На прошлой лекции:

f L-гладкая и μ-сильно выпуклая

для град. спуска

$O\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|^2}{\varepsilon}\right)$ итераций/оракул. вызовов

для получения решения с точн. ε: $\|x^k-x^*\|^2\le\varepsilon$

А можно ли лучше? Другой метод?

Метод тяжелого шарика (Heavy Ball/HB)

Б.Т. Поляк 1964

Алгоритм 1 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 $\nabla f(x^k)$

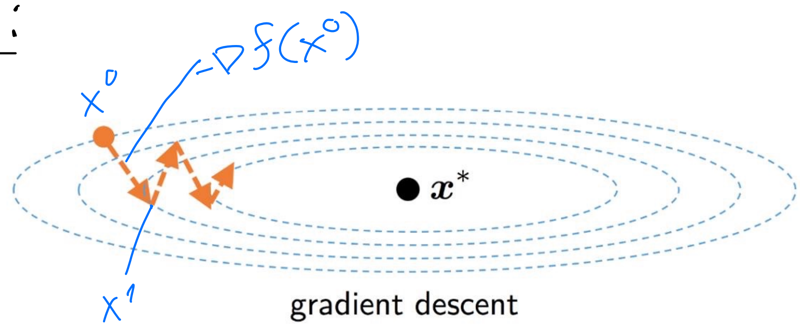
3: $x^{k+1} = x^k - \gamma_k \nabla f(x^k) \underbrace{+ \tau_k(x^k - x^{k-1})}_{\text{инерция}}$

4: **end for**

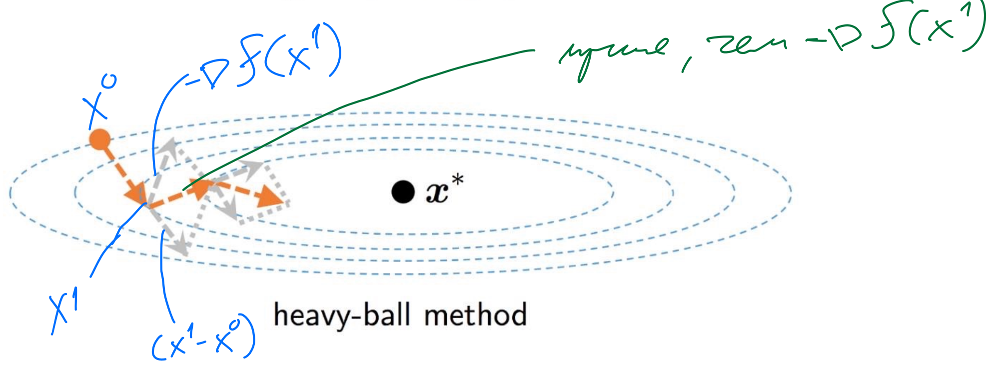
Выход: x^K

полож. моментум $\tau_k \ge 0$ $\tau_k \in [0;1]$

Физика:



gradient descent



heavy-ball method

Плюсы:

⊕ на практике лучше, чем град. спуск

⊕ уменьш. осцилляций траектории

⊕ легкость интер.

Минусы:

⊖ нужно подбирать параметры: γ и τ $\left(\tau \in (0{,}9;1)\right)$

⊖ для больших задач хранить x^{k-1} дорого

⊖ в теории метод HB не лучше, чем GD

Особенность:

⊙ немон. сходимость

Градиентный спуск с моментумом (pytorch):

$$
\begin{aligned}
v^{k+1} &= \beta v^k + \nabla f(x^k) \\
x^{k+1} &= x^k - \gamma v^{k+1}
\end{aligned}
\qquad \beta \in (0;1)
$$

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

пред. шт.: $x^k = x^{k-1} - \gamma v^k \Rightarrow -\gamma v^k = x^k - x^{k-1}$

$$x^{k+1} = x^k - \gamma \nabla f(x^k) + \beta\left(x^k - x^{k-1}\right) \qquad \text{HB}$$

Физика с точки зрения ML:

HB моментум = сохранение старых градиентов с уменьш. весом

Ускоренный градиентный спуск (Nesterov)

Ю. Е. Нестеров 1983

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

for $k = 0, 1, \ldots, K-1$ **do**
  Вычислить $\nabla f(y^k)$
  $x^{k+1} = y^k - \gamma_k \nabla f(y^k)$
  $y^{k+1} = x^{k+1} + \tau_k(x^{k+1} - x^k)$
end for

Выход: x^K

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

$$\text{Nesterov: } x^{k+1} = x^k + \tau_{k-1}(x^k - x^{k-1}) - \gamma_k \nabla f(x^k \underbrace{+ \tau_{k-1}(x^k - x^{k-1})}_{})$$
$$\text{HB: } x^{k+1} = x^k + \tau_{k-1}(x^k - x^{k-1}) - \gamma_k \nabla f(x^k)$$

итерация в точке подсчета град-

Плюсы:

⊕ плюсы HB

⊕ есть теория по подбору $\tau_k = \frac{k}{k+3}; \frac{k}{k+2}$

⊕ теор. лучше, чем град. спуск

Nesterov:
$$O\left(\sqrt{\frac{L}{\mu}} \log \frac{f(x^0) - f(x^*) + \|x^0 - x^*\|^2}{\varepsilon}\right)$$ итерац./орак. вызовов

лучше, чем у GD $O\left(\frac{L}{\mu}\right)$

⊕ оптимален

Минусы:

⊖ минусы HB

Особенности:

⊙ как у HB

А можно ли лучше?

Нижние оценки:

З М

μ-сильно вып.
L-гладкая

«память» M_k

- иниц: $M_0 = \{x^0\} = \{\vec{0}\}$
- обновл.: $M_k = \mathrm{span}\{x', \nabla f(x'')\}$
  $x', x'' \in M_{k-1}$
- выход: $x \in M_k$

GD, HB, Nesterov удовл.

Плохая задача (функция):

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

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

δ – подберем далее

- $A \succeq 0$, тогда f – μ-сильно вып.
- $\|A\|_2 \le 4$, тогда f L-гладкая

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

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

по компонентам

- первая компон.:

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

нач. услов.

- остальные (кроме первой и послед.):

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

линейная рекурр.

- последняя комп.:

.... (завис от ξ)

нач. условия

$x_{k+1} = A x_k + B x_{k-1}$

$\lambda^2 = A\lambda + B$

λ_1, λ_2 – корни

$\lambda_1 \neq \lambda_2$

$x_k = C_1 \lambda_1^k + C_2 \lambda_2^k$

C_1, C_2 из нач. условий

Решение рекурренты

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

$\cancel{+\tilde{q}^k}$ за счет ξ

2) Как работает метод из класса:

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

$$\sim Ax + x + e_1$$

- стартуем из 0

вызываем оракул ∇f

$\mathrm{span}\{0, \nabla f(0)\} = \mathrm{span}\{e_1\}$

разблокировали 1ую коорд.

- 1ая коорд. ненулевая

вызываем оракул ∇f

$\mathrm{span}\{e_1, \nabla f(x_1)\} = \mathrm{span}\{e_1, e_2\}$

$x \in \mathrm{span}\{e_1\}$

разбрасыв. 2ую коорд.

- k-ый вызов оракула $\nabla f \Rightarrow k$ ненулевых коорд.

  ост. все нулевые

3) Размерность задачи $d = 2\tilde{K}$ ($\tilde{K}$ кол-во вызовов ∇f)

В идеальном случае мы угадаем $\tilde{K}$ первых коорд

На все ост. может 0 в выходе

$$\underset{\text{выход}}{\|x^{\tilde{K}}} - x^*\|^2 \underset{\text{оценка снизу}}{\geq} \underbrace{\sum_{k=1}^{\tilde{K}} 0}_{\text{все угадали}} + \sum_{k=\tilde{K}+1}^{2\tilde{K}} (\underset{\text{выход}}{0} - \underset{x^*_k}{q^k})^2$$

$$= \sum_{k=\tilde{K}+1}^{2\tilde{K}} (q^k)^2 = q^{2\tilde{K}} \sum_{k=1}^{\tilde{K}} (q^k)^2$$

$$\|\underset{0}{x^0} - x^*\|^2 = \|x^*\|^2 = \sum_{k=1}^{2\tilde{K}} (q^k)^2 =$$

$$= \sum_{k=1}^{\tilde{K}} (q^k)^2 + \underbrace{\sum_{k=\tilde{K}+1}^{2\tilde{K}} (q^k)^2}_{q^{2\tilde{K}} \sum_{k=1}^{\tilde{K}} (q^k)^2}$$

$$= \left(1 + q^{2\tilde{K}}\right) \underbrace{\sum_{k=1}^{\tilde{K}} (q^k)^2}$$

$$\|x^{\bar{k}} - x^*\|^2 \geq \frac{q^{2\bar{k}}}{1+q^{2\bar{k}}}\|x^0-x^*\|^2$$

$$\geq \frac{q^{2\bar{k}}}{2}\|x^0-x^*\|^2$$

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

$\varepsilon \sim \|x^{\bar{k}}-x^*\|^2$

<u>Ф-ка нижней оценки</u>

f – L-гладкая, μ-сильно выпуклая, любой метод из класса найдет решение ($\|x^k - x^*\|^2 \le \varepsilon$) не быстрее, чем

$$\Omega\left(\sqrt{\frac{L}{\mu}}\log\frac{\|x^0-x^*\|^2}{\varepsilon}\right)$$ оракул вызовов

Метод Нестерова оптимален!