$f(x) = |x|$

⊕ выпукла
⊖ не L-гладкая
⊖ не дифф.

Вместо L-гладкости

Определение M-Липшецевой функции

Пусть дана функция $f : \mathbb{R}^d \to \mathbb{R}$. Будем говорить, что она является M-Липшицева, если для любых $x, y \in \mathbb{R}^d$ выполнено

$$|f(x) - f(y)| \leq M\|x - y\|_2.$$

Вместо градиента

Субградиент и субдифференциал

Пусть дана выпуклая функция $f : \mathbb{R}^d \to \mathbb{R}$. Вектор g будем называть субградиентом этой функции f в точке $x \in \mathbb{R}^d$, если для любого $y \in \mathbb{R}^d$ выполняется:

$$f(y) \geq f(x) + \langle g, y - x \rangle.$$

Множество $\partial f(x)$ всех субградиентов f в x будем называть субдифференциалом.

NB
$g = \nabla f$
выпуклость

Теорема (условие оптимальности)

x^* – минимум выпуклой функции f тогда и только тогда, когда

$$0 \in \partial f(x^*).$$

$\partial f(0) = [-1; 1]$

Док-во:

$\Rightarrow \quad f(x) \geq f(x^*) \quad \forall x \in \mathbb{R}^d$

$f(x) \geq f(x^*) + \langle 0; x - x^* \rangle \quad \forall x \in \mathbb{R}^d$

по опр субград $\quad 0 \in \partial f(x^*)$

$\Leftarrow \quad 0 \in \partial f(x^*)$ по выпуклости и опр субд.

$f(x) \geq f(x^*) + \langle g; x - x^* \rangle = f(x^*) \quad \forall x \in \mathbb{R}^d$

$g = 0$

Лемма (свойство M-Липшицевой функции)

Пусть дана выпуклая функция $f : \mathbb{R}^d \to \mathbb{R}$. Тогда функция f является M-Липшицевой тогда и только тогда, когда для любого $x \in \mathbb{R}^d$ и $g \in \partial f(x)$ имеем $\|g\|_2 \leq M$.

Док-во:

$\Rightarrow$ f выпукла и M-липшицева

по выпуклости и опред субград.

$f(y) - f(x) \geq \langle g; y - x \rangle \quad \forall g \in \partial f(x)$

M-липшицевость

$M\|x - y\|_2 \geq f(y) - f(x) \geq \langle g; y - x \rangle$

$y = x + g$

$M\|g\|_2 \geq \|g\|_2^2 \Rightarrow \|g\|_2 \leq M$

$\Leftarrow$ f выпукла $\forall g \in \partial f(x)$ $\|g\|_2 \leq M$

по выпуклости и по опред субград

$f(y) - f(x) \geq \langle g; y - x \rangle$

$f(x) - f(y) \leq \langle g; x - y \rangle$

КБШ

$f(x) - f(y) \leq \underbrace{\|g\|_2}_{M} \cdot \|x - y\|_2$

$|f(x) - f(y)| \leq M \cdot \|x - y\|_2$ ∎

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

Идея: в ур спуске брать субградиент

Алгоритм 2 Субградиентный метод

Вход: размеры шага $\gamma > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: $x^{k+1} = x^k - \gamma g^k$
4: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K-1} x^k$

Док-во:

- f выпукла
- f M- липшицева

$$\|x^{k+1}-x^*\|_2^2 = \|x^k - \gamma g^k - x^*\|_2^2$$
$$= \|x^k-x^*\|_2^2 - 2\gamma\langle g^k; x^k-x^*\rangle + \gamma^2\|g^k\|_2^2$$

M- липшицевость $\|g^k\|_2 \le M$

$$\le \|x^k-x^*\|_2^2 - 2\gamma\langle g^k; x^k-x^*\rangle + \gamma^2 M^2$$

$$2\gamma\langle g^k; x^k-x^*\rangle \le \|x^k-x^*\|_2^2 - \|x^{k+1}-x^*\|_2^2 + \gamma^2M^2$$

выпуклость

$$2\gamma\,(f(x^k)-f(x^*)) \le \|x^k-x^*\|_2^2 - \|x^{k+1}-x^*\|_2^2 + \gamma^2M^2$$

$\sum$ по всем итерациям

$$2\gamma\cdot\frac{1}{k}\sum_{k=0}^{k-1}(f(x^k)-f(x^*)) \le \frac{1}{k}\sum_{k=0}^{k-1}\|x^k-x^*\|_2^2 - \|x^{k+1}-x^*\|_2^2 + \gamma^2M^2$$

$$\|x^0-x^*\|_2^2 - \|x^k-x^*\|_2^2$$

$$\le \frac{\|x^0-x^*\|_2^2}{k} + \gamma^2M^2$$

0) в марте сразу $f(x^k) \le f(x^{k-1})$

сейчас нельзя

1) $\min\limits_k f(x^k)$

2) $\sum\limits_{k=0}^{k-1}\frac{1}{k}f(x^k)$

$\hat{x}$ - с.в. $\{x^k; p=\frac{1}{k}\}$

$\mathbb{E}f(\hat{x})$ - м.о выпуклой ф.

Йенсен $\mathbb{E}f(\hat{x}) \ge f(\mathbb{E}\hat{x})$

$$\sum_{k=0}^{k-1}\frac{1}{k}f(x^k) \ge f\Big(\underbrace{\frac{1}{k}\sum_{k=0}^{k-1}x^k}_{\text{средняя точка}}\Big)$$

$$2\gamma\,(f(\hat{x}^k)-f(x^*)) \le \frac{\|x^0-x^*\|_2^2}{k} + \gamma^2M^2 \quad |\cdot\gamma$$

$$f(\hat{x}^k)-f(x^*) \le \frac{\|x^0-x^*\|_2^2}{2\gamma k} + \frac{\gamma M^2}{2}$$

min требуемое число ш

$$\gamma_{opt} = \frac{\|x^0 - x^*\|}{M\sqrt{K}} \qquad \frac{\|x^0 - x^*\|_2^2}{2\gamma_{opt}^2 K} = \frac{M^2}{2}$$

$$f(\hat{x}^K) - f(x^*) \le \frac{M \cdot \|x^0 - x^*\|_2}{\sqrt{K}}$$

Теорема сходимость субградиентного спуска для M-Липшицевых и выпуклых функций

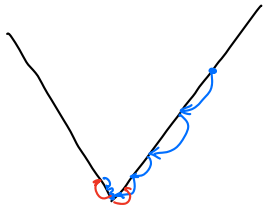
Пусть задача безусловной оптимизации с M-Липшицевой, выпуклой целевой функцией f решается с помощью субградиентного спуска. Тогда справедлива следующая оценка сходимости

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{M\|x^0 - x^*\|_2}{\sqrt{K}}$$

Более того, чтобы добиться точности ε по функции, необходимо

$$K = O\left(\frac{M^2\|x^0 - x^*\|_2^2}{\varepsilon^2}\right) \text{ итераций.}$$

Рычаги сходимости:

- $\gamma \sim \frac{1}{\sqrt{K}} \Rightarrow \gamma_k \sim \frac{1}{\sqrt{k}}$
- 

1) уменьш. шаг = меньш скор. (меньший отклик)

2) последняя точка – не обяз хорошая

⊕ проще, чем град спуск

⊕ оптимален для этого класса задач

⊖ подбор шага ($\|x^0 - x^*\|_2$, M, K)

$$\gamma = \frac{\|x^0 - x^*\|_2}{M\sqrt{K}}$$

← расст. до реш.

↑ M-ма ↑ кол. итер.

AdaGrad Norm – адаптивность под кол. св-ва M-ммы.

1) $\gamma_k = \frac{\|x^0 - x^*\|_2}{M\sqrt{k}} = \frac{\|x^0 - x^*\|_2}{\sqrt{k \cdot M^2}}$

2) Идея $\|g_k\| \le M$

$\gamma_k = \frac{\|x^0 - x^*\|_2^2}{\sqrt{\sum_{t=0}^{k} \|g^t\|_2^2}}$, $\sum_{t=0}^{k} \|g^t\|_2^2 \le k M^2$

3) $\gamma_k = \frac{D}{\sqrt{\sum_{t=0}^{k} \|g^t\|_2^2}}$ $\|x^0 - x^*\| \le D$

4) $\gamma_k = \frac{D}{\sqrt{\sum \|g^t\|_2^2 + e}}$ ← не делить на 0

$e > 0$ $e \sim 10^{-6} - 10^{-8}$

Алгоритм 3 AdaGradNorm

Вход: $D > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов норм градиентов $G^0 = 0$, параметр сглаживания e = 1e-8, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: Вычислить $G^{k+1} = G^k + \|g^k\|_2^2$
4: $x^{k+1} = x^k - \frac{D}{\sqrt{G^{k+1}+e}} g^k$
5: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K} x^k$

AdaGradNorm => AdaGrad адаптировать по коорд.

$\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k} (g_i^t)^2 + e}}$ шаг для коорд. i

Алгоритм 4 AdaGrad

Вход: $D_i > 0$, стартовая точка $x^0 \in \mathbb{R}^d$, сумма квадратов градиентов $G_i^0 = 0$, параметр сглаживания e = 1e-8, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: Для каждой координаты: $G_i^{k+1} = G_i^k + (g_i^k)^2$
4: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1}+\varepsilon}} g_i^k$
5: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K} x^k$

Док-во:

- f – выпуклая
- f – M-гладкая

$$|x_i^{k+1} - x_i^*|^2 = |x_i^k - \gamma_{k,i} g_i^k - x_i^*|^2$$

$$= |x_i^k - x_i^*|^2 - 2\gamma_{k,i} g_i^k \cdot (x_i^k - x_i^*) + \gamma_{k,i}^2 (g_i^k)^2$$

Разделим на $2\gamma_{k,i}$

$$g_i^k \cdot (x_i^k - x_i^*) = \frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i} (g_i^k)^2}{2}$$

сумм по всем коорд i

$$\langle g^k, x^k - x^* \rangle = \sum_{i=1}^d \left[\frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i} (g_i^k)^2}{2} \right]$$

выпуклость и субград

$$f(x^k) - f(x^*) \le \sum_{i=1}^d \left[\frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i} (g_i^k)^2}{2} \right]$$

$\frac{1}{K} \sum_{k=0}^{K-1}$

$$f(\hat{x}^K) - f(x^*) \le \frac{1}{K} \sum_{k=0}^{K-1} \sum_{i=1}^d \left[\frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i} (g_i^k)^2}{2} \right]$$

$$\sum_k \sum_d = \sum_d \sum_k$$

$$\le \frac{1}{K} \sum_{i=1}^d \sum_{k=0}^{K-1} \left[\frac{|x_i^k - x_i^*|^2 - |x_i^{k+1} - x_i^*|^2}{2\gamma_{k,i}} + \frac{\gamma_{k,i} (g_i^k)^2}{2} \right]$$

сгруппируем по $|x_i^k - x_i^*|^2$

$$\le \frac{1}{K} \sum_{i=1}^d \sum_{k=1}^{K-1} \left[\left(\frac{1}{2\gamma_{k,i}} - \frac{1}{2\gamma_{k-1,i}} \right) |x_i^k - x_i^*|^2 \right]$$

$$+ \frac{1}{K} \sum_{i=1}^d \left[\frac{1}{2\gamma_{0,i}} |x_i^0 - x_i^*|^2 - \frac{1}{2\gamma_{K-1,i}} |x_i^K - x_i^*|^2 \right]$$

$$+ \frac{1}{K} \sum_{i=1}^d \sum_{k=0}^{K-1} \frac{\gamma_{k,i} (g_i^k)^2}{2}$$

$\gamma_{-1,i} = +\infty$

$$\leq \frac{1}{k}\sum_{i=1}^{d}\sum_{k=0}^{k-1}\left[\left(\frac{1}{2\gamma_{k,i}}-\frac{1}{2\gamma_{k-1,i}}\right)|x_i^k-x_i^*|^2\right]$$

$$+\frac{1}{k}\sum_{i=1}^{d}\sum_{k=0}^{k-1}\frac{\gamma_{k,i}(g_i^k)^2}{2}$$

$$\boxed{|x_i^k-x_i^*|\leq D_i} \quad \leftarrow \text{предположение}$$

$$\leq \frac{1}{k}\sum_{i=1}^{d}\sum_{k=0}^{k-1}\left[\left(\frac{1}{2\gamma_{k,i}}-\frac{1}{2\gamma_{k-1,i}}\right)D_i^2\right]$$

$$+\frac{1}{k}\sum_{i=1}^{d}\sum_{k=0}^{k-1}\frac{\gamma_{k,i}(g_i^k)^2}{2}$$

$$\gamma_{k,i}=\frac{D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}$$

$$\leq \frac{1}{2k}\sum_{i=1}^{d}\sum_{k=0}^{k-1}\left[\left(\sqrt{\sum_{t=0}^{k}(g_i^t)^2}-\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}\right)D_i\right]$$

$$+\frac{1}{2k}\sum_{i=1}^{d}\sum_{k=0}^{k-1}\frac{(g_i^k)^2 D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}$$

$$=\frac{1}{2k}\sum_{i=1}^{d}D_i\sum_{k=0}^{k-1}\left[\sqrt{\sum_{t=0}^{k}(g_i^t)^2}-\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}\right]$$

$$+\frac{1}{2k}\sum_{i=1}^{d}D_i\sum_{k=0}^{k-1}\frac{(g_i^k)^2}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}$$

$$=\frac{1}{2k}\sum_{i=1}^{d}D_i\cdot\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}$$

$$+\frac{1}{2k}\sum_{i=1}^{d}D_i\sum_{k=0}^{k-1}\frac{(g_i^k)^2}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}}$$

$\{a_k\}$ $\quad \sum_{k=0}^{k-1}\frac{(a_k)^2}{\sqrt{\sum_{t=0}^{k}(a_t)^2}}\leq 2\sqrt{\sum_{k=0}^{k-1}(a_k)^2}$

$$\leq \frac{1}{2k}\sum_{i=1}^{d}D_i\cdot\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}$$

$$+\frac{1}{2k}\sum_{i=1}^{d}D_i\cdot 2\sqrt{\sum_{t=0}^{k-1}(g_i^t)^2}$$

$$= \frac{3}{2K} \sum_{i=1}^{d} D_i \sqrt{\sum_{t=0}^{K-1} (g_i^t)^2}$$

$(g_i^t)^2 \le M^2$

$$\le \frac{3}{2K} \cdot \sum_{i=1}^{d} D_i \cdot \sqrt{M^2 K}$$

$$= \frac{3M}{2\sqrt{K}} \underbrace{\sum_{i=1}^{d} D_i}_{\tilde{D}}$$

$$= \frac{3M\tilde{D}}{\sqrt{K}}$$

субЛинейная сход $\frac{1}{\sqrt{K}}$

Theorem

Пусть задача оптимизации с M-Липшицевой, выпуклой целевой функцией f решается с помощью AdaGrad на ограниченном множестве. Тогда справедлива следующая оценка сходимости:

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{3M\tilde{D}}{2\sqrt{K}},$$

где $\tilde{D} = \sum_{i=1}^{d} D_i$.
Более того, чтобы добиться точности ε по функции, необходимо

$$K = O\left(\frac{9M^2\tilde{D}^2}{4\varepsilon^2}\right) \text{ итераций.}$$

AdaGrad => RMSProp

$$\gamma_{k,i} = \frac{D_i}{\sqrt{\sum_{t=0}^{k}(g_i^t)^2}} \Rightarrow \frac{D_i}{\sqrt{h_i^k}}$$

$\beta_2 = 0{,}99$, $\in (0,1)$

идея пересчета: $h_i^k = \beta_2 h_i^{k-1} + (1-\beta_2)(g_i^k)^2$

= забываем старое значение

Алгоритм 5 RMSProp

Вход: шаг $D_i > 0$, параметр сглаживания $\beta_2 \in [0,1]$, стартовая точка $x^0 \in \mathbb{R}^d$, сглаженная сумма квадратов градиентов $G_i^0 = 0$, параметр сглаживания $\varepsilon =$ 1e-8, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: Для каждой координаты: $G_i^{k+1} = \beta_2 G_i^k + (1-\beta_2)(g_i^k)^2$
4: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1}+\varepsilon}} g_i^k$
5: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K} x^k$

RMSProp => Adam

$\beta_2 = const \Rightarrow \beta_{2,k}$

RMSProp: $h_i^k = \beta_2 h_i^{k-1} + (1-\beta_2)(g_i^k)^2$

$h_i^0 = 0 \quad h_i^1 = 0 + (1-\beta_2)(g_i^1)^2$

$\neq 1$

$\sum$ коэф. перед g в RMSProp $\neq 1$

$g_i^k \quad g_i^{k-1} \quad \ldots$

$1-\beta_2 \quad (1-\beta_2)\beta_2 \quad \ldots \leftarrow \sum \neq 1$

Adam $\sum = 1$ $\quad \beta_2 = \beta_{2,k}$

$$\sum_{t=0}^{k} (1-\beta_2)\beta^t = \frac{(1-\beta_2)^2}{(1-\beta_2)^{k+1}} \leftarrow \text{корректировка}$$

Алгоритм 6 Adam

Вход: шаг $D_i > 0$, параметры сглаживания $\beta_1 = 0.9$ и $\beta_2 = 0.99$, стартовая точка $x^0 \in \mathbb{R}^d$, сглаженная сумма квадратов градиентов $G_i^0 = 0$, сглаженная сумма градиентов $v^0 = 0$, параметр сглаживания $\varepsilon =$ 1e-8, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: Вычислить $g^k \in \partial f(x^k)$
3: Вычислить $v^{k+1} = \beta_1 v^k + (1-\beta_1) g^k$
4: Для каждой координаты: $G_i^{k+1} = \beta_2 G_i^k + (1-\beta_2)(g_i^k)^2$
5: Для каждой координаты: $x_i^{k+1} = x_i^k - \frac{D_i}{\sqrt{G_i^{k+1}+\varepsilon}} v_i^{k+1}$
6: **end for**

Выход: $\frac{1}{K}\sum_{k=0}^{K} x^k$

⊕ простота идеи.

⊖ доп. память

⊕ SOTA для обучения нейросетей