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

L – гладкая, μ – сильно выпуклая

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

нужно градиентному спуску

а можно лучше?

1964. Б. Т. Поляк

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**

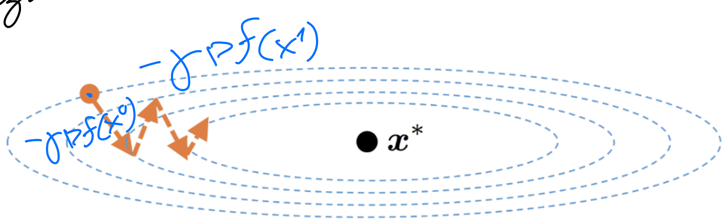
2: Вычислить $\nabla f(x^k)$

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

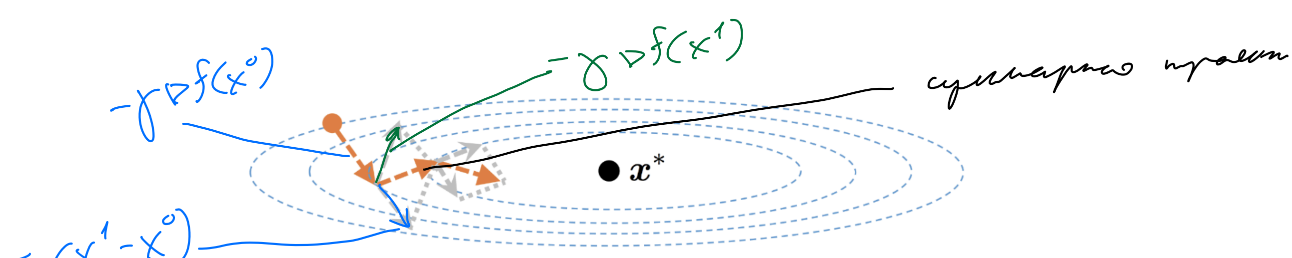
4: **end for**

Выход: x^K

Физический смысл:



gradient descent



heavy-ball method

Pytorch (основная библиотека DL)

GD с моментум
$$\begin{cases} v^{k+1} = \beta v^k + \nabla f(x^k) \\ x^{k+1} = x^k - \gamma v^{k+1} \end{cases} \quad \beta \in [0;1]$$

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

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

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

(+):

\+ интуиция и физика

\+ легкость имплементации

\+ дешевизна вычислений

(−):

− два параметра для подбора ($\tau \approx [0{,}85; 0{,}95]$)

− по памяти +1 вектор по сравнению с град. спуском

− в теории не лучше град. спуска

1983 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

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

Тяжелый шарик:

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

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

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

$$x^{k+1} = x^k + \underbrace{\tau (x^k - x^{k-1})}_{\text{т.ш.}} - \gamma_k \nabla f(x^k + \tau(x^k - x^{k-1}))$$

Другой метод, ускоряющий градиентный спуск

Алгоритм 4 Линейный каплинг: внутренний цикл

Вход: размер шагов $\{\gamma_k\}_{k=0} > 0$ и $\{\eta_k\}_{k=0} > 0$, моментумы $\{\tau_k\}_{k=0} \in [0;1]$, 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: $y^{k+1} = x^k - \eta_k \nabla f(x^k)$ ← стандартный шаг
4: $z^{k+1} = z^k - \gamma_k \nabla f(x^k)$ ← быстрый шаг
5: $x^{k+1} = \tau_k z^{k+1} + (1-\tau_k) y^{k+1}$ ← вып. комб.
6: **end for**

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

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

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

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

$$K = 4\sqrt{\frac{L}{\mu}}$$

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

Рестарт из новой стартовой точки

$$f(x_{final,new}) - f^* \leq \frac{1}{2}\left(f(\underbrace{x_{start,new}}_{= x_{final}}) - f^*\right)$$

Тогда нужно $\log_2 \frac{f(x^0) - f^*}{\varepsilon}$ рестартов, чтобы гарантировать точность ε по функции $f(x) - f^* \leq \varepsilon$

Итоговая сложность: $\underbrace{4\sqrt{\frac{L}{\mu}}}_{K} \log_2 \frac{f(x^0) - f^*}{\varepsilon}$

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

Ускор. град. метод такие же оценки → $$O\left(\sqrt{\frac{L}{\mu}}\log\frac{f(x^0) - f(x^*)}{\varepsilon}\right)$$ вызовов оракула.

лучше чем град спуск $\frac{L}{\mu}\log\ldots$

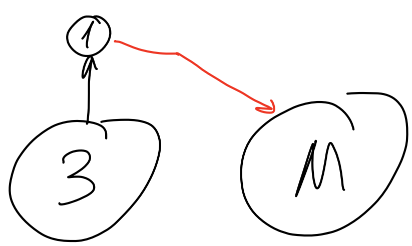
сходимость по функции:

- для сильно выпуклых задач ≈ эквивалентно
- для ускор. метода Нестерова

сходимость:

$$\|x^k - x^*\|_2^2 + f(x^k) - f^*$$

Можно ли еще лучше? Нижние оценки



Класс методов:

- $x^0 = 0$ стартовая точка

  $\mathcal{M}_0 = \{x^0\}$ ← послед. памяти
- $\nabla f(x^k)$ $x^k \in \mathcal{M}_k$
- $\mathcal{M}_{k+1} = \text{span}\{\mathcal{M}_k, \nabla f(x^k)\}$ ← лин. оболочка
- $\bar{K}$ вызовов оракула $x_{final} \in \mathcal{M}_{\bar{K}}$

Плохая функция

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

$$A = \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ 0 & & & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{pmatrix} \quad e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

1) f – L-гладкая, μ-сильно выпуклая

$\|A\|_2 \le 4$ $\quad A \succeq 0$ $\quad$ упражнение

2) x^* – решение найдем

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

$$\frac{L-\mu}{8} \cdot 2Ax + \frac{\mu}{2} \cdot 2x - \frac{L-\mu}{4} e_1 = 0$$

по-координатно:

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

другие $\frac{L-\mu}{4}\left(-x_{k-1} + 2x_k - x_{k+1}\right) + \mu x_k = 0$

последняя: ... ..

$$x_{k+1} = \left(2 + \frac{4\mu}{L-\mu}\right) x_k - x_{k-1}$$

линейная рекуррента

$$\lambda^2 = \left(2 + \frac{4\mu}{L-\mu}\right)\lambda - 1$$

$\lambda_{\text{решения}}$: $\lambda_1 = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$ $\quad \lambda_2 = \ldots$

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

зададим нач. условия так $C_2 = 0$ $C_1 = 1$

$$x_k = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^k$$

3)

$$\nabla f(x) = \frac{L-\mu}{8} \cdot 2Ax + \frac{\mu}{2} \cdot 2x - \frac{L-\mu}{4} e_1$$

1. $x^0 = 0$ $\quad \nabla f(x^0) \in \mathrm{span}\{e_1\}$

только 1ая коорд не нулевая

$$\mathcal{M}_1 \leq \mathrm{span}\{e_1\}$$

2. $x^1 \in \mathcal{M}_1$ $\quad \nabla f(x^1) \in \mathrm{span}\{e_1, e_2\}$

только 1ая и 2ая коорд ненулевые

$$\mathcal{M}_2 \in \mathrm{span}\{e_1, e_2\}$$

k. $x^{k-1} \in \mathcal{M}_{k-1}$ $\quad \nabla f(x^{k-1}) \in \mathrm{span}\{e_1 \ldots e_k\}$

Summary:

$$x_k^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^k = q^k$$

после $\bar{K}$ вызов оракула только $\bar{K}$ первых коорд ненулевые

$$\|x_{final}^{\bar{K}} - x^*\|^2 = \sum_{i=1}^{d} \left(x_{final,i}^{\bar{K}} - x_i^*\right)^2$$

$d = 2\bar{K}$ разбили не все коорд.

в лучшем случае метод угадает $\bar{K}$ коорд, дальше 0

$$\geq \sum_{i=\bar{k}+1}^{2\bar{k}=d} \left(x_{final,i}^{\bar{k}} - x_i^*\right)^2$$

$$\|x_{final}^{\bar{k}} - x^*\|^2 = \sum_{i=\bar{k}+1}^{2\bar{k}} (x_i^*)^2 = \sum_{i=\bar{k}+1}^{2\bar{k}} q^{2i}$$

$$\|x^0 - x^*\|_2^2 = \sum_{i=1}^{2\bar{k}} (x_i^*)^2 = \sum_{i=1}^{2\bar{k}} q^{2i}$$

$x^0 = 0$

$$\sum_{i=1}^{2\bar{k}} q^{2i} = \sum_{i=1}^{\bar{k}} q^{2i} + \underbrace{\sum_{i=\bar{k}+1}^{2\bar{k}} q^{2i}}_{q^{2\bar{k}} \sum_{i=1}^{\bar{k}} q^{2i}} = \left(1+q^{2\bar{k}}\right)\sum_{i=1}^{\bar{k}} q^{2i}$$

$$\sum_{i=\bar{k}+1}^{2\bar{k}} q^{2i} = q^{2\bar{k}} \cdot \sum_{i=1}^{\bar{k}} q^{2i}$$

$$\|x_{final}^{\bar{k}} - x^*\|_2^2 \geq \frac{q^{2\bar{k}}}{1+q^{2\bar{k}}} \|x^0 - x^*\|_2^2$$

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

K через ε $\qquad K \sim \sqrt{\frac{L}{\mu}} \log \ldots$

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

показывает оптимальность ускоренных методов

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