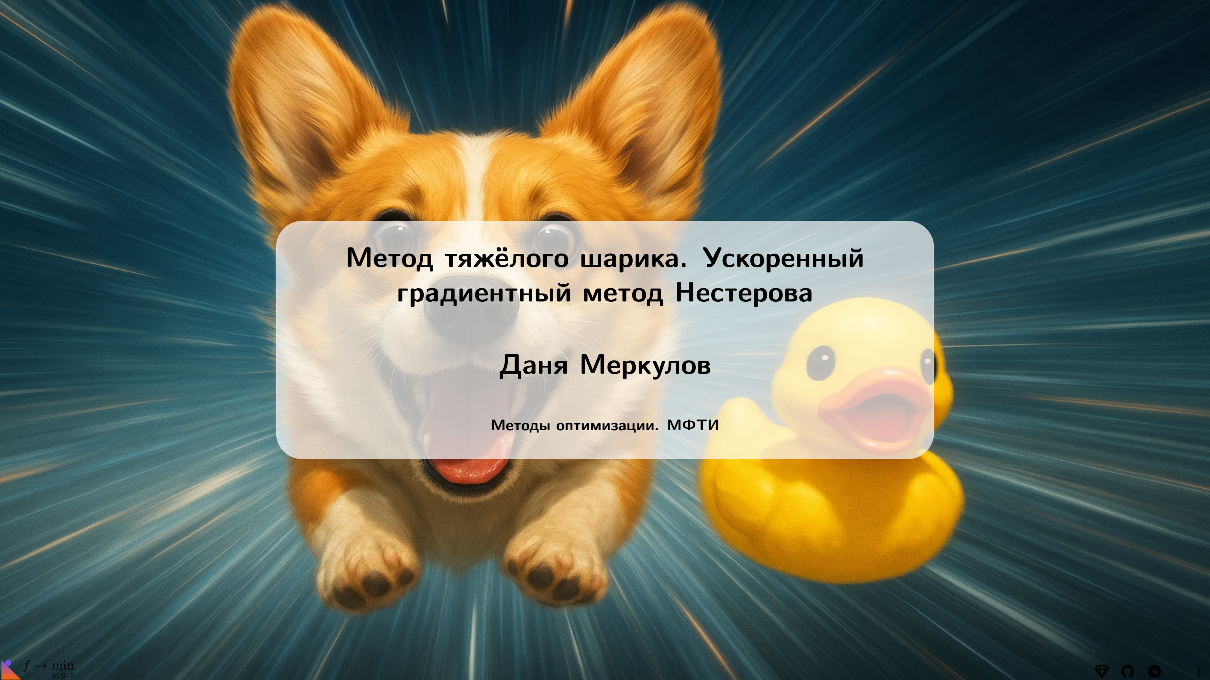


Метод тяжёлого шарика. Ускоренный градиентный метод Нестерова

Даня Меркулов

Методы оптимизации. МФТИ

Повторение

Результаты сходимости градиентного спуска для гладких функций

Градиентный спуск: $\min\limits_{x \in \mathbb{R}^n} f(x)$ $\quad x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ $\quad \lambda\left(\nabla^2 f(x)\right) \in [\mu, L], \varkappa = \frac{L}{\mu}$

выпуклая (негладкая)	гладкая (невыпуклая)	гладкая & выпуклая	гладкая & сильно выпуклая
$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$\min\limits_{0 \leq i \leq k} \|\nabla f(x_i)\| = \mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \mathcal{O}\left(\frac{1}{k}\right)$ $k_\varepsilon = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$	$\|x_k - x^*\|^2 = \mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^k\right)$ $k_\varepsilon = \mathcal{O}\left(\varkappa \log \frac{1}{\varepsilon}\right)$

Нижние оценки для методов I порядка на классе гладких функций

Произвольный метод I порядка: $\min_{x \in \mathbb{R}^n} f(x)$ $\quad x_{k+1} = x_k - \sum_{i=0}^{k} \alpha_i \nabla f(x_i)$ $\quad \lambda\left(\nabla^2 f(x)\right) \in [\mu, L], \varkappa = \frac{L}{\mu}$

выпуклая (негладкая)	гладкая (невыпуклая)[1]	гладкая & выпуклая[2]	гладкая & сильно выпуклая
$f(x_k) - f^* = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$\min_{0 \le i \le k} \|\nabla f(x_i)\| = \Omega\left(\frac{1}{\sqrt{k}}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\varepsilon^2}\right)$	$f(x_k) - f^* = \Omega\left(\frac{1}{k^2}\right)$ $k_\varepsilon = \Omega\left(\frac{1}{\sqrt{\varepsilon}}\right)$	$f(x_k) - f^* = \Omega\left(\left(\frac{\sqrt{\varkappa}-1}{\sqrt{\varkappa}+1}\right)^{2k}\right)$ $k_\varepsilon = \Omega\left(\sqrt{\varkappa} \log \frac{1}{\varepsilon}\right)$

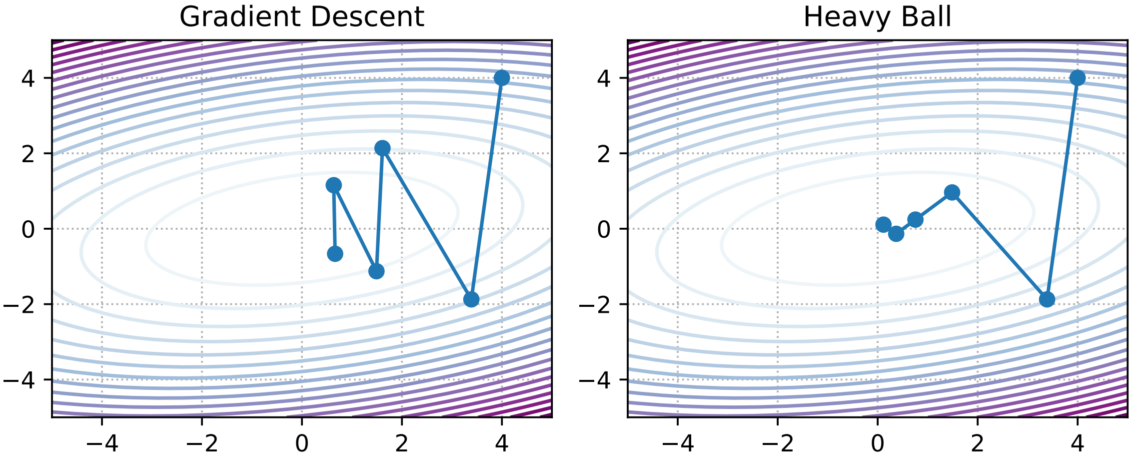
[1] Carmon, Duchi, Hinder, Sidford, 2017

[2] Nemirovski, Yudin, 1979

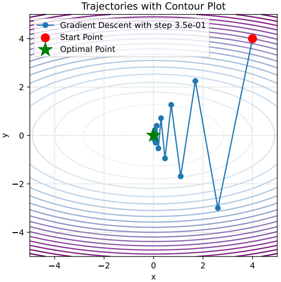
Метод тяжёлого шарика

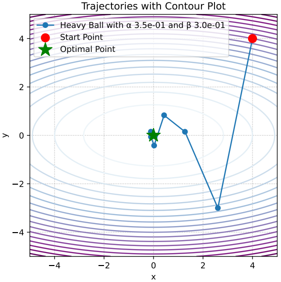
Колебания и ускорение

$$
x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).
$$



Метод тяжёлого шарика Поляка

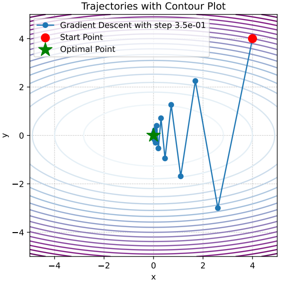


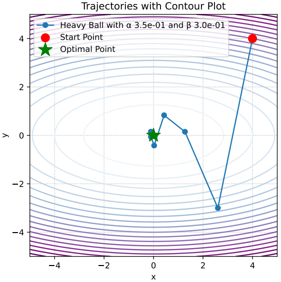


Рассмотрим идею моментума (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}).$$

Метод тяжёлого шарика Поляка



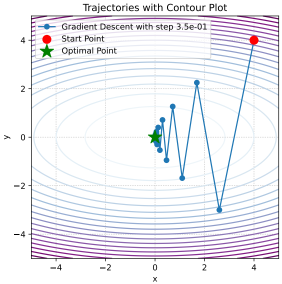


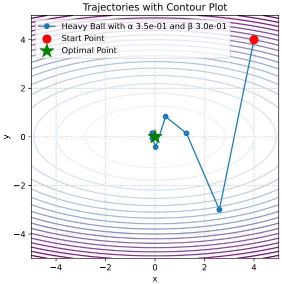
Рассмотрим идею моментума (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).$$

Давайте теперь подставим в итерацию предыдущую итерацию, т.е. $x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$, а так же отметим, что $x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$:

Метод тяжёлого шарика Поляка





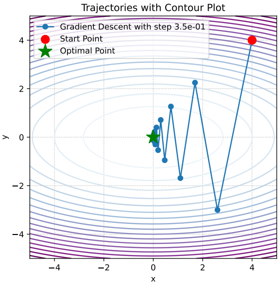
Рассмотрим идею моментума (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

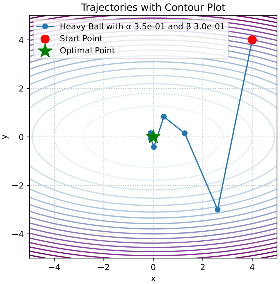
$$
x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}).
$$

Давайте теперь подставим в итерацию предыдущую итерацию, т.е. $x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$, а так же отметим, что $x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$:

$$
x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1})
$$

Метод тяжёлого шарика Поляка





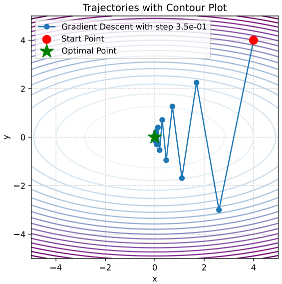
Рассмотрим идею моментума (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

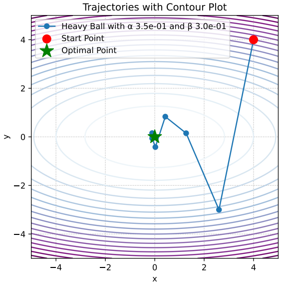
$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}).$$

Давайте теперь подставим в итерацию предыдущую итерацию, т.е. $x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})$, а так же отметим, что $x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})$:

$$\begin{aligned} x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}) \\ &= x_k - \alpha \nabla f(x_k) + \beta \left(-\alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2}) \right) \end{aligned}$$

Метод тяжёлого шарика Поляка





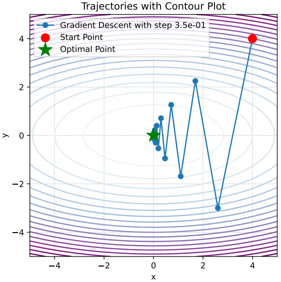
Рассмотрим идею моментума (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

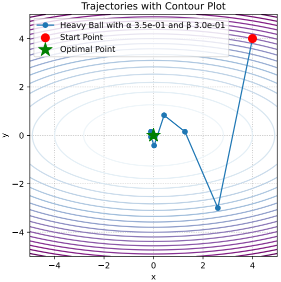
$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).$$

Давайте теперь подставим в итерацию предыдущую итерацию, т.е. $x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$, а так же отметим, что $x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$:

$$\begin{aligned} x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \\ &= x_k - \alpha \nabla f(x_k) + \beta\left(-\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})\right) \\ &= x_k - \alpha\left[\nabla f(x_k) + \beta \nabla f(x_{k-1})\right] + \beta^2(x_{k-1} - x_{k-2}) \end{aligned}$$

Метод тяжёлого шарика Поляка





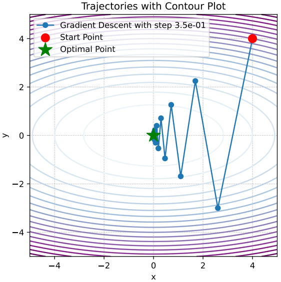
Рассмотрим идею моментума (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

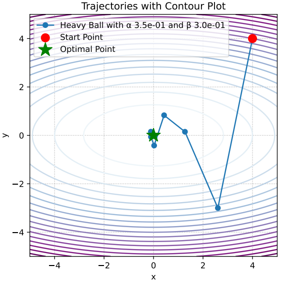
$$
x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}).
$$

Давайте теперь подставим в итерацию предыдущую итерацию, т.е. $x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})$, а так же отметим, что $x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})$:

$$
\begin{aligned}
x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}) \\
&= x_k - \alpha \nabla f(x_k) + \beta \left(-\alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})\right) \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1})\right] + \beta^2 (x_{k-1} - x_{k-2}) \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2})\right] + \beta^3 (x_{k-2} - x_{k-3})
\end{aligned}
$$

Метод тяжёлого шарика Поляка





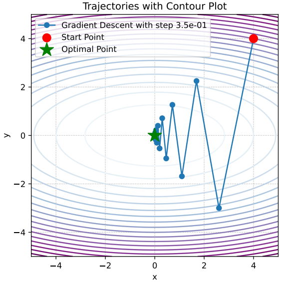
Рассмотрим идею моментума (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

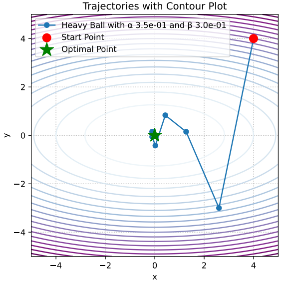
$$
x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).
$$

Давайте теперь подставим в итерацию предыдущую итерацию, т.е. $x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$, а так же отметим, что $x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})$:

$$
\begin{aligned}
x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) \\
&= x_k - \alpha \nabla f(x_k) + \beta\left(-\alpha \nabla f(x_{k-1}) + \beta(x_{k-1} - x_{k-2})\right) \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1})\right] + \beta^2 (x_{k-1} - x_{k-2}) \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2})\right] + \beta^3 (x_{k-2} - x_{k-3}) \\
&\vdots
\end{aligned}
$$

Метод тяжёлого шарика Поляка





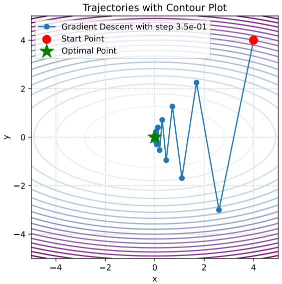
Рассмотрим идею моментума (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

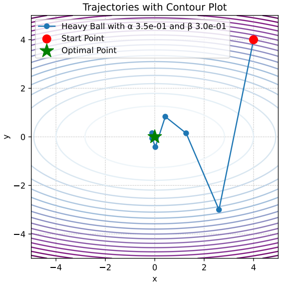
$$
x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}).
$$

Давайте теперь подставим в итерацию предыдущую итерацию, т.е. $x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})$, а так же отметим, что $x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})$:

$$
\begin{aligned}
x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}) \\
&= x_k - \alpha \nabla f(x_k) + \beta \left(-\alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})\right) \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1})\right] + \beta^2 (x_{k-1} - x_{k-2}) \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2})\right] + \beta^3 (x_{k-2} - x_{k-3}) \\
&\vdots \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2}) + \dots + \beta^k \nabla f(x_0)\right]
\end{aligned}
$$

Метод тяжёлого шарика Поляка





Рассмотрим идею моментума (импульса, тяжёлого шарика), предложенную Б.Т. Поляком в 1964 году. Обновление метода тяжёлого шарика имеет вид

$$
x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}).
$$

Давайте теперь подставим в итерацию предыдущую итерацию, т.е. $x_k = x_{k-1} - \alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})$, а так же отметим, что $x_k - x_{k-1} = -\alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2})$:

$$
\begin{aligned}
x_{k+1} &= x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}) \\
&= x_k - \alpha \nabla f(x_k) + \beta \left(-\alpha \nabla f(x_{k-1}) + \beta (x_{k-1} - x_{k-2}) \right) \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1}) \right] + \beta^2 (x_{k-1} - x_{k-2}) \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2}) \right] + \beta^3 (x_{k-2} - x_{k-3}) \\
&\vdots \\
&= x_k - \alpha \left[\nabla f(x_k) + \beta \nabla f(x_{k-1}) + \beta^2 \nabla f(x_{k-2}) + \dots + \beta^k \nabla f(x_0) \right]
\end{aligned}
$$

Таким образом, метод тяжёлого шарика учитывает все предудущие градиенты с тем меньшим весом, чем старше итерация ($0 \leq \beta < 1$).

Метод тяжёлого шарика Поляка для сильно выпуклой квадратичной функции

Мы уже до этого показывали, как для произвольной сильно выпуклой квадратичной функции можно сделать замену координат так, чтобы в новых координатах матрица квадратичной формы имела диагональный вид. Поэтому, можно рассмотреть функцию $f(x) = \frac{1}{2}x^T\Lambda x$ с диагональной матрицей Λ и собственными значениями $\lambda(\Lambda) \in [\mu, L]$. Тогда

$$x_{k+1} = x_k - \alpha\Lambda x_k + \beta(x_k - x_{k-1}) = (I - \alpha\Lambda + \beta I)x_k - \beta x_{k-1}$$

Метод тяжёлого шарика Поляка для сильно выпуклой квадратичной функции

Мы уже до этого показывали, как для произвольной сильно выпуклой квадратичной функции можно сделать замену координат так, чтобы в новых координатах матрица квадратичной формы имела диагональный вид. Поэтому, можно рассмотреть функцию $f(x) = \frac{1}{2}x^T \Lambda x$ с диагональной матрицей Λ и собственными значениями $\lambda(\Lambda) \in [\mu, L]$. Тогда

$$x_{k+1} = x_k - \alpha \Lambda x_k + \beta(x_k - x_{k-1}) = (I - \alpha\Lambda + \beta I)x_k - \beta x_{k-1}$$

Это можно переписать как

$$\begin{aligned}
\hat{x}_{k+1} &= (I - \alpha\Lambda + \beta I)\hat{x}_k - \beta\hat{x}_{k-1}, \\
\hat{x}_k &= \hat{x}_k.
\end{aligned}$$

Метод тяжёлого шарика Поляка для сильно выпуклой квадратичной функции

Мы уже до этого показывали, как для произвольной сильно выпуклой квадратичной функции можно сделать замену координат так, чтобы в новых координатах матрица квадратичной формы имела диагональный вид. Поэтому, можно рассмотреть функцию $f(x) = \frac{1}{2}x^T\Lambda x$ с диагональной матрицей Λ и собственными значениями $\lambda(\Lambda) \in [\mu, L]$. Тогда

$$
x_{k+1} = x_k - \alpha\Lambda x_k + \beta(x_k - x_{k-1}) = (I - \alpha\Lambda + \beta I)x_k - \beta x_{k-1}
$$

Это можно переписать как

$$
\begin{aligned}
\hat{x}_{k+1} &= (I - \alpha\Lambda + \beta I)\hat{x}_k - \beta\hat{x}_{k-1}, \\
\hat{x}_k &= \hat{x}_k.
\end{aligned}
$$

Давайте введем следующее обозначение: $\hat{z}_k = \begin{bmatrix} \hat{x}_{k+1} \\ \hat{x}_k \end{bmatrix}$. Следовательно, $\hat{z}_{k+1} = M\hat{z}_k$, где матрица итерации M имеет вид:

Метод тяжёлого шарика Поляка для сильно выпуклой квадратичной функции

Мы уже до этого показывали, как для произвольной сильно выпуклой квадратичной функции можно сделать замену координат так, чтобы в новых координатах матрица квадратичной формы имела диагональный вид. Поэтому, можно рассмотреть функцию $f(x) = \frac{1}{2}x^T\Lambda x$ с диагональной матрицей Λ и собственными значениями $\lambda(\Lambda) \in [\mu, L]$. Тогда

$$
x_{k+1} = x_k - \alpha\Lambda x_k + \beta(x_k - x_{k-1}) = (I - \alpha\Lambda + \beta I)x_k - \beta x_{k-1}
$$

Это можно переписать как

$$
\begin{aligned}
\hat{x}_{k+1} &= (I - \alpha\Lambda + \beta I)\hat{x}_k - \beta\hat{x}_{k-1}, \\
\hat{x}_k &= \hat{x}_k.
\end{aligned}
$$

Давайте введем следующее обозначение: $\hat{z}_k = \begin{bmatrix} \hat{x}_{k+1} \\ \hat{x}_k \end{bmatrix}$. Следовательно, $\hat{z}_{k+1} = M\hat{z}_k$, где матрица итерации M имеет вид:

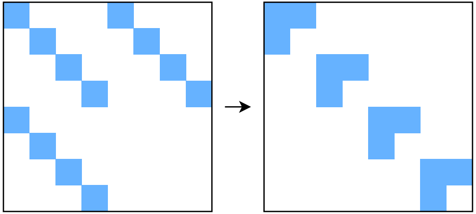
$$
M = \begin{bmatrix} I - \alpha\Lambda + \beta I & -\beta I \\ I & 0_d \end{bmatrix}.
$$

Сведение к скалярному случаю

Обратим внимание, что M является матрицей $2d \times 2d$ с четырьмя блочно-диагональными матрицами размера $d \times d$ внутри. Это означает, что мы можем изменить порядок координат, чтобы сделать M блочно-диагональной. Обратите внимание, что в уравнении ниже матрица M обозначает то же самое, что и в обозначении выше, за исключением описанной перестановки строк и столбцов. Мы используем эту небольшую перегрузку обозначений для простоты.

Сведение к скалярному случаю

Обратим внимание, что M является матрицей $2d \times 2d$ с четырьмя блочно-диагональными матрицами размера $d \times d$ внутри. Это означает, что мы можем изменить порядок координат, чтобы сделать M блочно-диагональной. Обратите внимание, что в уравнении ниже матрица M обозначает то же самое, что и в обозначении выше, за исключением описанной перестановки строк и столбцов. Мы используем эту небольшую перегрузку обозначений для простоты.



$$
\begin{bmatrix} \hat{x}_k^{(1)} \\ \vdots \\ \hat{x}_k^{(d)} \\ \hat{x}_{k-1}^{(1)} \\ \vdots \\ \hat{x}_{k-1}^{(d)} \end{bmatrix} \to \begin{bmatrix} \hat{x}_k^{(1)} \\ \hat{x}_{k-1}^{(1)} \\ \vdots \\ \hat{x}_k^{(d)} \\ \hat{x}_{k-1}^{(d)} \end{bmatrix} \qquad M = \begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \ldots & \\ & & & M_d \end{bmatrix}
$$

Рис. 1: Иллюстрация перестановки матрицы M

где $\hat{x}_k^{(i)}$ является i-й координатой вектора $\hat{x}_k \in \mathbb{R}^d$ и M_i обозначает матрицу размера 2×2. Переупорядочение позволяет нам исследовать динамику метода независимо от размерности. Асимптотическая скорость сходимости последовательности векторов $\hat{z}_k$ размерности $2d$ определяется наихудшей скоростью сходимости среди его блока координат. Следовательно, достаточно исследовать оптимизацию в одномерном случае.

Сведение к скалярному случаю

Для i-й координаты, где λ_i — i-е собственное значение матрицы A, имеем:

$$M_i = \begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix}.$$

Сведение к скалярному случаю

Для i-й координаты, где λ_i — i-е собственное значение матрицы A, имеем:

$$M_i = \begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix}.$$

Метод будет сходиться, если $\rho(M) < 1$, и оптимальные параметры могут быть вычислены путем оптимизации спектрального радиуса

$$\alpha^*, \beta^* = \arg\min_{\alpha,\beta} \max_i \rho(M_i), \quad \alpha^* = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}, \quad \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2.$$

Сведение к скалярному случаю

Для i-й координаты, где λ_i — i-е собственное значение матрицы A, имеем:

$$M_i = \begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix}.$$

Метод будет сходиться, если $\rho(M) < 1$, и оптимальные параметры могут быть вычислены путем оптимизации спектрального радиуса

$$\alpha^*, \beta^* = \arg\min_{\alpha,\beta} \max_i \rho(M_i), \quad \alpha^* = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}, \quad \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2.$$

Можно показать, что для таких параметров матрица M имеет комплексные собственные значения, которые образуют комплексно-сопряжённую пару, поэтому расстояние до оптимума (в этом случае $\|z_k\|$) обычно не убывает монотонно.

Характеристическое уравнение

Собственные значения матрицы M_i определяются из характеристического уравнения $\det(M_i - zI) = 0$:

Характеристическое уравнение

Собственные значения матрицы M_i определяются из характеристического уравнения $\det(M_i - zI) = 0$:

$$\det \begin{pmatrix} 1 - \alpha\lambda_i + \beta - z & -\beta \\ 1 & -z \end{pmatrix} = -z(1 - \alpha\lambda_i + \beta - z) + \beta = z^2 - (1 - \alpha\lambda_i + \beta)z + \beta = 0$$

Характеристическое уравнение

Собственные значения матрицы M_i определяются из характеристического уравнения $\det(M_i - zI) = 0$:

$$\det \begin{pmatrix} 1 - \alpha\lambda_i + \beta - z & -\beta \\ 1 & -z \end{pmatrix} = -z(1 - \alpha\lambda_i + \beta - z) + \beta = z^2 - (1 - \alpha\lambda_i + \beta)z + \beta = 0$$

Пусть z_1, z_2 — корни этого уравнения. По теореме Виета:

$$z_1 z_2 = \beta, \qquad z_1 + z_2 = 1 - \alpha\lambda_i + \beta$$

Характеристическое уравнение

Собственные значения матрицы M_i определяются из характеристического уравнения $\det(M_i - zI) = 0$:

$$\det \begin{pmatrix} 1 - \alpha\lambda_i + \beta - z & -\beta \\ 1 & -z \end{pmatrix} = -z(1 - \alpha\lambda_i + \beta - z) + \beta = z^2 - (1 - \alpha\lambda_i + \beta)z + \beta = 0$$

Пусть z_1, z_2 — корни этого уравнения. По теореме Виета:

$$z_1 z_2 = \beta, \qquad z_1 + z_2 = 1 - \alpha\lambda_i + \beta$$

Спектральный радиус $\rho(M_i) = \max(|z_1|, |z_2|)$. Для сходимости необходимо $\rho(M_i) < 1$, что подразумевает $\beta < 1$ (так как $z_1 z_2 = \beta$).

Анализ дискриминанта: Вещественные корни

Дискриминант квадратного уравнения $z^2 - (1 - \alpha\lambda_i + \beta)z + \beta = 0$:

$$D = (1 - \alpha\lambda_i + \beta)^2 - 4\beta$$

Рассмотрим случай **вещественных корней** $(D \geq 0)$. Корни вещественны и $z_1, z_2 = \frac{1-\alpha\lambda_i+\beta\pm\sqrt{D}}{2}$. Так как $z_1 z_2 = \beta$, то если корни различны, один из них по модулю должен быть больше $\sqrt{\beta}$ (если только они не равны $\pm\sqrt{\beta}$). Более того, если $D > 0$, то $\max(|z_1|, |z_2|) > \sqrt{\beta}$. Это означает, что скорость сходимости будет хуже, чем $\sqrt{\beta}$.

Анализ дискриминанта: Комплексные корни

Рассмотрим случай **комплексных корней** $(D < 0)$. Корни комплексно-сопряженные:

$$z_{1,2} = \frac{1 - \alpha\lambda_i + \beta \pm i\sqrt{4\beta - (1 - \alpha\lambda_i + \beta)^2}}{2}$$

Вычислим квадрат модуля корней:

$$|z_{1,2}|^2 = \left(\frac{1 - \alpha\lambda_i + \beta}{2}\right)^2 + \left(\frac{\sqrt{4\beta - (1 - \alpha\lambda_i + \beta)^2}}{2}\right)^2$$

$$= \frac{(1 - \alpha\lambda_i + \beta)^2 + 4\beta - (1 - \alpha\lambda_i + \beta)^2}{4} = \frac{4\beta}{4} = \beta$$

Следовательно, $|z_1| = |z_2| = \sqrt{\beta}$.

Вывод по дискриминанту

Вывод по дискриминанту

- В случае **комплексных корней** спектральный радиус $\rho(M_i) = \sqrt{\beta}$ и **не зависит от** λ_i.

Вывод по дискриминанту

* В случае **комплексных корней** спектральный радиус $\rho(M_i) = \sqrt{\beta}$ и **не зависит от** λ_i.
* В случае **вещественных корней** спектральный радиус $\rho(M_i) \geq \sqrt{\beta}$ и зависит от λ_i.

Вывод по дискриминанту

- В случае **комплексных корней** спектральный радиус $\rho(M_i) = \sqrt{\beta}$ и **не зависит от** λ_i.
- В случае **вещественных корней** спектральный радиус $\rho(M_i) \geq \sqrt{\beta}$ и зависит от λ_i.

Вывод по дискриминанту

- В случае **комплексных корней** спектральный радиус $\rho(M_i) = \sqrt{\beta}$ и **не зависит от** λ_i.
- В случае **вещественных корней** спектральный радиус $\rho(M_i) \geq \sqrt{\beta}$ и зависит от λ_i.

Стратегия: Мы хотим минимизировать худший спектральный радиус по всем λ_i. Наилучшая ситуация достигается, когда для всех λ_i корни комплексные (или на границе $D = 0$), и мы минимизируем $\sqrt{\beta}$. Поэтому мы требуем выполнения условия $D \leq 0$ для всех $\lambda_i \in [\mu, L]$.

Постановка задачи оптимизации

Мы ищем $\alpha > 0, \beta \geq 0$, минимизирующие спектральный радиус $\rho(\alpha, \beta) = \max_{\lambda \in [\mu, L]} \max(|z_1(\lambda)|, |z_2(\lambda)|)$.
Радиус корней для фиксированного λ:

$$r(\lambda) = \begin{cases} \frac{1}{2}\left(|1 + \beta - \alpha\lambda| + \sqrt{(1 + \beta - \alpha\lambda)^2 - 4\beta}\right), & \text{если } D > 0 \\ \sqrt{\beta}, & \text{если } D \leq 0 \end{cases}$$

Постановка задачи оптимизации

Мы ищем $\alpha > 0, \beta \geq 0$, минимизирующие спектральный радиус $\rho(\alpha, \beta) = \max_{\lambda \in [\mu, L]} \max(|z_1(\lambda)|, |z_2(\lambda)|)$. Радиус корней для фиксированного λ:

$$
r(\lambda) = \begin{cases} \frac{1}{2}\left(|1 + \beta - \alpha\lambda| + \sqrt{(1 + \beta - \alpha\lambda)^2 - 4\beta}\right), & \text{если } D > 0 \\ \sqrt{\beta}, & \text{если } D \leq 0 \end{cases}
$$

Обозначим $\alpha_{opt} = \left(\frac{2}{\sqrt{L} + \sqrt{\mu}}\right)^2$. Заметим, что $D \leq 0 \iff \beta \geq (1 - \sqrt{\alpha\lambda})^2$. Также $|1 - \sqrt{\alpha\mu}| < |1 - \sqrt{\alpha L}| \iff \alpha > \alpha_{opt}$.

Анализ случаев

Рассмотрим 4 случая в зависимости от α и β:

1. $0 < \alpha \leq \alpha_{opt}$ и $\beta \geq (1 - \sqrt{\alpha\mu})^2$. Тогда $\rho = \sqrt{\beta} \geq 1 - \sqrt{\alpha\mu} \geq \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.

Анализ случаев

Рассмотрим 4 случая в зависимости от α и β:

1. $0 < \alpha \leq \alpha_{opt}$ и $\beta \geq (1 - \sqrt{\alpha\mu})^2$. Тогда $\rho = \sqrt{\beta} \geq 1 - \sqrt{\alpha\mu} \geq \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.

Анализ случаев

Рассмотрим 4 случая в зависимости от α и β:

1. $0 < \alpha \leq \alpha_{opt}$ и $\beta \geq (1 - \sqrt{\alpha\mu})^2$. Тогда $\rho = \sqrt{\beta} \geq 1 - \sqrt{\alpha\mu} \geq \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.
2. $0 < \alpha \leq \alpha_{opt}$ и $\beta < (1 - \sqrt{\alpha\mu})^2$. Тогда $\rho \geq r(\mu) > 1 - \sqrt{\alpha\mu} \geq \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. (Здесь $r(\mu)$ убывает по β).

Анализ случаев (продолжение)

3. $\alpha > \alpha_{opt}$ и $\beta \geq (\sqrt{\alpha L} - 1)^2$. Тогда $\rho = \sqrt{\beta} \geq \sqrt{\alpha L} - 1 > \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$.

Анализ случаев (продолжение)

3. $\alpha > \alpha_{opt}$ и $\beta \geq (\sqrt{\alpha L} - 1)^2$. Тогда $\rho = \sqrt{\beta} \geq \sqrt{\alpha L} - 1 > \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.

Анализ случаев (продолжение)

3. $\alpha > \alpha_{opt}$ и $\beta \geq (\sqrt{\alpha L} - 1)^2$. Тогда $\rho = \sqrt{\beta} \geq \sqrt{\alpha L} - 1 > \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$.
4. $\alpha > \alpha_{opt}$ и $\beta < (\sqrt{\alpha L} - 1)^2$. Тогда $\rho \geq r(L) > \sqrt{\alpha L} - 1 > \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$. (Здесь $r(L)$ убывает по β).

Оптимальные параметры

Во всех случаях $\rho(\alpha, \beta) \geq \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. Равенство достигается только в первом случае на границе:

$$\alpha^* = \left(\frac{2}{\sqrt{L}+\sqrt{\mu}}\right)^2, \quad \beta^* = \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2$$

Оптимальные параметры

Во всех случаях $\rho(\alpha, \beta) \geq \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. Равенство достигается только в первом случае на границе:

$$\alpha^* = \left(\frac{2}{\sqrt{L} + \sqrt{\mu}}\right)^2, \quad \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2$$

При этом оптимальная скорость сходимости:

$$\rho_{opt} = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} = \frac{\sqrt{\varkappa} - 1}{\sqrt{\varkappa} + 1}$$

Это соответствует сложности $O(\sqrt{\varkappa} \log \frac{1}{\varepsilon})$.

Сходимость метода тяжёлого шарика для квадратичной функции

Theorem

Предположим, что f является μ-сильно выпуклой и L-гладкой квадратичной функцией. Тогда метод тяжёлого шарика с параметрами

$$\alpha = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}, \beta = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2$$

сходится линейно:

$$\|x_k - x^*\|_2 \leq \left(\frac{\sqrt{\varkappa} - 1}{\sqrt{\varkappa} + 1}\right)^k \|x_0 - x^*\|$$

Глобальная сходимость метода тяжёлого шарика [3]

Theorem

Предположим, что f является гладкой и выпуклой и что

$$
\beta \in [0, 1), \quad \alpha \in \left(0, \frac{2(1-\beta)}{L}\right).
$$

Тогда последовательность $\{x_k\}$, генерируемая итерациями тяжёлого шарика, удовлетворяет

$$
f(\overline{x}_T) - f^\star \leq \begin{cases} \frac{\|x_0 - x^\star\|^2}{2(T+1)} \left(\frac{L\beta}{1-\beta} + \frac{1-\beta}{\alpha} \right), & \text{if } \alpha \in (0, \dfrac{1-\beta}{L}], \\ \frac{\|x_0 - x^\star\|^2}{2(T+1)(2(1-\beta)-\alpha L)} \left(L\beta + \frac{(1-\beta)^2}{\alpha} \right), & \text{if } \alpha \in [\dfrac{1-\beta}{L}, \dfrac{2(1-\beta)}{L}), \end{cases}
$$

где $\overline{x}_T$ среднее Чезаро последовательности итераций, т.е.

$$
\overline{x}_T = \frac{1}{T+1} \sum_{k=0}^{T} x_k.
$$

[3] Глобальная сходимость метода тяжёлого шарика для выпуклой оптимизации, Euhanna Ghadimi et al.

Глобальная сходимость метода тяжёлого шарика [4]

Theorem

Предположим, что f является гладкой и сильно выпуклой и что

$$
\alpha \in \left(0, \frac{2}{L}\right), \quad 0 \leq \beta < \frac{1}{2}\left(\frac{\mu\alpha}{2} + \sqrt{\frac{\mu^2\alpha^2}{4} + 4(1 - \frac{\alpha L}{2})}\right).
$$

Тогда последовательность $\{x_k\}$, генерируемая итерациями методатяжёлого шарика, сходится линейно к единственному оптимальному решению $x^\star$. В частности,

$$
f(x_k) - f^\star \leq q^k (f(x_0) - f^\star),
$$

где $q \in [0, 1)$.

[4] Глобальная сходимость метода тяжёлого шарика для выпуклой оптимизации, Euhanna Ghadimi et al.

Итоги по методу тяжёлого шарика

- Обеспечивает ускоренную сходимость для сильно выпуклых квадратичных задач.

Итоги по методу тяжёлого шарика

- Обеспечивает ускоренную сходимость для сильно выпуклых квадратичных задач.
- Локально ускоренная сходимость была доказана в оригинальной статье.

Итоги по методу тяжёлого шарика

- Обеспечивает ускоренную сходимость для сильно выпуклых квадратичных задач.
- Локально ускоренная сходимость была доказана в оригинальной статье.
- Недавно [5] было доказано, что глобального ускорения сходимости для метода не существует.

[5] Provable non-accelerations of the heavy-ball method

Итоги по методу тяжёлого шарика

- Обеспечивает ускоренную сходимость для сильно выпуклых квадратичных задач.
- Локально ускоренная сходимость была доказана в оригинальной статье.
- Недавно [5] было доказано, что глобального ускорения сходимости для метода не существует.
- Метод не был чрезвычайно популярен до ML-бума.

[5] Provable non-accelerations of the heavy-ball method

Итоги по методу тяжёлого шарика

- Обеспечивает ускоренную сходимость для сильно выпуклых квадратичных задач.
- Локально ускоренная сходимость была доказана в оригинальной статье.
- Недавно [5] было доказано, что глобального ускорения сходимости для метода не существует.
- Метод не был чрезвычайно популярен до ML-бума.
- Сейчас он фактически является стандартом для практического ускорения методов градиентного спуска, в том числе для невыпуклых задач (обучение нейронных сетей).

[5] Provable non-accelerations of the heavy-ball method

Ускоренный градиентный метод Нестерова

Концепция ускоренного градиентного метода Нестерова

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \qquad x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta (x_k - x_{k-1}) \qquad \begin{cases} y_{k+1} = x_k + \beta (x_k - x_{k-1}) \\ x_{k+1} = y_{k+1} - \alpha \nabla f(y_{k+1}) \end{cases}$$

Концепция ускоренного градиентного метода Нестерова

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \qquad x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$$

$$\begin{cases} y_{k+1} = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_{k+1} - \alpha \nabla f(y_{k+1}) \end{cases}$$

Давайте определим следующие обозначения

$$x^+ = x - \alpha \nabla f(x) \qquad \text{Градиентный шаг}$$

$$d_k = \beta_k(x_k - x_{k-1}) \qquad \text{Импульс}$$

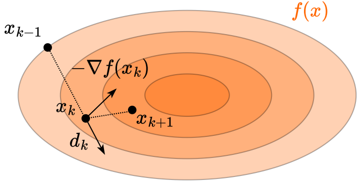
Тогда мы можем записать:

$$x_{k+1} = x_k^+ \qquad \text{Градиентный спуск}$$

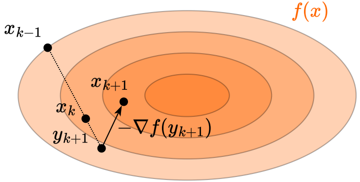
$$x_{k+1} = x_k^+ + d_k \qquad \text{Метод тяжёлого шарика}$$

$$x_{k+1} = (x_k + d_k)^+ \qquad \text{Ускоренный градиентный метод Нестерова}$$

Polyak momentum



Nesterov momentum



Сходимость для выпуклых функций

Theorem

Предположим, что $f : \mathbb{R}^n \to \mathbb{R}$ является выпуклой и L-гладкой. Ускоренный градиентный метод Нестерова (NAG) предназначен для решения задачи минимизации, начиная с начальной точки $x_0 = y_0 \in \mathbb{R}^n$ и $\lambda_0 = 0$. Алгоритм выполняет следующие шаги:

$$
\begin{aligned}
&\textbf{Обновление градиента:} && x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k) \\
&\textbf{Вес экстраполяции:} && \lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2} \\
& && \gamma_k = \frac{\lambda_k - 1}{\lambda_{k+1}} \\
&\textbf{Экстраполяция:} && y_{k+1} = x_{k+1} + \gamma_k \left(x_{k+1} - x_k\right)
\end{aligned}
$$

Последовательность $\{f(x_k)\}_{k \in \mathbb{N}}$, генерируемая алгоритмом, сходится к оптимальному значению f^* со скоростью $\mathcal{O}\left(\frac{1}{k^2}\right)$, в частности:

$$
f(x_k) - f^* \leq \frac{2L\|x_0 - x^*\|^2}{k^2}
$$

Ускоренная сходимость для сильно выпуклых функций

Theorem

Предположим, что $f : \mathbb{R}^n \to \mathbb{R}$ является μ-сильно выпуклой и L-гладкой. Ускоренный градиентный метод Нестерова (NAG) предназначен для решения задачи минимизации, начиная с начальной точки $x_0 = y_0 \in \mathbb{R}^n$ и $\lambda_0 = 0$. Алгоритм выполняет следующие шаги:

Обновление градиента: $$x_{k+1} = y_k - \frac{1}{L} \nabla f(y_k)$$

Экстраполяция: $$y_{k+1} = x_{k+1} - \gamma \left(x_{k+1} - x_k\right)$$

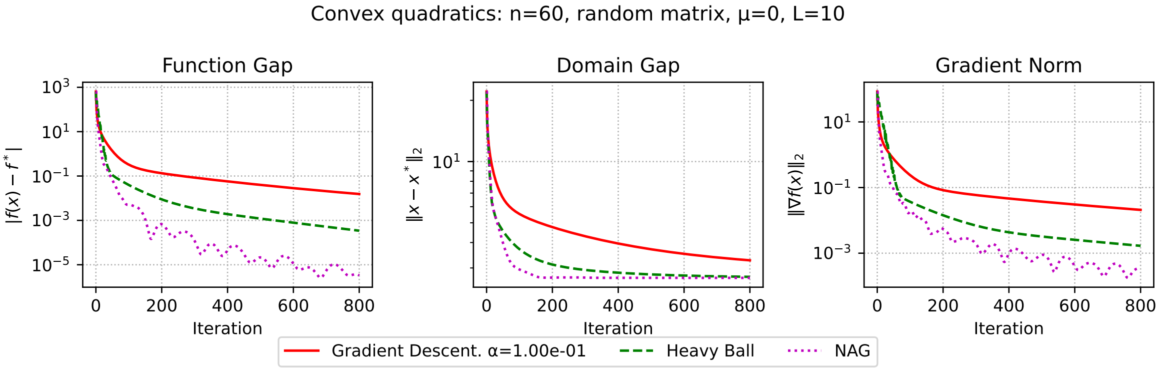
Вес экстраполяции: $$\gamma = \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}$$

Последовательность $\{f(x_k)\}_{k \in \mathbb{N}}$, генерируемая алгоритмом, сходится к оптимальному значению f^* линейно:

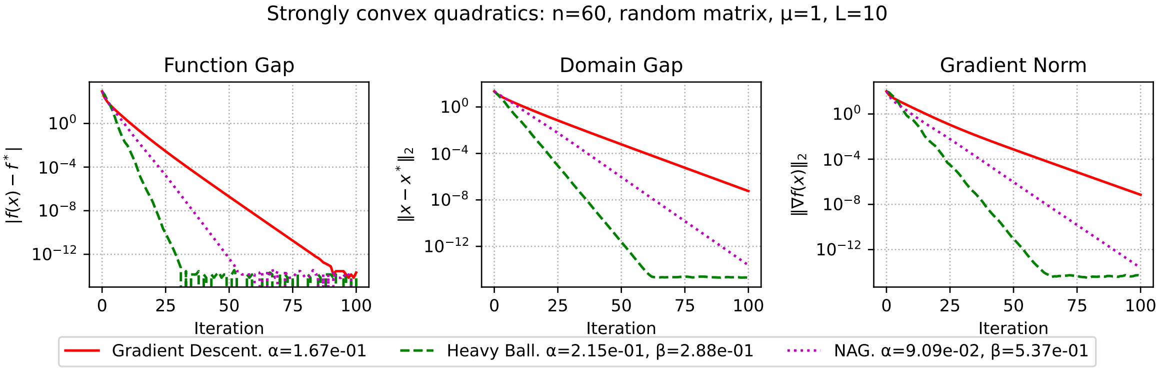
$$f(x_k) - f^* \leq \frac{\mu + L}{2} \|x_0 - x^*\|_2^2 \exp\left(-\frac{k}{\sqrt{\varkappa}}\right)$$

Численные эксперименты

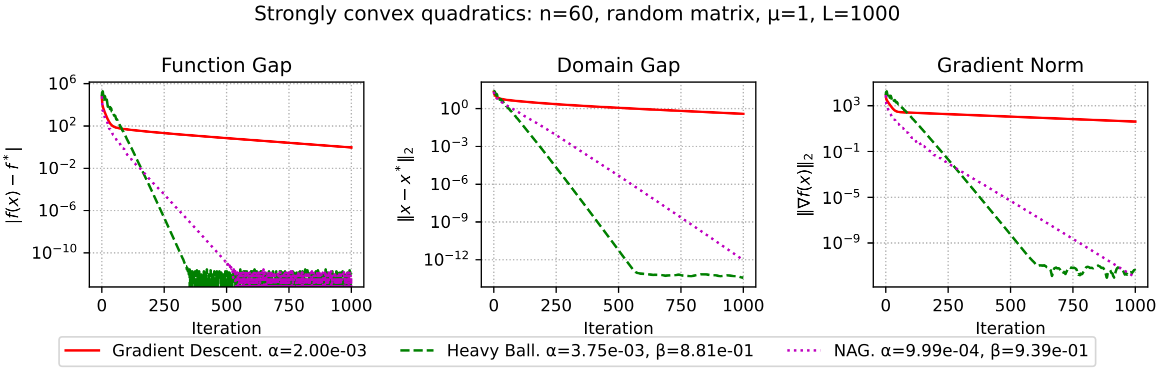
Выпуклая квадратичная задача (линейная регрессия)



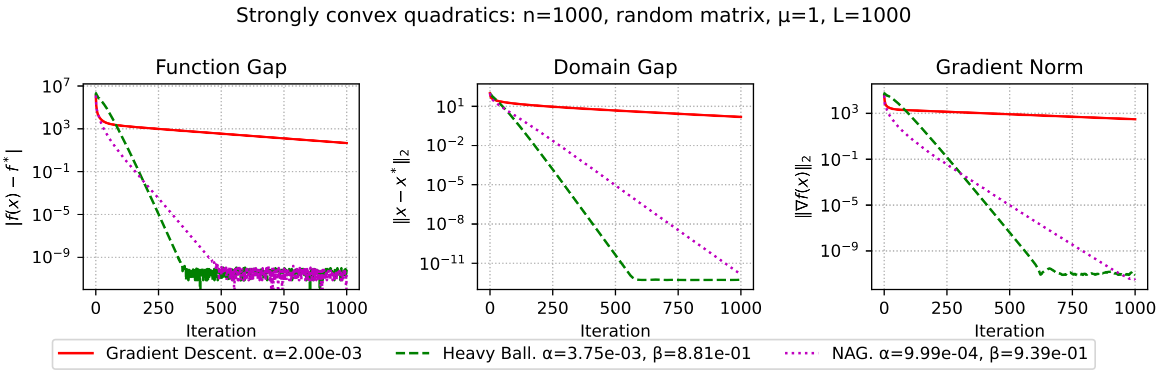
Сильно выпуклая квадратичная задача (регуляризованная линейная регрессия)



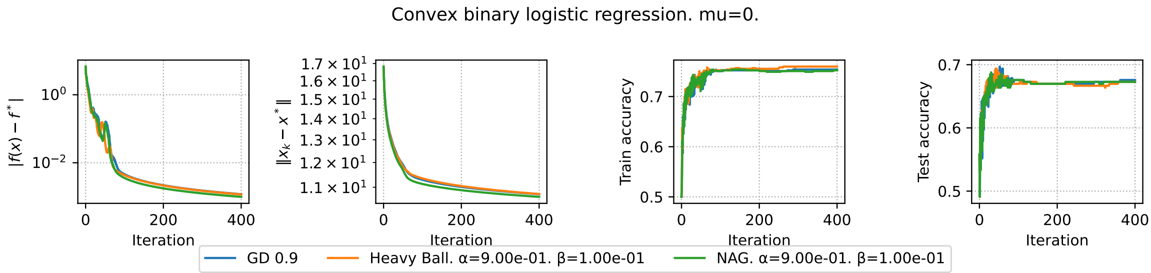
Сильно выпуклая квадратичная задача (регуляризованная линейная регрессия)



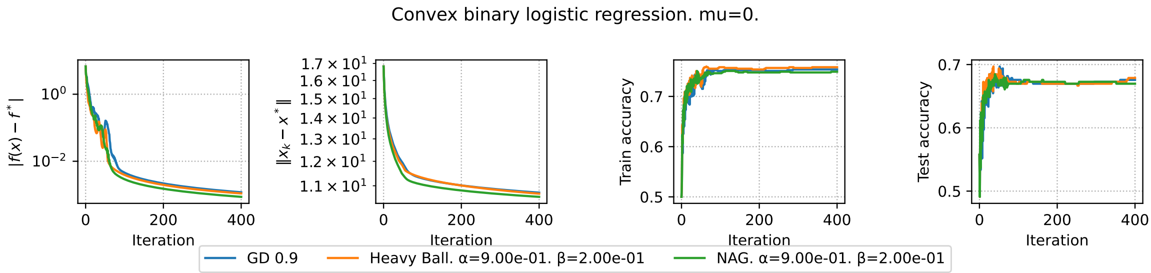
Сильно выпуклая квадратичная задача (регуляризованная линейная регрессия)



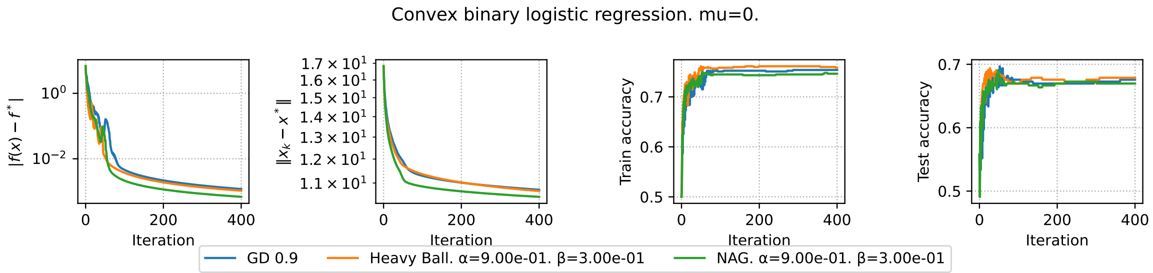
Выпуклая бинарная логистическая регрессия



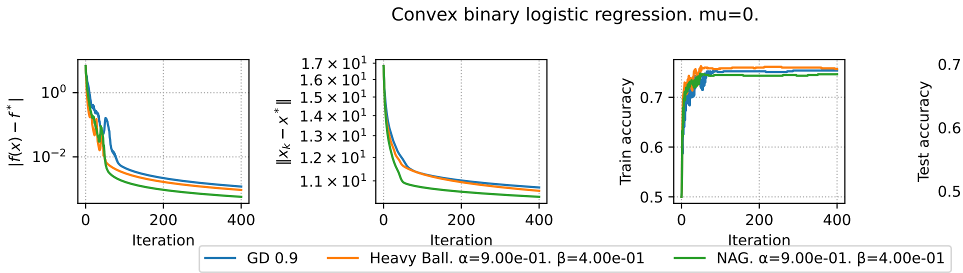
Выпуклая бинарная логистическая регрессия



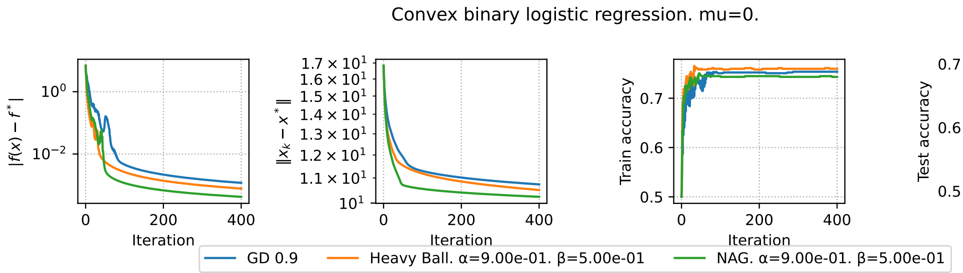
Выпуклая бинарная логистическая регрессия



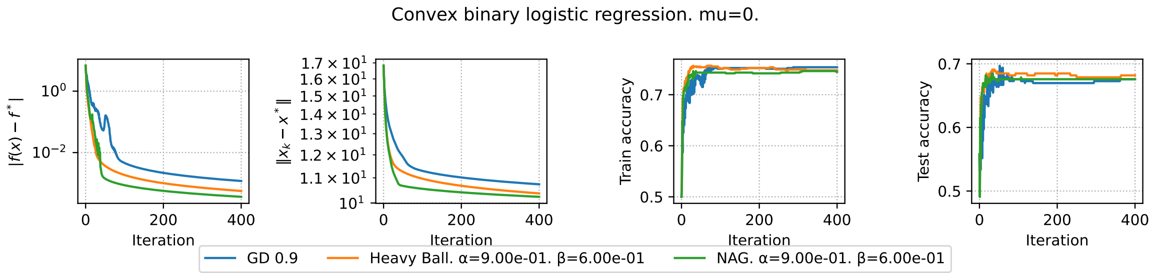
Выпуклая бинарная логистическая регрессия



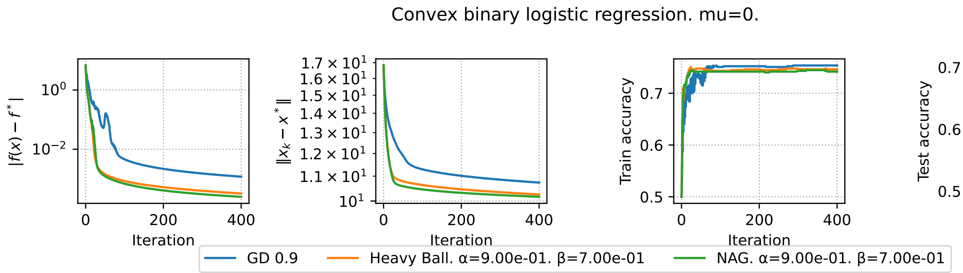
Выпуклая бинарная логистическая регрессия



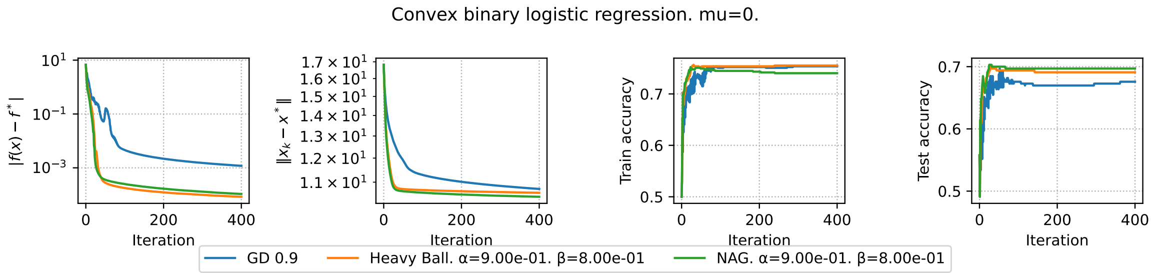
Выпуклая бинарная логистическая регрессия



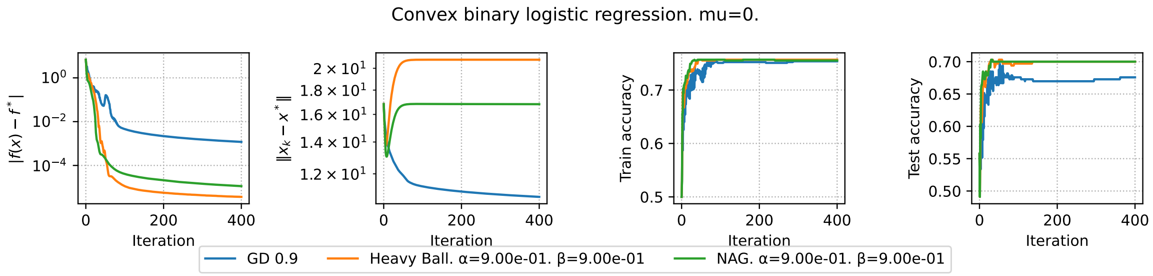
Выпуклая бинарная логистическая регрессия



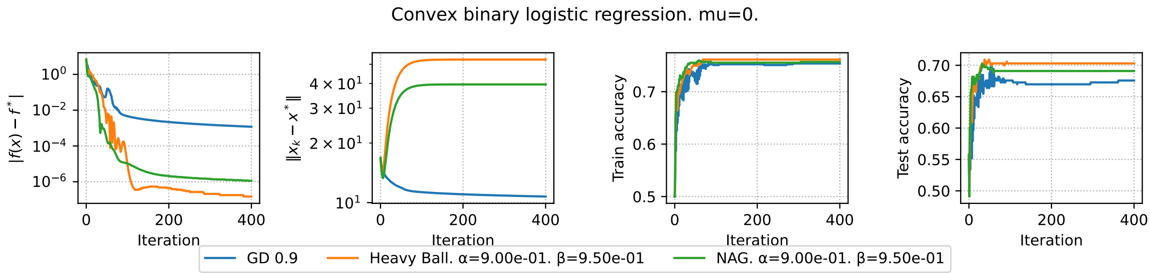
Выпуклая бинарная логистическая регрессия



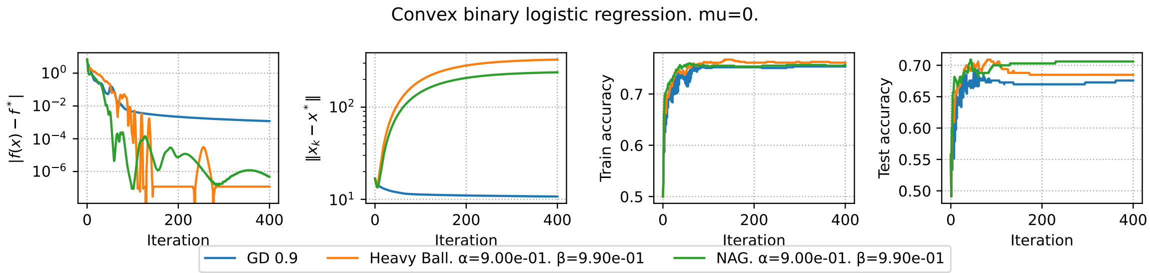
Выпуклая бинарная логистическая регрессия



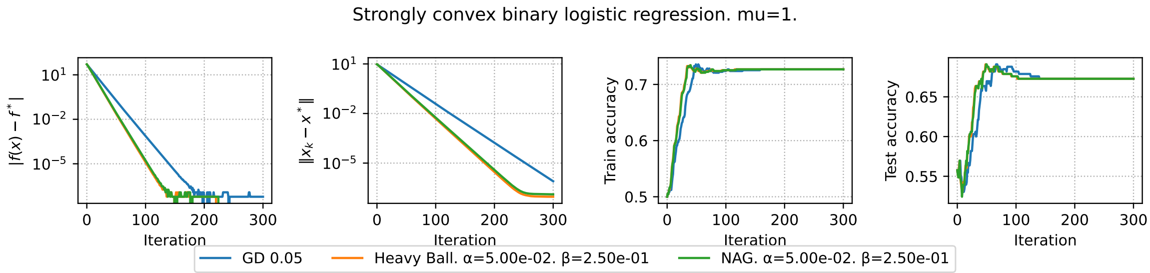
Выпуклая бинарная логистическая регрессия



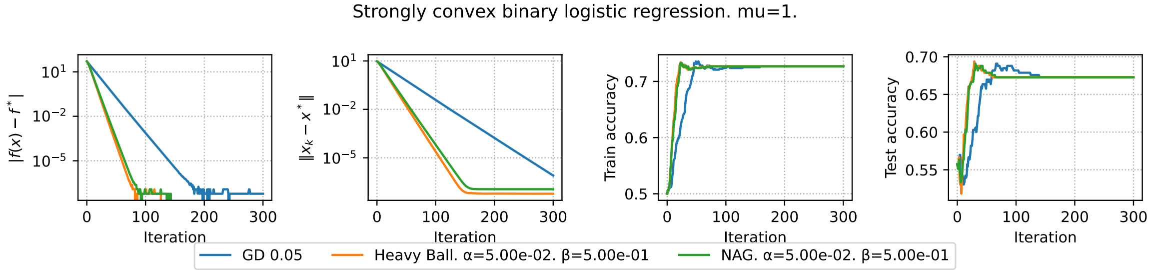
Выпуклая бинарная логистическая регрессия



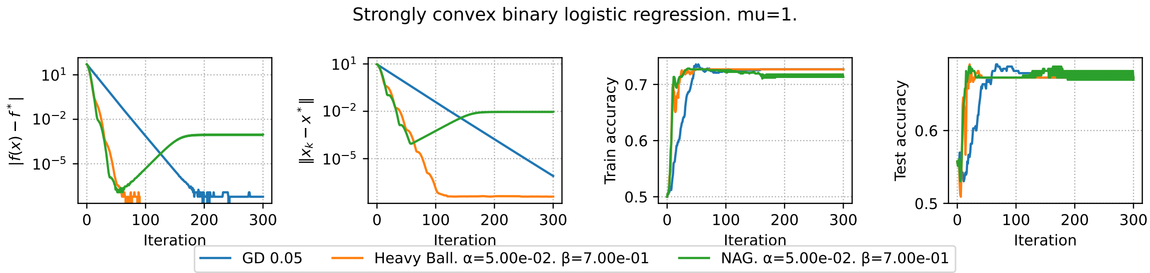
Сильно выпуклая бинарная логистическая регрессия



Сильно выпуклая бинарная логистическая регрессия



Сильно выпуклая бинарная логистическая регрессия



Сильно выпуклая бинарная логистическая регрессия