

## Второе задание по методам оптимизации

Иван Ермаков

### 3

Логистическая регрессия

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{m} \langle \ln [1 + \exp(-b \odot (Ax))] , 1_m \rangle + \frac{\lambda}{2} \|x\|^2 \right\} \quad (1)$$

Здесь  $b$  — вектор из  $b_i$ ,  $A^T$  — матрица  $m \times n$ , в которой строки являются векторами  $a_i$ .

$$\begin{aligned} Df(x) &= \lambda x - \frac{A^T}{m} \left( b \odot \frac{\exp(-b \odot (Ax))}{1 + \exp(-b \odot (Ax))} \right) = \lambda x - \frac{A^T}{m} (b \odot u[-b \odot (Ax)]) \\ u(y) &= \frac{e^y}{1 + e^y} = \frac{1}{1 + e^{-y}} \end{aligned} \quad (2)$$

Здесь использовано много свойств произведения Адамара.

$$\frac{du}{dy} = \frac{e^y(1 + e^y) - e^{2y}}{(1 + e^y)^2} = \frac{e^y}{(1 + e^y)^2} \quad (3)$$

$$D^2 f[dx] = \lambda dx + \frac{A^T}{m} \left( b \odot \frac{e^y}{(1 + e^y)^2} \odot b \odot (Adx) \right) \quad (4)$$

$$D^2 f = \lambda + \frac{A^T}{m} \left\{ \left[ \left( b \odot b \odot \frac{e^y}{(1 + e^y)^2} [-b \odot (Ax)] \right) 1_n^T \right] \odot A \right\} \quad (5)$$

Потратив еще *немного* времени, можно понять, что это выражение можно переписать как

$$D^2 f = \lambda + \frac{1}{m} \cdot A^T \text{diag} \{ b \odot b \odot v(-b \odot (Ax)) \} A \quad (6)$$

где

$$v(y) = \frac{e^y}{(1 + e^y)^2} \quad (7)$$

## 1 Эксперимент: траектория градиентного спуска

Построение графиков выполнено при помощи скрипта “gradplots/trajectory.py”. Для хорошо обусловленной матрицы поиск останавливается за 109000 шагов, а для плохо обусловленной за 69000. Скорее всего так происходит из-за того, что в хорошо обусловленной матрице метод постоянно проскакивает оптимальную точку из-за высокой требуемой точности.

Теперь все встало на свои места: для хорошо обусловленной матрицы метод сходится всего за 18 шагов. Для плохо обусловленной требуется 327 шагов. Видно, что в первом случае довольно быстро траектория становится прямой, а во втором случае она все время колеблется вокруг оптимального пути и совершает больше шагов, чем нужно.

Для условия Вольфа ситуация аналогичная первому случаю: для хорошо обусловленной матрицы метод работает “слишком хорошо” и проскакивает точку экстремума. Метод сходится за 18 шагов. Для плохо обусловленной матрицы требуется всего 4 шага. Понятно, что в общем случае для хорошо обусловленной матрицы метод Вольфа тоже работает хорошо, и некоторое  $O(1)$  количество операций в конце, связанное с проскакиванием, роли не играет.

При этом, конечно, чем ближе прямая градиента к оптимальной точке, тем быстрее будут работать все методы, потому что траектория будет ближе к прямой.

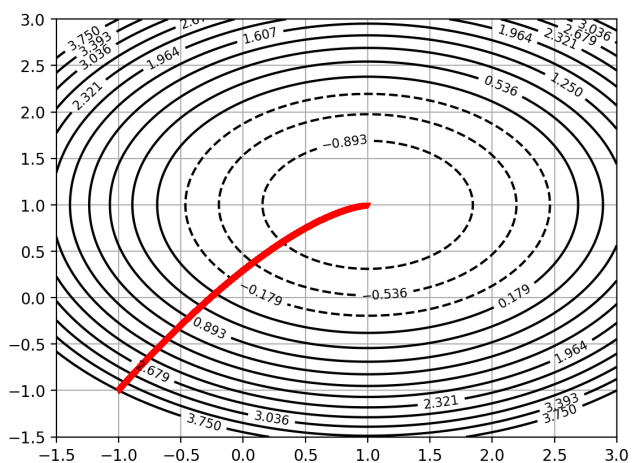


Рис. 1: Постоянный шаг, хорошо обусловленная матрица.

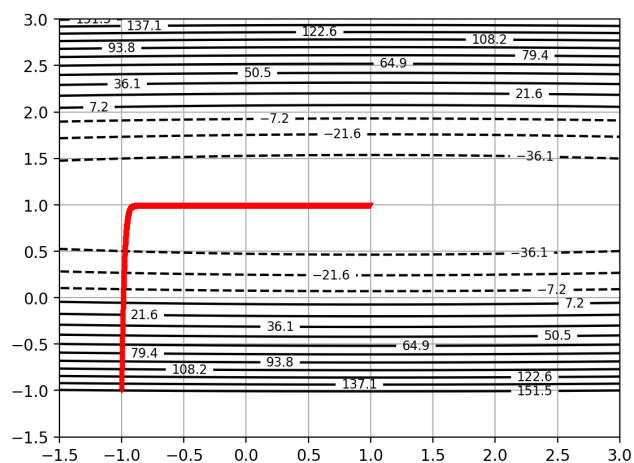


Рис. 2: Постоянный шаг, плохо обусловленная матрица.

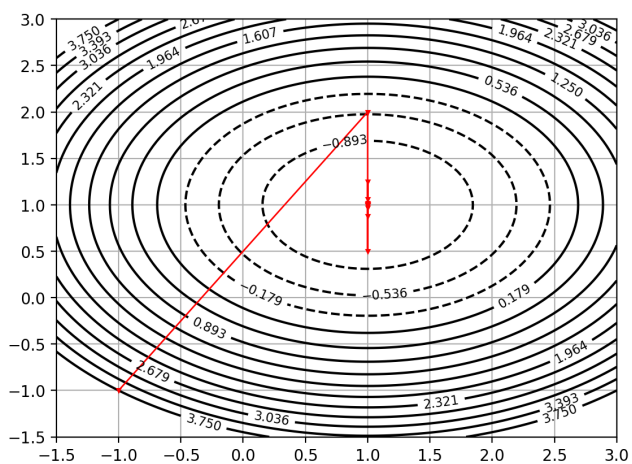


Рис. 3: Условие Армихо, хорошо обусловленная матрица.

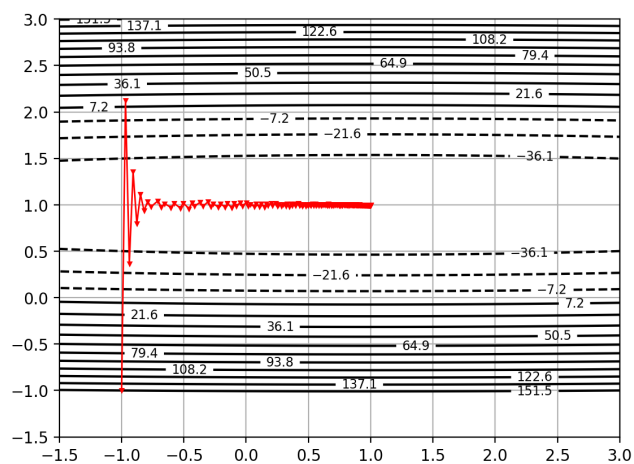


Рис. 4: Условие Армихо, плохо обусловленная матрица.

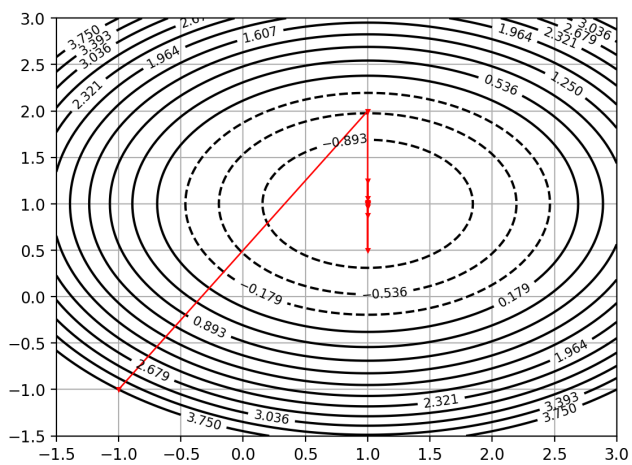


Рис. 5: Условие Вольфа, хорошо обусловленная матрица.

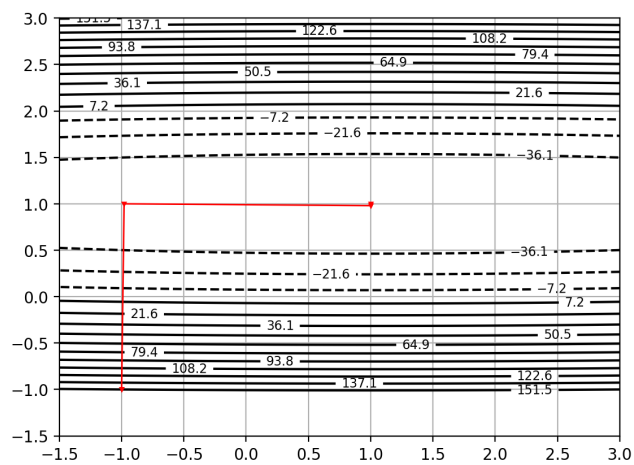


Рис. 6: Условие Вольфа, плохо обусловленная матрица.