В прошлый раз

GD сходится за

$$\frac{L}{\mu}\log\frac{\|x^0-x^*\|_2^2}{\varepsilon}$$ итераций / градиентных вызовов

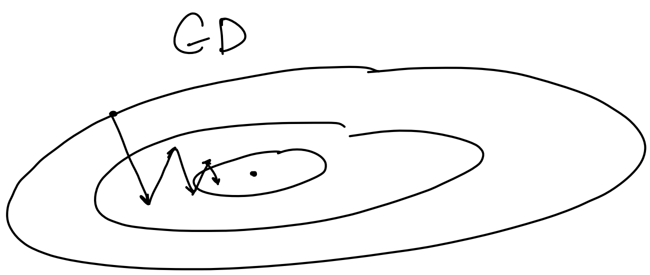
до точности $\varepsilon \sim \|x^k - x^*\|_2^2$

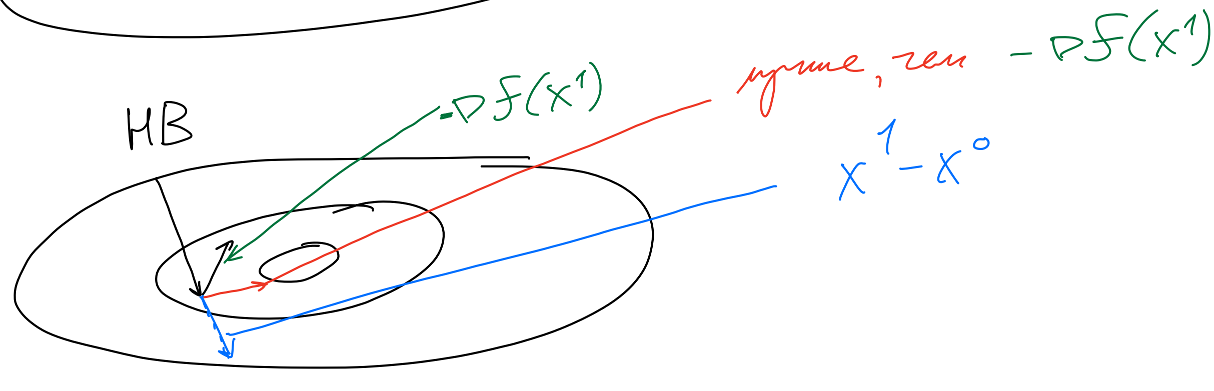
1964 г. Б.Т. Поляк

Метод тяжелого шарика (HB):

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

момент инерции
(иногд. момент
моментум)





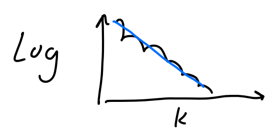
Плюсы:

- физика инерции
- легко кодить (как у GD)
- дешевизна выч.

Минусы:

$\tau_k \in (0{,}9; 1)$

- подбирать γ_k и τ_k
- немонот сходимость



- а лучше GD? нет

1983 г. Ю.Е. Нестеров

Ускоренный градиентный спуск (AGD/Nesterov)

$$\boxed{\begin{aligned} x^{k+1} &= y^k - \gamma_k \nabla f(y^k) \\ y^{k+1} &= x^{k+1} + \tau_k (x^{k+1} - x^k) \end{aligned}}$$

$$y^k = x^k + \tau_{k-1}(x^k - x^{k-1})$$

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

Nesterov: $\quad x^{k+1} = x^k - \gamma_k \nabla f(x^k + \tau_{k-1}(x^k - x^{k-1})) + \tau_{k-1}(x^k - x^{k-1})$

Линейный каплинг (другой ускоренный метод):

$$\boxed{\begin{aligned} y^{k+1} &= x^k - \eta \nabla f(x^k) \\ z^{k+1} &= z^k - \gamma \nabla f(x^k) \\ x^{k+1} &= \tau z^{k+1} + (1-\tau) y^{k+1} \end{aligned}}$$

← медленная динамика GD $\left(\eta \sim \frac{1}{L}\right)$

← быстрая динамика $\left(\gamma \sim \frac{1}{\sqrt{\mu L}}\right)$

← связь динамик

Предположения: f L-гладкая и μ-сильно выпуклая

$$\frac{\mu}{2}\|x-y\|_2^2 \le f(y) - f(x) - \langle \nabla f(x); y-x\rangle \le \frac{L}{2}\|y-x\|_2^2$$

Док-во сходимости:

$$\|z^{k+1} - x^*\|_2^2 = \|z^k - \gamma \nabla f(x^k) - z^*\|_2^2$$

$$= \|z^k - z^*\|_2^2 - 2\gamma \langle \nabla f(x^k); z^k - x^*\rangle + \gamma^2 \|\nabla f(x^k)\|_2^2$$

в док-ве GD $\quad -\langle \nabla f(x^k); x^k - x^*\rangle$ оценивали с μ-сильной вып

$$= \|z^k - z^*\|_2^2 - 2\gamma \langle \nabla f(x^k); x^k - x^*\rangle$$

$$- 2\gamma \langle \nabla f(x^k); z^k - x^k\rangle + \gamma^2 \|\nabla f(x^k)\|_2^2 \qquad (1)$$

L – гладкость, но с использ. y:

$$f(y^{k+1}) \le f(x^k) + \langle \nabla f(x^k); y^{k+1}-x^k \rangle + \frac{L}{2}\|y^{k+1}-x^k\|_2^2$$

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

$$= f(x^k) + \langle \nabla f(x^k); -\eta \nabla f(x^k) \rangle + \frac{L}{2}\|\eta \nabla f(x^k)\|_2^2$$

$$= f(x^k) - \eta\left(1-\frac{L\eta}{2}\right)\|\nabla f(x^k)\|_2^2$$

$$\eta\left(1-\frac{L\eta}{2}\right)\|\nabla f(x^k)\|_2^2 \le f(x^k) - f(y^{k+1})$$

$\eta \in \left(0; \frac{2}{L}\right)$

$$\|\nabla f(x^k)\|_2^2 \le \frac{2}{\eta(2-L\eta)}\left(f(x^k) - f(y^{k+1})\right) \qquad (2)$$

Соединяем (1) и (2)

$$\|z^{k+1}-x^*\|_2^2 \le \|z^k - z^*\|_2^2 - 2\gamma\langle \nabla f(x^k); x^k - x^* \rangle - 2\gamma\langle \nabla f(x^k); z^k - x^k \rangle + \frac{2\gamma^2}{\eta(2-L\eta)}\left(f(x^k) - f(y^{k+1})\right) \qquad (3)$$

Осталось $-\langle \nabla f(x^k); z^k - x^k \rangle$, для него 3 строку алгоритма

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

$$\langle \nabla f(x^k); x^k - z^k \rangle = \langle \nabla f(x^k); \underbrace{x^k - \frac{1}{\tau}\left(x^k - (1-\tau)y^k\right)}_{\frac{(1-\tau)x^k}{\tau}} \rangle$$

$$= \frac{(1-\tau)}{\tau}\langle \nabla f(x^k); y^k - x^k \rangle$$

Выпуклость

$$\le \frac{1-\tau}{\tau}\left(f(y^k) - f(x^k)\right) \qquad (4)$$

Объединяем (3) и (4)

$$\|z^{k+1} - x^*\|_2^2 \le \|z^k - z^*\|_2^2 - 2\gamma \langle \nabla f(x^k); x^k - x^* \rangle + 2\gamma \frac{1-\tau}{\tau} \left(f(y^k) - f(x^k)\right) + \frac{2\gamma^2}{\eta(2-L\eta)} \left(f(x^k) - f(y^{k+1})\right)$$

$$\frac{1-\tau}{\tau} = \frac{\gamma}{\eta(2-L\eta)}$$ (подбор τ)

$$= \|z^k - z^*\|_2^2 - 2\gamma \langle \nabla f(x^k); x^k - x^* \rangle \quad \frac{2\gamma^2}{\eta(2-L\eta)} \left(f(y^k) - f(y^{k+1})\right)$$

$$\langle \nabla f(x^k); x^k - x^* \rangle \le \frac{\|z^k - x^*\|_2^2 - \|z^{k+1} - x^*\|_2^2}{2\gamma} + \frac{\gamma}{\eta(2-L\eta)} \left(f(y^k) - f(y^{k+1})\right)$$

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

$$\sum_{k=0}^{K-1} \langle \nabla f(x^k); x^k - x^* \rangle \le \frac{\|z^0 - x^*\|_2^2 - \|z^K - x^*\|_2^2}{2\gamma} + \frac{\gamma}{\eta(2-L\eta)} \left(f(y^0) - f(y^K)\right)$$

$\|z^K - x^*\|_2^2 \ge 0 \qquad f(y^K) \ge f(x^*)$

$$\le \frac{\|z^0 - x^*\|_2^2}{2\gamma} + \frac{\gamma}{\eta(2-L\eta)} \left(f(y^0) - f(x^*)\right)$$

Выпуклость и $\frac{1}{k}$ (усреднение):

$$\frac{1}{k}\sum_{k=0}^{k-1} f(x^k) - f(x^*) \le \frac{\|z^0 - x^*\|_2^2}{2\gamma k} + \frac{\gamma}{\eta(2-L\eta)k}\left(f(y^0) - f(x^*)\right)$$

Инициализация: $y^0 = z^0 = x^0$

$$\frac{1}{k}\sum_{k=0}^{k-1} \left(f(x^k) - f(x^*)\right) \le \frac{\|x^0 - x^*\|_2^2}{2\gamma k} + \frac{\gamma}{\eta(2-L\eta)k}\left(f(x^0) - f(x^*)\right)$$

Йенсен:

$$f\left(\frac{1}{k}\sum_{k=0}^{k-1} x^k\right) - f(x^*) \le \frac{\|x^0 - x^*\|_2^2}{2\gamma k} + \frac{\gamma}{\eta(2-L\eta)k}\left(f(x^0) - f(x^*)\right)$$

Сублинейно ? $\frac{1}{k}$?

μ - сильная выпуклость $\quad \frac{\mu}{2}\|x^0 - x^*\|_2^2 \le f(x^0) - f(x^*)$.

$$f\left(\frac{1}{k}\sum_{k=0}^{k-1} x^k\right) - f(x^*) \le \left(\frac{1}{\mu\gamma k} + \frac{\gamma}{\eta(2-L\eta)k}\right)\left(f(x^0) - f(x^*)\right)$$

Оптимизируем η : $\eta = \frac{1}{L}$

$$= \left(\frac{1}{\mu\gamma k} + \frac{\gamma L}{k}\right)\left(f(x^0) - f(x^*)\right)$$

Оптимизируем γ : $\gamma = \frac{1}{\sqrt{\mu L}}$

$$= \sqrt{\frac{4L}{\mu k^2}}\left(f(x^0) - f(x^*)\right)$$

сублинейно $\frac{1}{k}$!

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \sqrt{\frac{4L}{\mu K^2}}\left(f(x^0) - f(x^*)\right)$$

Рестарты:

- 1 запуск: вход $x^0 \longrightarrow$ выход $\frac{1}{K}\sum_{k=0}^{K-1} x^k$

$$K = \sqrt{\frac{16L}{\mu}}$$

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{1}{2}\left(f(x^0) - f(x^*)\right)$$

- 2 запуск: вход = выход 1-го запуска

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{1}{4}\left(f(x^0) - f(x^*)\right)$$

. . . .

$$\frac{\mu}{2}\|x^0 - x^*\| \le f(x^0) - f(x^*) \le \frac{L}{2}\|x^0 - x^*\|_2^2$$

- T запусков

$$f\left(\frac{1}{K}\sum_{k=0}^{K-1} x^k\right) - f(x^*) \le \frac{1}{2^T}\left(f(x^0) - f(x^*)\right)$$

Тогда $\varepsilon \sim (f(x) - f(x^*))$

$$T \sim \log_2 \frac{(f(x^0) - f(x^*))}{\varepsilon} \quad \text{запусков}$$

Общее число итераций (оракульная сложность)

$$K \cdot T = \boxed{4\sqrt{\frac{L}{\mu}} \log \frac{(f(x^0) - f(x^*))}{\varepsilon}}$$

а можно ли еще? $\Leftarrow$ лучше, чем у GD $\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$

Нижние оценки

- Класс алгоритмов

  1) Стартовая $x^0 = 0$, $M = \{x^0\}$

  2) изменения в M_k:

  добавлять в M_k точку вида

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

По-другому: | линейное

$$M_{k+1} = \text{span}\{x', \nabla f(x'')\} \text{ по всем } x', x'' \in M_k$$

(более общее)

- Плохая функция

$$f(x) = \frac{L-\mu}{8} x^T A x + \frac{\mu}{2}\|x\|_2^2 - \frac{L-\mu}{4} e_1^T x$$

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

f — L-гладкая, μ-сильно выпуклая

($A \succeq 0$, $\|A\|_2 \le 4$)

Док-во нижних оценок:

- найти x^* для f: $\nabla f(x^*) = 0$

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

По координатам:

1 коорд.:

$$2x_1^* - x_2^* + \frac{4\mu}{L-\mu} x_1^* - 1 = 0 \Rightarrow \frac{2(L+\mu)}{L-\mu} x_1^* - x_2^* = 1$$

Остальные координаты:

$$-x_{k-1}^* + \frac{2(L+\mu)}{L-\mu} x_k^* - x_{k+1}^* = 0$$

Линейная рекуррента

$$-1 + \frac{2(L+\mu)}{L-\mu}\lambda - \lambda^2 = 0$$

$$\boxed{x_k^* = q^k \qquad q = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}}$$

$$x_k^* = \lambda_1^k \underbrace{C_1}_{?} + \lambda_2^k \underbrace{C_2}_{?}$$

Δ в пределе (второе начальное условие)

за счет Δ подбираем $C_2 = 0$

- Идея

0) $x^0 = 0$

$$\nabla f(x^0) = \frac{L-\mu}{4} Ax + \mu x - \frac{L-\mu}{4} e_1$$

сколько ненулевых координат? 1 ненулевая

0 ненулевых коорд $\rightarrow$ 1 ненул. коорд (первая)

1) x – первая коорд. ненулевая

в $\nabla f(x)$ 2 ненул. коорд!

$1 \longrightarrow 2$ (первая и вторая)

$\boxed{K \text{ вызовов градиента} = K \text{ ненулевых коорд (от 1 до } K)}$

Размерность задачи $d = 2K$ (заведомо больше)

$\Downarrow$

K последних коорд. нулевые

выход алгоритма

$$\|x^K - x^*\|_2^2 \ge \underbrace{\sum_{i=K+1}^{2K} q^{2i}}_{\text{вторая часть}} = q^{2K}\sum_{i=1}^{K} q^{2i} = \frac{q^{2K}}{1+q^{2K}}\|x^0 - x^*\|_2^2 \, ?$$

вторая часть K последних коорд

алгоритм угадал только K коорд.

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

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

Нижняя оценка:

$$\boxed{\text{не менее } \Omega\left(\sqrt{\frac{L}{\mu}} \log \frac{\|x^0 - x^*\|_2^2}{\varepsilon}\right) \text{ оракульных вызовов}}$$

- Поляк и Nesterov оптимальные методы в введенном классе задач
- аналогично в выпуклом случае