- L-гладкой, μ-сильно выпуклой f

  $\|x^k - x^*\|_2^2 \le \varepsilon$

  $$O\left(\frac{L}{\mu}\log\frac{\|x^0 - x^*\|_2^2}{\varepsilon}\right)$$

- L-гладкая, выпуклая, $f^* > -\infty$

  $f(x^k) - f^* \le \varepsilon$

  $$O\left(\frac{L\|x^0 - x^*\|_2^2}{\varepsilon}\right)$$

- L-гладкая, невыпуклая, $f^* > -\infty$ (уникальный)

  $\|\nabla f(x^k)\|_2 \le \varepsilon$

  $$O\left(\frac{L(f(x^0) - f^*)}{\varepsilon^2}\right)$$

А можно ли лучше?

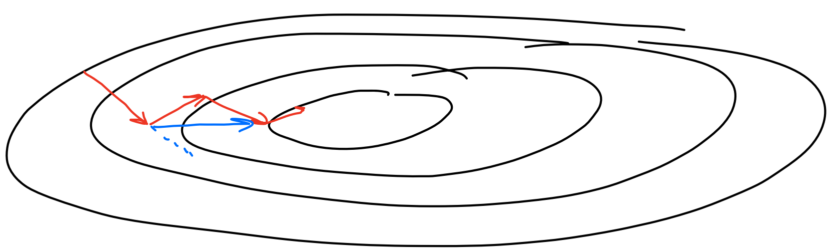
1) Поляк Б.Т. 1964

метод тяжелого шарика

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

$\tau \in (0;1)$

$\tau \in (0,9;1)$ лучше



- в теории для выпуклых задач метод тяжелого шарика не лучше GD
- на практике: + лучше у GD

  \+ немонот. сход.

  \+ уменьшение осцилляций при сход.

  \+ легкая итер., как у GD

  − нужно подбирать γ и τ

2) Нестеров Ю.Е. 1983

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

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

перепишем

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

инерцию в подсчет точки

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

GD спуск с моментумом (pytorch):

$$v^{k+1} = \beta v^k + \nabla f(x^k) \qquad \beta \in (0;1)$$

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

$$\Downarrow$$

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

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

$$\boxed{\begin{aligned} x^k &= x^{k-1} - \gamma v^k \\ -\gamma v^k &= x^k - x^{k-1} \end{aligned}}$$

инерция HB = запоминание старых град. с уменьш. весом

Линейный каплинг

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

L-гладкая, 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:

$$\underline{\|z^{k+1} - x^*\|_2^2} = \|z^k - \gamma \nabla f(x^k) - x^*\|_2^2$$
$$= \|z^k - x^*\|_2^2 - 2\gamma \underbrace{\langle \nabla f(x^k); z^k - x^*\rangle}_{\nabla f(x^k);\ x^k - x^*} + \gamma^2 \|\nabla f(x^k)\|_2^2$$
$$= \underline{\|z^k - x^*\|_2^2} - \underline{2\gamma\langle \nabla f(x^k); x^k - x^*\rangle} + \underline{\gamma^2\|\nabla f(x^k)\|_2^2}$$
$$\underbrace{-2\gamma\langle \nabla f(x^k); z^k - x^k\rangle}_{???}$$

$\|\nabla f(x^k)\|_2^2$:

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

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

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

$$\underbrace{\eta\left(1 - \frac{L\eta}{2}\right)}_{>?\quad 0<\eta<\frac{2}{L}}\|\nabla f(x^k)\|_2^2 \le f(x^k) - f(y^{k+1})$$

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

*

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

$\langle \nabla f(x^k); x^k - z^k \rangle$:

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

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

$$\underset{\text{выпуклость}}{\le} \frac{1-\tau}{\tau} \left(f(y^k) - f(x^k) \right)$$

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

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

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

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

выпуклость

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

$$\underbrace{\frac{2\gamma}{K}\sum_{k=0}^{K-1}\left(f(x^k) - f(x^*)\right)}_{\text{„}\mathbb{E} f\text{"}} \le \frac{\|z^0 - x^*\|_2^2}{K} + \frac{2\gamma^2}{K\eta(2-L\eta)}\left(f(y^0) - f(x^*)\right)$$

$f(y^K) \ge f(x^*)$

Нер-во Йенсена

$$\mathbb{E} f(\xi) \ge f(\mathbb{E}\xi) \qquad f \text{ – выпуклая}$$

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

сублинейная $\frac{1}{K}$ — хуже GD?

$y^0 = z^0 = x^0$

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

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

$$f(\bar{x}^K) - f(x^*) \le \underbrace{\left[\frac{1}{\gamma\mu K} + \frac{\gamma}{K\eta(2-L\eta)}\right]}_{\text{подобрать}} \left(f(x^0) - f(x^*)\right)$$

делаем как можно меньше

$$\eta = \frac{1}{L}$$

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

делаем min по γ

$$\gamma = \sqrt{\frac{1}{\mu L}}$$

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

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

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

За K итераций мы в 2 раза приблизились к решению

Рестарты:

$x^0 \xrightarrow{K \text{ итер}} \bar{x}^K$ 1/2 ближе к решению

$x^0 = \bar{x}^K \xrightarrow{K} \bar{x}$ 1/2 ближе (всего 1/4)

$\log\frac{1}{\varepsilon}$ рестартов K итераций

Итого:

$$T = K \cdot R = O\left(\sqrt{\frac{L}{\mu}}\,\log\frac{f(x^0)-f(x^*)}{\varepsilon}\right)$$

K ↑ кол-во рест

$\sqrt{\frac{L}{\mu}}$ ↑ лучше чем GD (для Нестерова это же самое)

ΛK:

+ оптим. оценки (быстрее GD)

− рестарты

− подбор 3х параметров вместо 2х в HB

Нестеров:

+ оптим. оценка

+ подбором только шаг $\quad \tau_k = \frac{k}{k+3}\,;\ \frac{k}{k+2}$

Вопрос оптимальности?

З || μ-сильн. L-гладк.

М || память, пополнение

GD, HB, Нестеров, ΛK

$$M_0 = \{x^0\} = \{\vec{0}\}$$

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

Нижние оценки

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

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

• решение $\min\limits_{x\in\mathbb{R}^d} f(x)$

$\nabla f(x^*) = 0 \qquad \underbrace{Ax^* + \frac{4\mu}{L-\mu}x^* - e_1 = 0}_{\text{по компонентам}}$

1) $\frac{2(L+\mu)}{L-\mu}x_1^* - x_2^* = 1$ для первой

2) $\underbrace{-x_{k-1}^* + \frac{2(L+\mu)}{L-\mu}x_k^* - x_{k+1}^* = 0}_{\text{линейная рек: } x_{k+1} = \mathrm{Lin}(x_k;\, x_{k-1})}$ для остальных

Решение рекур.

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

• как работает любой метод:

$\nabla f(x) \sim Ax + x + e_1$

стартуем из 0 → в градиенте только 1 коорд не нулевая

имеем 1 коорд не 0 → 2 ненулевые коорд.

k коорд → k+1 ненул. коорд.

• размерность задачи $d = 2K$ K – кол. вызовов ∇f

K – ненул. коорд. (первых), но последние K коорд только нулевые

$$\underset{\text{выход ал.}}{\|x^K - x^*\|_2^2} \geq \sum_{k=K+1}^{2K} q^k \geq \left(1-\frac{2\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^{2K}\frac{\|\overset{0}{\overset{\|}{x^0}}-x^*\|_2^2}{2}$$

$$\Omega\left(\frac{\sqrt{L}+\sqrt{\mu}}{2\sqrt{\mu}}\log\frac{1}{\varepsilon}\right) \quad \text{вычислений град.}$$

$$\gg$$

$$\Omega\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$$

- для выпуклых аналогично: ЛК и Нестеров оптимальны
- для невыпуклых и гладких: GD оптимален