На прошлой лекции

градиентный спуск $\|x^k - x^*\|_2^2 \sim \varepsilon$

$$\frac{L}{\mu} \log\left(\frac{\|x^0 - x^*\|_2^2}{\varepsilon}\right)$$ итераций / оракульн. слож.

а можно лучше?

1964 г. - Б.Т. Поляк

Алгоритм 2 Метод тяжелого шарика

Вход: размер шагов $\{\gamma_k\}_{k=0} > 0$, моментумы $\{\tau_k\}_{k=0} \in [0;1]$, стартовая точка $x^0 = x^{-1} \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

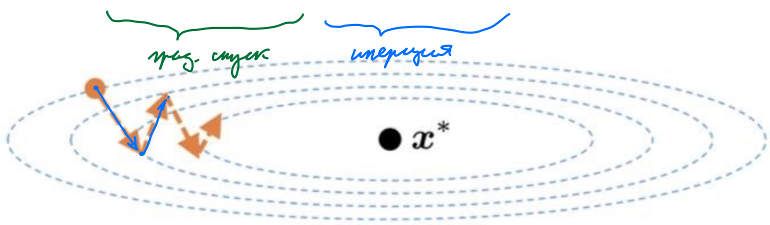
2: Вычислить $\nabla f(x^k)$

3: $x^{k+1} = x^k - \gamma_k \nabla f(x^k) + \tau_k (x^k - x^{k-1})$

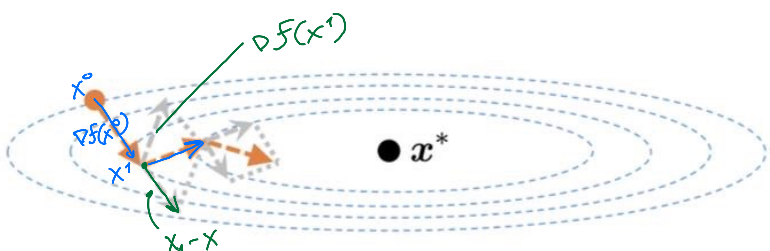
4: **end for**

Выход: x^K

Физика:



gradient descent



heavy-ball method

Пример: в pytorch

$$v^{k+1} = \beta v^k + \nabla f(x^k)$$
$$x^{k+1} = x^k - \gamma v^{k+1}$$

$\beta \in [0;1)$

во второе подст. первое

$$x^{k+1} = x^k - \gamma \beta \underline{v}^k - \gamma \nabla f(x^k)$$

на k шаге v^k: $x^k = x^{k-1} - \gamma v^k$

$$-\gamma v^k = x^k - x^{k-1}$$

$$x^{k+1} = x^k + \beta(x^k - x^{k-1}) - \gamma \nabla f(x^k)$$

(+) понятная физика и интуиция

(+) легко имплементировать

(+) дешевизна вычислений

(−) два параметра: γ_k и τ_k $\boxed{\tau_k \in [0,9; 0,95)}$

(−) неоптим. сходимость

(−) метод не лучше, чем град. спуск

1983 г. — Ю.Е. Нестеров

Алгоритм 3 Ускоренный градиентный метод

Вход: размер шагов $\{\gamma_k\}_{k=0} > 0$, моментумы $\{\tau_k\}_{k=0} \in [0;1]$, стартовая точка $x^0 = y^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \dots, K-1$ **do**

2: Вычислить $\nabla f(y^k)$

3: $x^{k+1} = y^k - \gamma_k \nabla f(y^k)$

4: $y^{k+1} = x^{k+1} + \tau_k(x^{k+1} - x^k)$

5: **end for**

Выход: x^K

Тяжелый шарик:

$$x^{k+1} = x^k - \gamma \nabla f(x^k) + \tau(x^k - x^{k-1})$$

Метод Нестерова:

$$x^{k+1} = y^k - \gamma \nabla f(y^k) \quad | \quad y^k = x^k + \tau(x^k - x^{k-1})$$

$$x^{k+1} = x^k + \underbrace{\tau(x^k - x^{k-1})}_{\text{инерция}} - \gamma \nabla f(x^k \underbrace{+ \tau(x^k - x^{k-1})}_{\text{инерция в точке подсчета}})$$

Другой ускоренный метод:

Алгоритм 5 Линейный каплинг: внутренний цикл

Вход: размер шагов $\gamma > 0$ и $\eta > 0$, моментум $\tau \in [0;1]$, стартовая точка $x^0 = y^0 = z^0 \in \mathbb{R}^d$, количество итераций K

1: **for** $k = 0, 1, \dots, K-1$ **do**

2: Вычислить $\nabla f(x^k)$

3: $y^{k+1} = x^k - \eta \nabla f(x^k)$ ← „медленный" $\eta = \frac{1}{L}$ град спуск

4: $z^{k+1} = z^k - \gamma \nabla f(x^k)$ ← „быстрый" $\gamma = \frac{1}{\sqrt{\mu L}}$

5: $x^{k+1} = \tau z^{k+1} + (1-\tau) y^{k+1}$

6: **end for**

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

О сходимости линейного каплинга

Пусть задача безусловной оптимизации с L-гладкой, μ-сильно выпуклой целевой функцией f решается с помощью реставрированного линейного каплинга. Тогда при $\eta = \frac{1}{L}$, $\gamma = \sqrt{\frac{1}{\mu L}}$ и $K = \sqrt{\frac{16L}{\mu}}$, чтобы добиться точности ε по функции ($f(x) - f(x^*) \le \varepsilon$), необходимо

для метода Нестерова оптимально → $$O\left(\sqrt{\frac{L}{\mu}} \log \frac{f(x^0) - f(x^*)}{\varepsilon}\right)$$ вызовов оракула. лучше, чем

Град спуск: $\frac{L}{\mu} \log \frac{f(x^0) - f(x^*)}{\varepsilon}$ порядок вызовов

А можно ли еще лучше?

нижние оценки

Нужен класс алгоритмов:

1) x^0 – стартовая точка $\qquad M = \{x^0\}$ ← „память"

2) $\nabla f(x')$, где $x' \in M$

добавлять в M могу

$$x \in \mathrm{span}\{x'', \nabla f(x')\} \qquad x' \in M,\ x'' \in M$$

$$M = \mathrm{span}\{x, M\}$$

3) выход за k шагов это $x_{out} \in M$

- изученные методы здесь лежат
- далеко не все алгоритмы учтены

<u>Нижние оценки</u> для этого класса

„Плохая" функция $\qquad x \in \mathbb{R}^d$ ← размерность

$$f(x) = \frac{L-\mu}{8}x^T A x + \frac{\mu}{2}x^T x - \frac{L-\mu}{4}e_1^T x,$$

$$A = \begin{pmatrix} 2 & -1 & & 0 & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & 0 & & -1 & \zeta \end{pmatrix}, \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

0) f – μ-сильно выпуклая и L-гладкая

1) Решение f: $\arg\min\limits_{x \in \mathbb{R}^d} f(x)$

$$\nabla f(x^*) = 0$$

$$\left(\frac{L-\mu}{4} A + \mu I\right) x^* - \frac{L-\mu}{4} e_1 = 0$$

$$A x^* + \frac{4\mu}{L-\mu} x^* - e_1 = 0$$

Распишем по координатам:

$$2x_1^* - x_2^* + \frac{4\mu}{L-\mu}x_1^* - 1 = 0$$ ← 1 строка

$$-x_{k-1}^* + \frac{2(L+\mu)}{L-\mu}x_k^* - x_{k+1}^* = 0$$ ← все кроме 1 и посл.

для последней

рекуррента 2 порядка:

$$-1 + \frac{2(L+\mu)}{L-\mu}\lambda - \lambda^2 = 0 \qquad x_k^* = C_1\lambda_1^k + C_2\lambda_2^k$$

NB f задана так, что

$$x_k^* = \lambda_1^k \qquad \lambda_1 = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$$ ← один из корней

2) $x^0 = 0$ ← точка $\quad M = \{0\}$

$$\nabla f(x) = \frac{L-\mu}{4}Ax + \mu x - \frac{L-\mu}{4}e_1$$

$$\nabla f(x^0) \in \operatorname{span}\{e_1\}$$

за один подсчет град $+1$ ненулевая коорд.

за каждый подсчет градиента $+1$ ненул. коор. в M

Объединяем: $d = 2K \quad$ K – кол-во вызовов ∇

за K вызов в лучшем случае $M = \operatorname{span}\{e_1 \dots e_K\}$

в лучшем случае K первых в x^* угадаем идеально, а все остальные вернем 0

$$\|x^K - x^*\|_2^2 \geq \sum_{i=K+1}^{2K}\lambda_1^{2i} = \lambda^{2K}\sum_{i=1}^{K}\lambda_1^{2i} =$$

выход

не угадали $K+1$ до $2K$

$$\|x^0 - x^*\|_2^2 = \|x^*\|_2^2 = \sum_{i=1}^{2K}\lambda_1^{2i} = (1+\lambda_1^{2K})\sum_{i=1}^{K}\lambda_1^{2i}$$

$$= \frac{\lambda_1^{2k}}{\lambda_1^{2k}+1}\|x^0 - x^*\|_2^2 \geq \frac{\lambda_1^{2k}}{2}\|x^0 - x^*\|_2^2$$

$$= \left(1 - \frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^{2k} \frac{\|x^0 - x^*\|_2^2}{2}$$

Нижняя оценка на оракульную сложность

Для любого метода из класса, описанного выше, существует безусловная задача оптимизации с L-гладкой, μ-сильно выпуклой целевой функцией f такая, что для решения этой задачи методу необходимо

$$\Omega\left(\sqrt{\frac{L}{\mu}}\log\frac{\|x^0 - x^*\|_2}{\varepsilon}\right) \text{ вызовов оракула.}$$

(+) метод Нестерова / линейный случай оптимален и неулучшаем (для сильно выпуклых и выпуклых) (NB в классе span)

(+) плюсы тяжелого шарика

(−) немонотон. сходимость

(−) два параметра