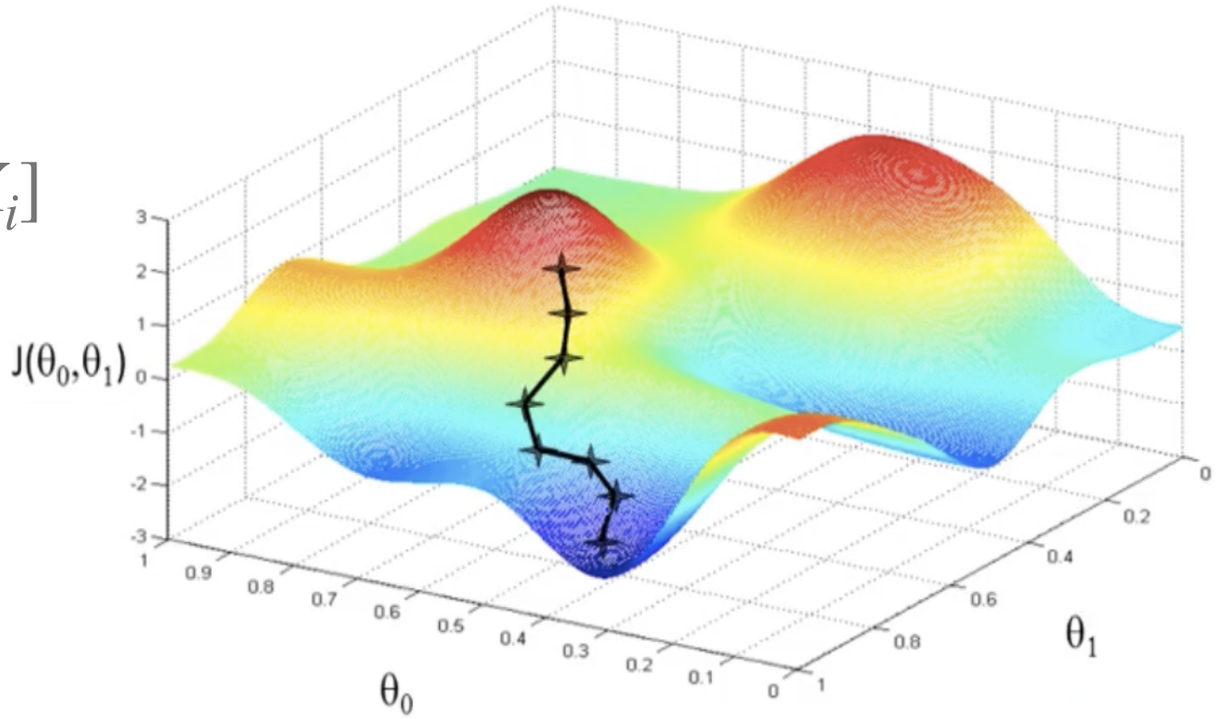
5. Сходимост на спускането по стохастичен градиент (20 мин)

6. Партиден стохастичен градиент и спускане по стохастичен градиент с адаптация и инерционен момент (20 мин)

Обучение чрез спускане по градиента

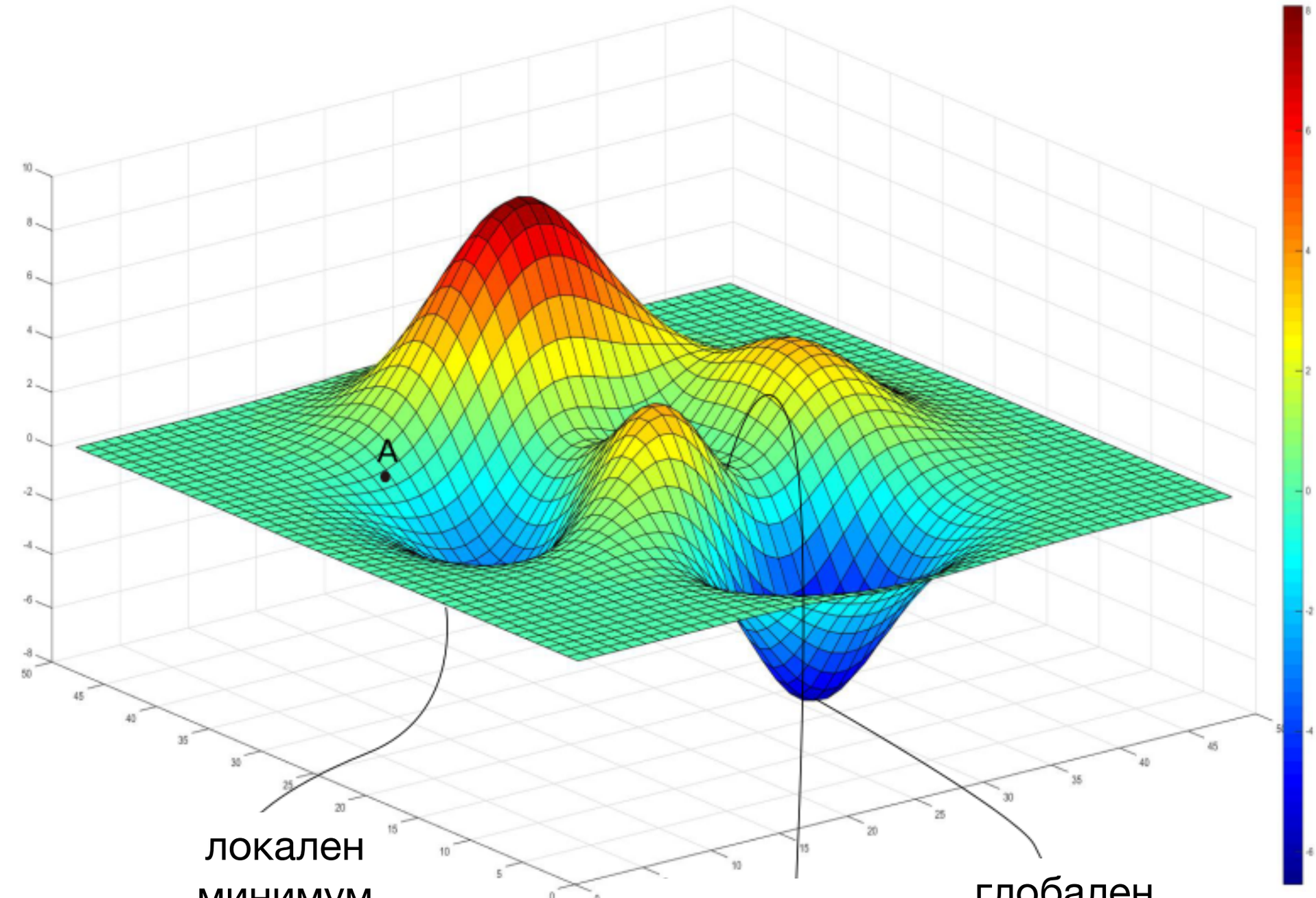
- При машинното обучение се стремим да минимизираме кросентропията върху корпус $X = \{X_1, X_2, \ldots, X_{|X|}\}$:

$$H_X[\Pr \| \Pr_\theta] = -\frac{1}{|X|} \sum_{i=1}^{|X|} \log \Pr_\theta[X_i]$$



- Нека означим $f(\theta) = H_X[\Pr \| \Pr_\theta]$.
- Започваме от начално θ_0
- На всяка стъпка се спускаме по градиента: $\theta_{k+1} = \theta_k - \alpha \nabla_\theta f(\theta_k)$

Минимизация на многомерна функция



Градиент, Хесиан, развиване в ред на Тейлър

Теорема (Тейлър) в $\mathbb{R}$: Нека $g : \mathbb{R} \to \mathbb{R}$ е два пъти диференцируема с непрекъснати производни в околност на точката $t_0 \in \mathbb{R}$. Нека $t \in \mathbb{R}$ е произволна точка в тази околност. Тогава съществува $\bar{t} \in (t_0, t)$, така че:

$$g(t) = g(t_0) + g'(t_0)(t - t_0) + \frac{1}{2} g''(\bar{t})(t - t_0)^2.$$

Ще изведем теоремата на Тейлър за многомерния случай.

Нека $f : \mathbb{R}^n \to \mathbb{R}$. **Градиент** и **Хесиан** на f наричаме съответно:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \frac{\partial}{\partial x_2} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}) \end{bmatrix}, \nabla^2_{\mathbf{x}} f(\mathbf{x}) = \frac{\partial^2}{\partial \mathbf{x}^2} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} f(\mathbf{x}) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} f(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} f(\mathbf{x}) & \cdots & \frac{\partial^2}{\partial x_n \partial x_n} f(\mathbf{x}) \end{bmatrix}$$

Нека $f : \mathbb{R}^n \to \mathbb{R}$. Т.е. за $\mathbf{v} \in \mathbb{R}^n$ имаме, че $f(\mathbf{v}) \in \mathbb{R}$. Разглеждаме функциите $\mathbf{v} : \mathbb{R} \to \mathbb{R}^n$, $\mathbf{v}(t) = \mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0)$ и $g : \mathbb{R} \to \mathbb{R}$, $g(t) = f(\mathbf{v}(t)) = f(\mathbf{x}_0 + t(\mathbf{x} - \mathbf{x}_0))$ и прилагаме теоремата на Тейлър за g при $t_0 = 0,\ t = 1$. В такъв случай имаме, че

$$g'(t) = (f(\mathbf{v}(t)))' = \left(\frac{\partial f(\mathbf{v}(t))}{\partial \mathbf{v}}\right)^{\top} \frac{\partial \mathbf{v}(t)}{\partial t} = (\nabla_{\mathbf{v}} f(\mathbf{v}(t)))^{\top}(\mathbf{x} - \mathbf{x}_0),$$

$$g''(t) = \frac{\partial}{\partial t}\left(\left(\frac{\partial f(\mathbf{v}(t))}{\partial \mathbf{v}}\right)^{\top}(\mathbf{x} - \mathbf{x}_0)\right) = (\mathbf{x} - \mathbf{x}_0)^{\top} \frac{\partial^2 f(\mathbf{v}(t))}{\partial \mathbf{v}^2}(\mathbf{x} - \mathbf{x}_0) = (\mathbf{x} - \mathbf{x}_0)^{\top} \nabla^2_{\mathbf{v}} f(\mathbf{v}(t))(\mathbf{x} - \mathbf{x}_0)$$

<u>Теорема</u> (Тейлър) в $\mathbb{R}^n$: Нека $f : \mathbb{R}^n \to \mathbb{R}$ е два пъти диференцируема с непрекъснати производни в околност на точката $\mathbf{x}_0 \in \mathbb{R}^n$. Нека $\mathbf{x} \in \mathbb{R}^n$ е произволна точка в тази околност. Тогава съществува $\bar{t} \in (0,1)$, така че ако $\bar{\mathbf{x}} = \mathbf{x}_0 + \bar{t}(\mathbf{x} - \mathbf{x}_0)$, то:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + (\nabla_{\mathbf{x}} f(\mathbf{x}_0))^{\top}(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^{\top} \nabla^2_{\mathbf{x}} f(\bar{\mathbf{x}})(\mathbf{x} - \mathbf{x}_0).$$

Операторна норма на матрица

- **Дефиниция:** Нека $A \in \mathbb{R}^{M \times N}$ е матрица. **Операторната норма** на A дефинираме като $\|A\| = \sup_{\mathbf{x} \in \mathbb{R}^N \setminus \{0\}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}$.
- **Свойства:**
  - $\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$
  - $\|AB\| \leq \|A\|\|B\|$, за всеки $A \in \mathbb{R}^{M \times N}$ и $B \in \mathbb{R}^{N \times K}$
  - $\|A\| = \sqrt{\lambda_{\max}(A^\top A)}$, където $\lambda_{\max}(A^\top A)$ е най-голямата по модул собствена стойност на матрицата $A^\top A$.
  - $\|A\| \leq \sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} A_{i,j}^2}$ — **Задача:** Докажете го използвайки неравенството на Коши.

План на лекцията

1. Формалности за курса (5 мин)

2. Спускане по градиент, развиване в ред на Тейлър и операторна норма (15 мин)

3. **Сходимост на спускането по градиента (15 мин)**

4. Спускане по стохастичен градиент и условно очакване (15 мин)

5. Сходимост на спускането по стохастичен градиент (20 мин)

6. Партиден стохастичен градиент и спускане по стохастичен градиент с адаптация и инерционен момент (20 мин)

Сходимост на спускането по градиента

- **<u>Теорема:</u>** Нека функцията $f : \mathbb{R}^n \to \mathbb{R}$ е два пъти диференцируема, ограничена отдолу и нека съществува $L > 0$, така че $\|\nabla^2 f(\theta)\| \leq L$. Нека $\theta_0 \in \mathbb{R}^n$ и $\theta_{k+1} = \theta_k - \alpha \nabla f(\theta_k)$ за $k = 0,1,2,\ldots$. Тогава съществува $\alpha > 0$, така че $\lim_{k\to\infty} \|\nabla f(\theta_k)\|^2 = 0$.

- Ще докажем теоремата на следващите страници.

- **<u>Свойство:</u>** Ако функцията $f : \mathbb{R}^n \to \mathbb{R}$ е два пъти диференцируема и има <u>ограничени</u> втори производни, то съществува $L > 0$, така че $\|\nabla^2 f(\theta)\| \leq L$. (Следва от последното свойство за операторната норма.)

- **<u>Следствие:</u>** При условията на теоремата редицата $\{\theta_k\}_{k=0}^{\infty}$ е сходяща и ако $f : \mathbb{R}^n \to \mathbb{R}$ е строго изпъкнала, то редицата клони към глобалния минимум.
  <u>Доказателство:</u> $\|\theta_k - \theta_{k+1}\| = \alpha\|\nabla f(\theta_k)\| \to 0$

Доказателство на теоремата за сходимост на спускането по градиента

- Спускаме се по градиента: $\theta_{k+1} = \theta_k - \alpha \nabla f(\theta_k)$

- От развиването в ред на Тейлър около точката θ_k получаваме:
$$f(\theta_{k+1}) = f(\theta_k) - \nabla f(\theta_k)^\top (\alpha \nabla f(\theta_k)) + \frac{1}{2}(\alpha \nabla f(\theta_k))^\top \nabla^2 f(\bar{\theta})(\alpha \nabla f(\theta_k))$$

- От неравенството на Коши и $\|\nabla^2 f(\bar{\theta})\| \leq L$ имаме, че за всяко $u \in \mathbb{R}^n$ е изпълнено:
$\|u^\top \nabla^2 f(\bar{\theta})\ u\| \leq \|u\|\ \|\nabla^2 f(\bar{\theta})\ u\| \leq L\|u\|^2$.

- Следователно $(\alpha \nabla f(\theta_k))^\top \nabla^2 f(\bar{\theta})(\alpha \nabla f(\theta_k)) \leq \alpha^2 L \|\nabla f(\theta_k)\|^2$.

- Заместваме и получаваме:

$$f(\theta_{k+1}) \leq f(\theta_k) - \alpha \|\nabla f(\theta_k)\|^2 + \frac{\alpha^2 L}{2} \|\nabla f(\theta_k)\|^2$$

$$f(\theta_k) - f(\theta_{k+1}) \geq \alpha(1 - \frac{\alpha L}{2}) \|\nabla f(\theta_k)\|^2$$

- Нека подберем α, така че $0 < \alpha L < 1$. Тогава:

$$f(\theta_k) - f(\theta_{k+1}) \geq \frac{\alpha}{2} \|\nabla f(\theta_k)\|^2 \geq 0$$

- Следователно, на всяка стъпка **стойността на f намалява.** Тъй като редицата $\left\{f(\theta_k)\right\}_{k=0}^{\infty}$ е ограничена и монотонно намаляваща, то тя е сходяща.

- Разликата на последователните членове на редицата е $\geq \frac{\alpha}{2} \|\nabla f(\theta_k)\|^2$.

Следователно $0 = \lim_{k \to \infty} f(\theta_k) - f(\theta_{k+1}) \geq \frac{\alpha}{2} \lim_{k \to \infty} \|\nabla f(\theta_k)\|^2 \geq 0$. ■

План на лекцията

1. Формалности за курса (5 мин)
2. Спускане по градиент, развиване в ред на Тейлър и операторна норма (15 мин)
3. Сходимост на спускането по градиента (15 мин)
4. **Спускане по стохастичен градиент и условно очакване (15 мин)**
5. Сходимост на спускането по стохастичен градиент (20 мин)
6. Партиден стохастичен градиент и спускане по стохастичен градиент с адаптация и инерционен момент (20 мин)

Стандартен стохастичен градиент
Standard Stochastic Gradient

- $f(\theta) = H_X[\Pr \| \Pr_\theta] = -\frac{1}{|X|}\sum_{i=1}^{|X|} \log \Pr_\theta[X_i].$

- $f_i(\theta) = H_{X_i}[\Pr \| \Pr_\theta] = -\log \Pr_\theta[X_i]$ поточковата кросентропия в точката $X_i \in X$.

Спускане по стандартен стохастичен градиент със скорост α_k на стъпка $k = 0,1,2,\ldots$

- Започваме от някое начално: $\theta_0 \in \mathbb{R}^n$.
- На стъпка $k = 0,1,2,\ldots$: $\theta_{k+1} = \theta_k - \alpha_k \nabla f_{i_k}(\theta_k),$

където i_k е случйна величина със стойности в $\{1,2,\ldots,|X|\}$ с **равномерно** случайно разпределение. Т.е. вероятността $\Pr[i_k = n] = \frac{1}{|X|}$.

Условно очакване на случайни величини

Дефиниция: Нека X, Y са случайни величини със съвместна функция на разпределение $(x, y) \mapsto \Pr[X = x, Y = y]$. Условното очакване на $f(X)$ при условие Y наричаме **случайната величина**

$$\mathbb{E}_X[f(X) \mid Y] = \sum_{x \in X(\Omega)} \Pr[X = x \mid Y = y]\, f(x) = \sum_{x \in X(\Omega)} \frac{\Pr[X = x, Y = y]}{\Pr[Y = y]}\, f(x)$$

Теорема (за пълното очакване): $\mathbb{E}[\mathbb{E}_X[f(X) \mid Y]] = \mathbb{E}[f(X)]$

$$\begin{aligned}
\mathbb{E}[\mathbb{E}_X[f(X) \mid Y]] &= \sum_{y \in Y(\Omega)} \Pr[Y = y]\, \mathbb{E}_X[f(X) \mid Y] = \\
&= \sum_{y \in Y(\Omega)} \Pr[Y = y] \sum_{x \in X(\Omega)} \frac{\Pr[X = x, Y = y]}{\Pr[Y = y]}\, f(x) = \\
&= \sum_{x \in X(\Omega)} f(x) \sum_{y \in Y(\Omega)} \Pr[X = x, Y = y] = \\
&= \sum_{x \in X(\Omega)} f(x) \Pr[X = x] = \mathbb{E}[f(X)]
\end{aligned}$$

Помощни свойства:

(1) $$\mathbb{E}[f_i(\theta)] = \sum_{i=1}^{m} \Pr[i]\, f_i(\theta) = \sum_{i=1}^{m} \frac{1}{m} f_i(\theta) = f(\theta)$$

$$\mathbb{E}_i[f_i(\theta)\,|\,\theta] = \sum_{i=1}^{m} \frac{\Pr[i,\theta]}{\Pr[\theta]}\, f_i(\theta) \overset{\text{незав.}}{=} \sum_{i=1}^{m} \frac{\Pr[i]\Pr[\theta]}{\Pr[\theta]} f_i(\theta) = f(\theta)$$

(2) $$\mathbb{E}[\,\nabla f_i(\theta)] = \nabla \mathbb{E}[f_i(\theta)] = \nabla f(\theta),$$

$$\mathbb{E}_i[\,\nabla f_i(\theta)\,|\,\theta] = \nabla \mathbb{E}_i[f_i(\theta)\,|\,\theta] = \nabla f(\theta)$$

(3)
$$\begin{aligned}
\mathbb{E}[\|\nabla f_i(\theta) - \nabla f(\theta)\|^2] &= \\
&= \mathbb{E}[\|\nabla f_i(\theta)\|^2] - 2\mathbb{E}[\,\nabla f_i(\theta) \cdot \nabla f(\theta)] + \mathbb{E}[\|\nabla f(\theta)\|]^2 = \\
&= \mathbb{E}[\|\nabla f_i(\theta)\|^2] - 2\mathbb{E}[\,\nabla f_i(\theta)] \cdot \nabla f(\theta) + \mathbb{E}[\|\nabla f(\theta)\|]^2 = \\
&= \mathbb{E}[\|\nabla f_i(\theta)\|^2] - \|\nabla f(\theta)\|^2
\end{aligned}$$

(4) $$\mathbb{E}[\|\nabla f_i(\theta) - \nabla f(\theta)\|^2] \le \sigma^2 \Longrightarrow \mathbb{E}[\|\nabla f_i(\theta)\|^2] \le \sigma^2 + \|\nabla f(\theta)\|^2$$

План на лекцията

1. Формалности за курса (5 мин)
2. Спускане по градиент, развиване в ред на Тейлър и операторна норма (15 мин)
3. Сходимост на спускането по градиента (15 мин)
4. Спускане по стохастичен градиент и условно очакване (15 мин)
5. **Сходимост на спускането по стохастичен градиент (20 мин)**
6. Партиден стохастичен градиент и спускане по стохастичен градиент с адаптация и инерционен момент (20 мин)

Сходимост на стандартен стохастичен градиент

- **Теорема:** Нека функцията $f : \mathbb{R}^n \to \mathbb{R}$ е два пъти диференцируема, ограничена отдолу и $f(\theta) = \frac{1}{m}\sum_{i=1}^{m} f_i(\theta)$. Нека съществува $L > 0$, така че $\|\nabla^2 f(\theta)\| \leq L$. Нека за всяко $k = 0,1,2,\ldots$ е дадена случайна величина i_k със стойности в $\{1,2,\ldots,m\}$ с равномерното случайно разпределение $\Pr[i_k = n] = \frac{1}{m}$. Нека съществува $\sigma > 0$ така че за всяко $\theta \in \mathbb{R}^n$ $\mathbb{E}[\|\nabla f_{i_k}(\theta) - \nabla f(\theta)\|^2] \leq \sigma^2$. Тогава съществува редица от скорости $\alpha_k > 0,\ k = 0,1,2,\ldots$, така че ако $\theta_0 \in \mathbb{R}^n$ и $\theta_{k+1} = \theta_k - \alpha_k \nabla f_{i_k}(\theta_k)$, то $\lim_{t\to\infty} \min_{k=0}^{t} \mathbb{E}[\|\nabla f(\theta_k)\|^2] = 0$.

Доказателство на теоремата:

От теоремата на Тейлър получаваме:

$$f(\theta_{k+1}) = f(\theta_k) - \nabla f(\theta_k)^\top (\alpha_k \nabla f_{i_k}(\theta_k)) + \frac{1}{2}(\alpha_k \nabla f_{i_k}(\theta_k))^\top \nabla^2 f(\bar{\theta})(\alpha_k \nabla f_{i_k}(\theta_k))$$

$$f(\theta_{k+1}) \leq f(\theta_k) - \alpha_k \nabla f(\theta_k)^\top \nabla f_{i_k}(\theta_k) + \frac{\alpha_k^2 L}{2} \|\nabla f_{i_k}(\theta_k)\|^2$$

- $\nabla f(\theta_k)^\top \nabla f_{i_k}(\theta_k)$ може да бъде и отрицателно, следователно нямаме никаква гаранция, че стойността на f намалява на всяка стъпка.

- $f(\theta_{k+1})$ зависи от случайните величини θ_k и i_k. Ще подходим **вероятностно** —ще изследваме условното очакването $\mathbb{E}_{i_k}[f(\theta_{k+1}) \mid \theta_k]$.

$$\mathbb{E}_{i_k}[f(\theta_{k+1}) \mid \theta_k] \leq \mathbb{E}_{i_k}\left[f(\theta_k) - \alpha_k \nabla f(\theta_k)^\top \nabla f_{i_k}(\theta_k)] + \frac{\alpha_k^2 L}{2} \|\nabla f_{i_k}(\theta_k)\|^2 \;\middle|\; \theta_k \right]$$

$$\mathbb{E}_{i_k}[f(\theta_{k+1}) \mid \theta_k] \leq f(\theta_k) - \alpha_k \nabla f(\theta_k)^\top \mathbb{E}_{i_k}[\nabla f_{i_k}(\theta_k) \mid \theta_k] + \frac{\alpha_k^2 L}{2} \mathbb{E}_{i_k}[\|\nabla f_{i_k}(\theta_k)\|^2 | \theta_k]$$

$$\mathbb{E}_{i_k}[f(\theta_{k+1}) \mid \theta_k] \leq f(\theta_k) - \alpha_k \|\nabla f(\theta_k)\|^2 + \frac{\alpha_k^2 L}{2} (\sigma^2 + \|\nabla f(\theta_k)\|^2)$$

$$\mathbb{E}_{i_k}[f(\theta_{k+1}) \mid \theta_k] \leq f(\theta_k) - \alpha_k (1 - \frac{\alpha_k L}{2}) \|\nabla f(\theta_k)\|^2 + \frac{\alpha_k^2 L \sigma^2}{2}$$

пак ще подберем α_k, така че: $0 < \alpha_k L < 1 \Rightarrow \left(1 - \frac{\alpha_k L}{2}\right) < \frac{1}{2}$

$$\mathbb{E}_{i_k}[f(\theta_{k+1}) \mid \theta_k] \leq f(\theta_k) - \frac{\alpha_k}{2} \|\nabla f(\theta_k)\|^2 + \frac{\alpha_k^2 L \sigma^2}{2}$$

$$\frac{\alpha_k}{2} \|\nabla f(\theta_k)\|^2 \leq f(\theta_k) - \mathbb{E}_{i_k}[f(\theta_{k+1}) \mid \theta_k] + \frac{\alpha_k^2 L \sigma^2}{2}$$

$$\alpha_k \|\nabla f(\theta_k)\|^2 \leq 2\left(f(\theta_k) - \mathbb{E}_{i_k}[f(\theta_{k+1}) \,|\, \theta_k]\right) + \alpha_k^2 L\sigma^2$$

сумираме горното неравенство за $k = 0,1,2,\ldots,t$:

$$\sum_{k=0}^{t} \alpha_k \|\nabla f(\theta_k)\|^2 \leq 2\sum_{k=0}^{t}\left(f(\theta_k) - \mathbb{E}_{i_k}[f(\theta_{k+1}) \,|\, \theta_k]\right) + L\sigma^2 \sum_{k=1}^{t} \alpha_k^2$$

$$\mathbb{E}\left[\sum_{k=0}^{1} \alpha_k \|\nabla f(\theta_k)\|^2\right] \leq \mathbb{E}\left[2\sum_{k=0}^{t}\left(f(\theta_k) - \mathbb{E}_{i_k}[f(\theta_{k+1}) \,|\, \theta_k]\right) + L\sigma^2 \sum_{k=1}^{1} \alpha_k^2\right]$$

Теорема за пълно очакване

$$\sum_{k=0}^{t} \alpha_k \mathbb{E}[\|\nabla f(\theta_k)\|^2] \leq 2\sum_{k=0}^{t}\left(\mathbb{E}[f(\theta_k)] - \mathbb{E}[f(\theta_{k+1})]\right) + L\sigma^2 \sum_{k=1}^{t} \alpha_k^2$$

$$\min_{k=0}^{t} \mathbb{E}[\|\nabla f(\theta_k)\|^2] \sum_{k=0}^{t} \alpha_k \leq 2(f(\theta_0) - \mathbb{E}[f(\theta_{t+1})]) + L\sigma^2 \sum_{k=1}^{t} \alpha_k^2$$

Нека f^* е долна граница на функцията f:

$$\min_{k=0}^{t} \mathbb{E}[\|\nabla f(\theta_k)\|^2] \leq 2\frac{f(\theta_0) - f^*}{\sum_{k=0}^{t} \alpha_k} + L\sigma^2 \frac{\sum_{k=0}^{t} \alpha_k^2}{\sum_{k=0}^{t} \alpha_k}$$

Ако $\alpha_k = \alpha$, то $\sum_k \alpha_k = \alpha(t+1)$ и $\sum_k \alpha_k^2 = \alpha^2(t+1)$ и грешката е $O(1/t) + O(\alpha)$ — не следва сходимост към 0.

Ако $\alpha_k = \alpha/k$, то $\sum_k \alpha_k = O(\log(t))$ и $\sum_k \alpha_k^2 = O(1)$ и грешката е $O(1/\log(t))$.

Ако $\alpha_k = \alpha/\sqrt{k}$, то $\sum_k \alpha_k = O(\sqrt{t})$ и $\sum_k \alpha_k^2 = O(\log(t))$ и грешката е $O(\log(t)/\sqrt{t})$.■

Свойства на стандартния стохастичен градиент

- Изчисляването на градиента става $|X|$ пъти по-бързо.
- Стойността на f може на някои стъпки да се увеличава.
- Това че $\lim_{t\to\infty} \min_{k=0}^{t} \mathbb{E}[\|\nabla f(\theta_k)\|^2] = 0$ не означава, че редицата $\mathbb{E}[\|\nabla f(\theta_k)\|^2]$ е непременно е сходяща.
- Теоремата казва, че ще достигнем стойност на очакването на нормата на градиента, която е произволно близка до 0.
- Сходимостта зависи от θ_0, α_k, L и σ^2.
- Ако функцията f е изпъкнала, то може да се докаже, че стохастичния градиент е сходящ и клони към глобалния минимум.

План на лекцията

1. Формалности за курса (5 мин)
2. Спускане по градиент, развиване в ред на Тейлър и операторна норма (15 мин)
3. Сходимост на спускането по градиента (15 мин)
4. Спускане по стохастичен градиент и условно очакване (15 мин)
5. Сходимост на спускането по стохастичен градиент (20 мин)
6. **Партиден стохастичен градиент и спускане по стохастичен градиент с адаптация и инерционен момент (20 мин)**

Партиден стохастичен градиент
Batched Stochastic Gradient

Как да намалим вариацията на стохастичния градиент σ^2 ?.

- Разглеждаме **партида** (batch, minibatch) — извадка от b на брой случайни наблюдения с равномерно разпределение от X: $X_B = X_{i_1}, X_{i_2}, \ldots, X_{i_b}$. Тогава ако означим с $f_B(\theta) = \frac{1}{b}\sum_{i=1}^{b} f_{i_b}(\theta_k)$ кросентропията на партидата X_B, то дефинираме спускането по партиден стохастичен градиент като:

- $\theta_{k+1} = \theta_k - \alpha_k \nabla f_{B_k}(\theta_k)$.

- Можем да повторим същите разсъждения като при стандартния стохастичен градиент в случая на партида. Разликите са, че времето за намиране на градиента нараства с фактор b, но за сметка на това **вариацията** на партидния градиент ще **намалее** с фактор b.

Методи за адаптиране на скоростта на спускане по отделните параметри

- Може да има големи разлики в големината на производните за отделните параметри. Затова ще дефинираме скоростта поотделно за всеки параметър: $\eta_k \in \mathbb{R}^n$, $\theta_{k+1} = \theta_k - \eta_k \odot \nabla f_{B_k}(\theta_k)$

- Ще използваме итеративни формули за средно аритметично и експоненциално средно $m_0 = 0$, $\beta \in (0,1)$:

  Средно аритметично: $$m_k = \frac{k-1}{k} m_{k-1} + \frac{1}{k} x_k = \frac{1}{k} \sum_{i=0}^{k-1} x_{k-i}$$

  Екпоненциално средно: $$m_k = \beta m_{k-1} + (1-\beta) x_k = (1-\beta) \sum_{i=0}^{k-1} \beta^i x_{k-i}$$

ADAGrad (диагонален):

$$\begin{cases} (v_k)_i = \frac{k-1}{k}(v_{k-1})_i + \frac{1}{k}(\nabla f_{B_k}(\theta_k))_i^2 \\ (\eta_k)_i = \frac{\gamma_k}{\sqrt{(v_k)_i + \epsilon}} \end{cases}$$

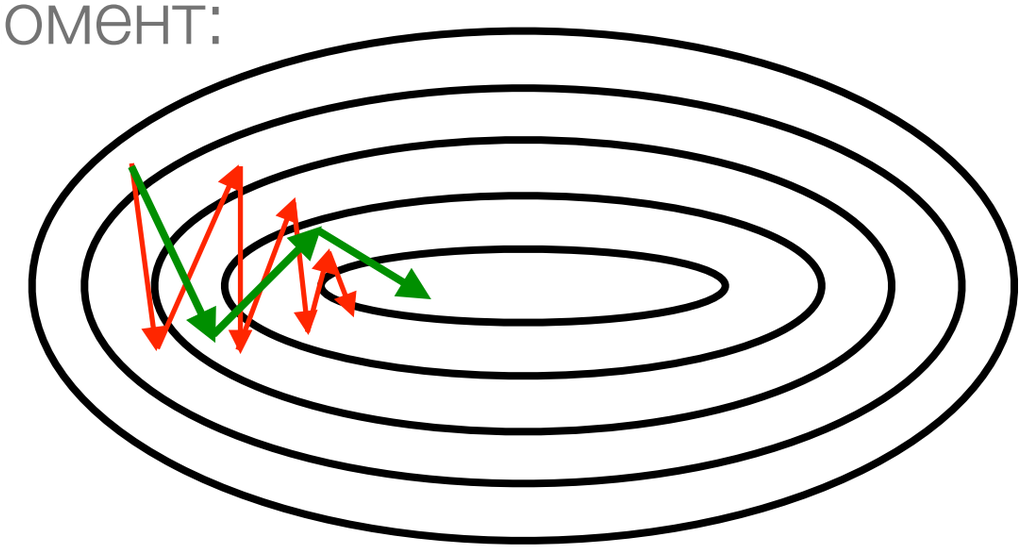
RMSProp:

$$\begin{cases} (v_k)_i = \beta(v_{k-1})_i + (1-\beta)(\nabla f_{B_k}(\theta_k))_i^2 \\ (\eta_k)_i = \frac{\gamma_k}{\sqrt{(v_k)_i + \epsilon}} \end{cases}$$

Спускане по стохастичен градиент с инерционен момент и метода ADAM

- Спускане с класически инерционен момент:

$$\begin{aligned} m_0 &= 0 \\ m_k &= \beta m_{k-1} + (1-\beta)\nabla f_{B_k}(\theta_k) \\ \theta_{k+1} &= \theta_k - \eta_k m_k \end{aligned}$$



- ADAM:

$$\begin{aligned} m_k &= \beta_1 m_{k-1} + (1-\beta_1)\nabla f_{B_k}(\theta_k) \\ (v_k)_i &= \beta_2 (v_{k-1})_i + (1-\beta_2)(\nabla f_{B_k}(\theta_k))_i^2 \\ (\eta_k)_i &= \frac{\gamma_k}{\sqrt{(v_k)_i + \epsilon}} \\ \theta_{k+1} &= \theta_k - \eta_k \odot m_k \end{aligned}$$

Методи за спускане по стохастичен градиент

- Съществуват разнообразни варианти, целящи да подобрят сходимостта и скоростта на сходимост на спускането по стохастичен градиент.

- Изследванията в тази област продължават и методите още търпят развитие.

- За по-задълбочено изучаване:
  Yurii Nesterov: Lectures on Convex Optimization, Springer 2018

- На практика методът ADAM е сред най-често използваните.

Заключение

- Спускането по пълния градиент е гарантирано сходящо, но се изчислява бавно.
- Партидният стохастичен градиент има много висока вероятност за схождане, като е значително по-ефективен от пълния градиент.
- Партидният стохастичен градиент (и неговите вариации) е дефакто стандартният подход в съвременните системи за дълбоко машинно обучение.