

МЕТОДЫ ОПТИМИЗАЦИИ

Regularized Nonlinear Acceleration

Чернис Константин, группа 694

Содержание

1	Anderson acceleration	2
2	Regularized nonlinear acceleration	4
3	Численные эксперименты	5
3.1	Ускорение градиентного спуска для квадратичной функции	5
3.2	Ускорение Ridge регрессии	6
3.3	Ускорение метода Ньютона	7
4	Список литературы	8

В данном проекте представлен алгоритм нелинейного ускорения итерационных методов. Также теоретически и экспериментально исследовано влияние алгоритма на скорость сходимости различных методов первого и второго порядков.

Для начала рассмотрим более простой алгоритм ускорения, а именно, ускорение Андерсона.

1. Anderson acceleration

Пусть необходимо решить задачу оптимизации:

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

где $f(x)$ сильно выпукла с константой μ , а её градиент липшецев с константой L . Будем искать решение с помощью метода неподвижной точки:

$$x_{i+1} = g(x_i), \quad i = \overline{0, k} \quad (1)$$

Пусть $g(x)$ дифференцируема и G — матрица Якоби функции g в точке x^* . Далее считаем G симметричной положительно определённой матрицей, $G \preceq \sigma I$, где $\sigma < 1$. Тогда из (1) получаем линейный метод неподвижной точки:

$$x_{i+1} = g(x^*) + G(x_i - x^*) + O(\|x_i - x^*\|_2^2), \quad i = \overline{1, n} \quad (2)$$

Пренебрегая вторым порядком малости и учитывая, что x^* — неподвижная точка $g(x)$, то есть $g(x^*) = x^*$, получаем

$$x_{i+1} - x^* = G(x_i - x^*)$$

В силу того, что $\|G\|_2 \leq \sigma < 1$, получаем линейную скорость сходимости:

$$\|x_i - x^*\|_2 \leq \sigma \|x_{i-1} - x^*\|_2 \leq \sigma^i \|x_0 - x^*\|_2$$

Рассмотрим линейную комбинацию x_i после k итераций:

$$\sum_{i=0}^k c_i x_i = \sum_{i=0}^k c_i x^* + \sum_{i=0}^k c_i G(x_i - x^*) = \left(\sum_{i=0}^k c_i \right) x^* + \left(\sum_{i=0}^k c_i G^i \right) (x_0 - x^*)$$

и определим многочлен

$$p(z) := \sum_{i=0}^k c_i z^i$$

Теперь линейную комбинацию можно записать с помощью матричного многочлена $p(G)$, добавив ограничение $p(1) = \sum_{i=0}^k c_i = 1$:

$$\sum_{i=0}^k c_i x_i = x^* + \underbrace{p(G)(x_0 - x^*)}_{\text{ошибка}} \quad (3)$$

Будем искать коэффициенты c (или p соответственно), которые минимизируют ошибку:

$$c^* = \arg \min_{\{c \in \mathbb{R}^{k+1} : c^T \mathbf{1} = 1\}} \left\| \sum_{i=0}^k c_i G^i (x_0 - x^*) \right\|_2 = \arg \min_{\{p \in \mathbb{R}_k[x] : p(1) = 1\}} \|p(G)(x_0 - x^*)\|_2$$

где $\mathbb{R}_k[x]$ — пространство многочленов степени не выше k .

Следующее предложение даёт оценку сверху на размер ошибки:

Предложение 1.1. Пусть

1. последовательность x_i , $i = \overline{0, n}$, получена из (2);
2. G — симметричная матрица Якоби g , для которой выполнено $0 \preceq G \preceq \sigma I$, $\sigma < 1$;
3. x^* — неподвижная точка g .

Тогда l_2 норма ошибки (3) ограничена:

$$\left\| \sum_{i=0}^k c_i^* x_i - x^* \right\|_2 \leq \begin{cases} \frac{2\beta^k}{1 + \beta^{2k}} \|x_0 - x^*\|_2, & \text{если } k < m \\ 0 & \text{иначе,} \end{cases} \quad (4)$$

где m — число различных собственных значений G и

$$\beta = \frac{1 - \sqrt{1 - \sigma}}{1 + \sqrt{1 - \sigma}} < 1$$

Данная оценка получена из полу-итерационного метода Чебышёва, после преобразования получаем

$$\left\| \sum_{i=0}^k c_i^* x_i - x^* \right\|_2 \leq (1 - \sqrt{1 - \sigma})^k \|x_0 - x^*\|_2 \ll \sigma^k \|x_0 - x^*\|_2,$$

так что достигнуто ускорение. В то же время, для применения этого метода требуется знание оценки σ матрицы Якоби G в точке x^* , кроме того, коэффициенты в линейной комбинации могут быть очень большими, что влияет на численную стабильность алгоритма.

В связи с перечисленными выше проблемами далее сосредоточимся на методе, который будет приближённо минимизировать ошибку (3). В силу того, что G и x^* неизвестны, будем работать с невязкой:

$$r_i = x_{i+1} - x_i = g(x_i) - x_i,$$

тогда линейный метод неподвижной точки (2) принимает вид

$$r_i = x_{i+1} - x_i = (G - I)(x_i - x^*),$$

а линейная комбинация записывается как

$$\sum_{i=0}^k c_i r_i = (G - I) \sum_{i=0}^k c_i (x_i - x^*) = (G - I)p(G)(x_0 - x^*),$$

что равно ошибке $p(G)(x_0 - x^*)$, умноженной на $(G - I)$, то есть использование этих коэффициентов будет приближённо минимизировать ошибку.

Предложение 1.2. Пусть

$$c^* = \arg \min_{\{c \in \mathbb{R}^{k+1} : c^T \mathbf{1} = 1\}} \left\| \sum_{i=0}^k c_i r_i \right\|_2$$

Тогда последовательность x_i , $i = \overline{0, k}$, усреднённая с помощью коэффициентов c^* , удовлетворяет соотношению

$$\left\| \sum_{i=0}^k c_i^* x_i - x^* \right\|_2 \leq \frac{1}{1 - \sigma} \arg \min_{\{c \in \mathbb{R}^{k+1} : c^\top \mathbf{1} = 1\}} \left\| \sum_{i=0}^k c_i G^i (x_0 - x^*) \right\|_2, \quad (5)$$

где $0 \preceq G \preceq \sigma I$, $\sigma < 1$.

Отсюда получаем ускорение Андерсона:

Algorithm 1: Anderson acceleration

Data: Последовательность $x_0, x_1, \dots, x_{k+1} \in \mathbb{R}^d$.

- 1 Составить матрицу $R = [r_0, \dots, r_k]$;
- 2 Решить задачу

$$c^* = \arg \min_{\{c \in \mathbb{R}^{k+1} : c^\top \mathbf{1} = 1\}} \left\| \sum_{i=0}^k c_i r_i \right\|_2$$

Result: Аппроксимация $\hat{x}^* = \sum_{i=0}^k c_i^* x_i$, удовлетворяющая (5).

Из ККТ можно вывести, что решение получается в 2 шага:

1. Решить $R^\top R z = \mathbf{1}$
2. $c^* = z / (\mathbf{1}^\top z)$

2. Regularized nonlinear acceleration

Чтобы повысить численную стабильность алгоритма и обеспечить его работу, если функция g содержит шум вида $x_{i+1} - x_i = g(x_i) - x_i = G(x_i - x^*) + e_i$, добавим регуляризацию в предложенный выше алгоритм:

Algorithm 2: Regularized Nonlinear Acceleration (RNA)

Data: Последовательность $x_0, x_1, \dots, x_{k+1} \in \mathbb{R}^d$, полученная из метода неподвижной точки, и регуляризационный параметр $\lambda > 0$.

- 1 Составить матрицу $R = [r_0, \dots, r_k]$, где $r_i = x_{i+1} - x_i$;
- 2 Решить задачу

$$c^* = \arg \min_{c^\top \mathbf{1} = 1} \|Rc\|_2^2 + \lambda \|c\|_2^2,$$

или, что эквивалентно, решить $(R^\top R + \lambda I)z = \mathbf{1}$ и взять $c_\lambda^* = z / \mathbf{1}^\top z$.

Result: Аппроксимация $\hat{x}^* = \sum_{i=0}^k (c_\lambda^*)_i x_i$, удовлетворяющая (5).

Наконец, добавим поиск по сетке для выбора λ и backtracking для выбора размера шага:

$$\min_{t>0} f(x_0 + t(x_{extr}(\lambda) - x_0))$$

Algorithm 3: Adaptive Regularized Nonlinear Acceleration (ARNA)

Data: Последовательность $x_0, x_1, \dots, x_{k+1} \in \mathbb{R}^d$, полученная из метода неподвижной точки, границы $[\lambda_{\min}, \lambda_{\max}]$ и целевая функция $f(x)$.

- 1 Разбить отрезок $[\lambda_{\min}, \lambda_{\max}]$ на k частей, используя логарифмический масштаб;
 - 2 Составить матрицу $R = [r_0, \dots, r_k]$, где $r_i = x_{i+1} - x_i$;
 - 3 Построить матрицу $M = R^T R / \|R^T R\|_2$;
 - 4 **for** $j = \overline{1, k}$ **do**
 - 5 | Решить систему $(M + \lambda_j)z = \mathbf{1}$;
 - 6 | Нормировать решение: $c_{\lambda_j}^* = z / \mathbf{1}^T z$;
 - 7 | Вычислить $x_{extr}(\lambda_j) = \sum_{i=0}^k (c_{\lambda_j}^*)_i x_i$;
 - 8 **end**
 - 9 Выбрать $x_{extr}^* = \arg \min_{j=\overline{1, k}} f(x_{extr}(\lambda_j))$;
 - 10 Задать $F_t = f(x_0 + t(x_{extr}^* - x_0))$;
 - 11 $t := 1$;
 - 12 **while** $F_{2t} < F_t$ **do**
 - 13 | $t = 2t$;
 - 14 **end**
- Result:** Аппроксимация $(x_0 + t(x_{extr}^* - x_0))$

3. Численные эксперименты

3.1 Ускорение градиентного спуска для квадратичной функции

Рассмотрим квадратичную функцию $f = \frac{1}{2}x^T A x - b^T x$, тогда её градиент имеет вид

$$\nabla f(x) = Ax - b = A(x - x^*),$$

где последний переход верен в силу $Ax^* = b$. Пусть $\mu I \preceq A \preceq \sigma I$, тогда градиентный спуск с шагом $1/L$ записывается как

$$x_{i+1} = x_i - \frac{1}{L} \nabla f(x_i) = x_i - \frac{1}{L} A(x_i - x^*) = \underbrace{(I - A/L)}_G (x_i - x^*) + x^*,$$

откуда $0 \preceq G \preceq \left(1 - \frac{\mu}{L}\right) I$

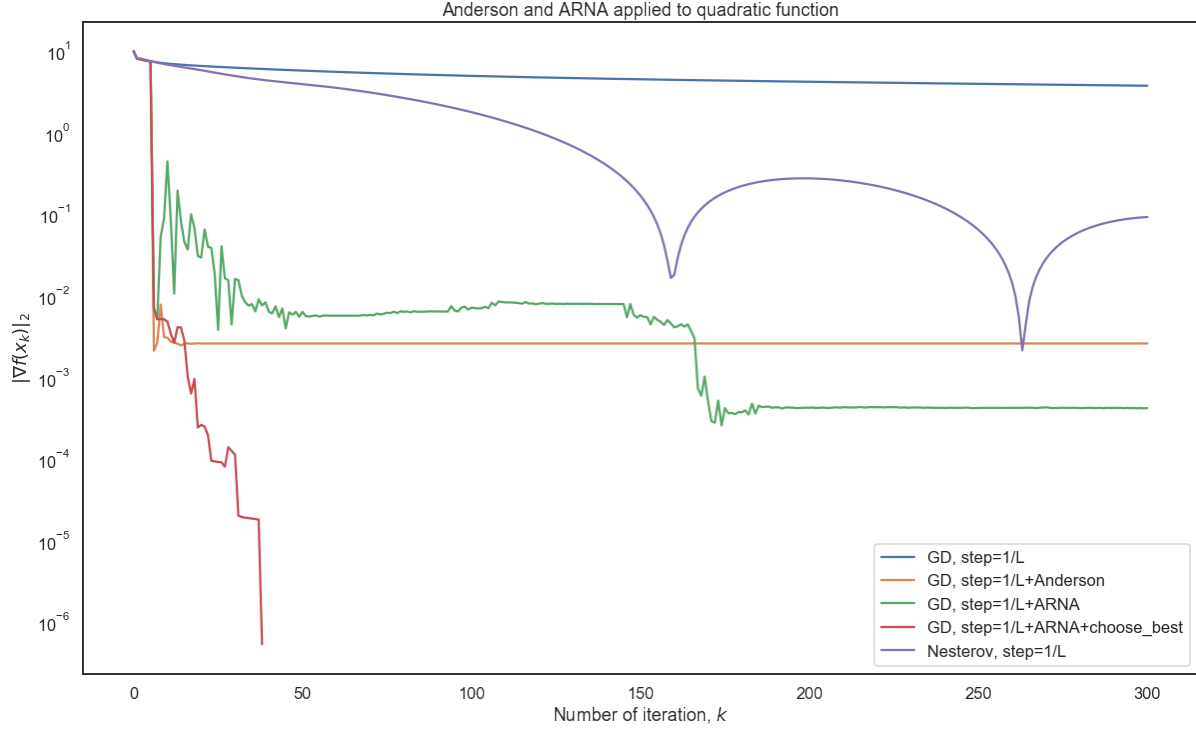
Используя Предложение 1.1 и Предложение 1.2, получаем следующую оценку на скорость сходимости градиентного спуска, ускоренного алгоритмом Андерсона:

$$\left\| \sum_{i=0}^N c_i^* x_i - x^* \right\|_2 \leq \frac{L}{\mu} \frac{2\beta^k}{1 + 2\beta^k} \|x_0 - x^*\|_2, \quad \text{где } \beta = \frac{1 - \sqrt{\frac{\mu}{L}}}{1 + \sqrt{\frac{\mu}{L}}}$$

Преобразовывая, получаем

$$\left\| \sum_{i=0}^N c_i^* x_i - x^* \right\|_2 \lesssim \frac{L}{\mu} \cdot 2\beta^k \|x_0 - x^*\|_2,$$

что соответствует оценке сходимости оптимального метода Нестерова [2]. Продемонстрируем скорость сходимости на практике:



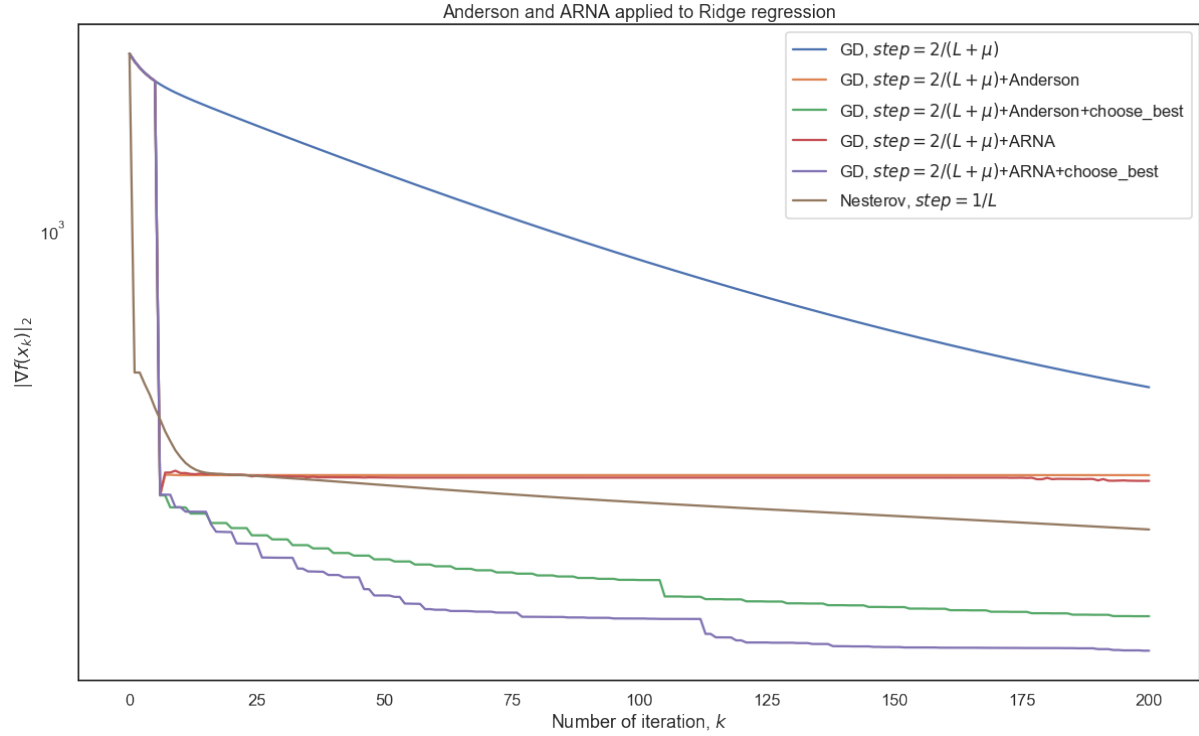
Более медленную сходимость ARNA по сравнению с Anderson на начальном этапе можно объяснить неточными значениями коэффициентов из-за присутствия регуляризации. В остальном ARNA демонстрирует отличную скорость сходимости (для Anderson *choose_best* не применялся в силу того, что эта опция не исправляет застревание в этом случае).

3.2 Ускорение Ridge регрессии

Применим ускорение к оптимизации Ridge регрессии для стандартного датасета boston. Функция потерь задаётся как

$$f(w, \lambda) = \sum_{i=1}^m (Z^T w - y)^2 + \frac{\lambda}{2} \|w\|_2^2,$$

она строго выпукла с $\mu = \lambda$, а её градиент липшецев с константой $L = \|2Z^T Z + \lambda I\|_2$. Снова получаем сходимость, аналогичную методу Нестерова:



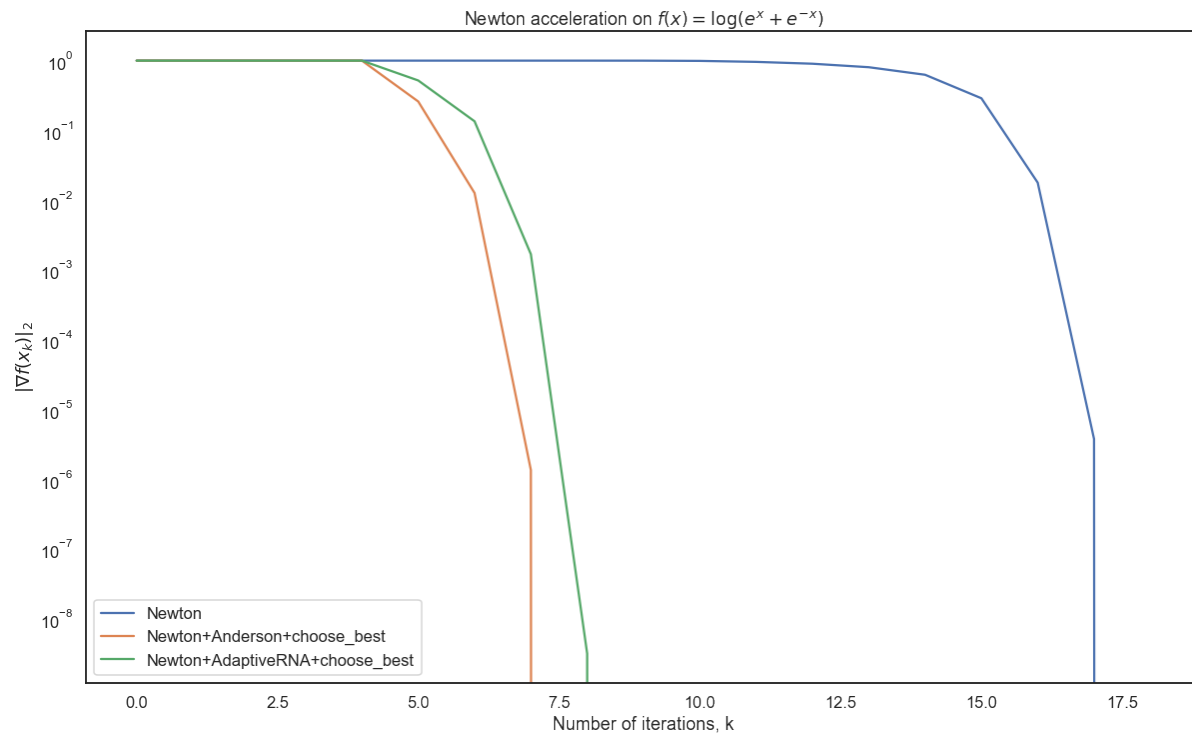
Заметим, что хоть на начальном этапе ускоренные алгоритмы и вырываются вперёд, в дальнейшем характер сходимости почти не отличается от метода Нестерова. Также видно, что *choose_best* оказывает критическое влияние на сходимость.

3.3 Ускорение метода Ньютона

Важным отличием представленных в проекте методов ускорения является их применимость практически к любым методам, основанным на принципе неподвижной точки. Для демонстрации этого свойства оптимизируем функцию

$$f(x) = \log(e^x + e^{-x})$$

с помощью метода Ньютона и его ускоренных версий. В данном случае методы помогают достичь области квадратичной сходимости сразу после накопления последовательности $\{x_i\}$ достаточной длины, чуть более медленная сходимость ARNA объясняется более удачным попаданием в эту область, ибо в дальнейшем, как видно в сравнении с обычным методом Ньютона, ускорение не участвует в сходимости в силу опции *choose_best*.



4. Список литературы

- [1] Scieur, D., d'Aspremont, A., and Bach, F. (2016). "Regularized Nonlinear Acceleration". ArXiv e-prints, arXiv:1606.04133.
- [2] Nesterov, Y. (2013). "Introductory lectures on convex optimization: A basic course". Vol. 87, Springer Science & Business Media.