Адаптация стохастического метода оптимизации SFO для линейных моделей в машинном обучении

Родоманов Антон Олегович,
Кропотов Дмитрий Александрович

Студент, научный сотрудник

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: anton.rodomanov@gmail.com, dmitry.kropotov@gmail.com

Многие задачи, возникающие в машинном обучении, можно переформулировать как задачи оптимизации вида

$$\sum_{j=1}^{N} f_j(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^D}, \tag{1}$$

где f_j есть функция потерь на j-м объекте обучающей выборки. Одной из основных сложностей, возникающих при оптимизации подобного рода функций, является то, что количество объектов обучающей выборки N может быть очень большим, в то время как среди самих объектов имеется много избыточности.

Эффективным методом оптимизации функций вида (1) является предложенный в 2014 г. метод SFO [1]. Для применения этого метода обучающая выборка случайным образом равномерно разбивается на $M \approx \sqrt{N}/10$ групп, т. е. задача (1) преобразуется к следующему эквивалентному виду:

$$\sum_{i=1}^{M} F_i(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^D} \tag{2}$$

$$F_i(\mathbf{x}) := \sum_{j\in\mathcal{S}_i} f_j(\mathbf{x}), \tag{3}$$

где $\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_M = \{1, \ldots, N\}$, $\mathcal{S}_{i_1} \cap \mathcal{S}_{i_2} = \varnothing$, $i_1 \neq i_2$. Для каждой функции F_i дополнительно хранится следующая квадратичная модель:

$$Q_i^k(\mathbf{x}) := c_i^k + (\mathbf{g}_i^k)^\top(\mathbf{x} - \mathbf{v}_i^k) + \frac{1}{2}(\mathbf{x} - \mathbf{v}_i^k)^\top \mathbf{H}_i^k(\mathbf{x} - \mathbf{v}_i^k), \tag{4}$$

где $\mathbf{H}_i^k$ есть симметричная положительно определенная матрица. Итерация метода SFO заключается в применении одного шага (демп-

фированного) метода Ньютона к функции $Q^k(\mathbf{x}) := \sum_{i=1}^{M} Q_i^k(\mathbf{x})$; затем происходит обновление одной случайно выбранной модели Q_i: $(\mathbf{v}_i^{k+1}, c_i^{k+1}, \mathbf{g}_i^{k+1}, \mathbf{H}_i^{k+1}) := (\mathbf{x}_{k+1}, F_i(\mathbf{x}_{k+1}), \nabla \mathbf{F}_i(\mathbf{x}_{k+1}), \nabla^{\mathbf{2}} \mathbf{F}_i(\mathbf{x}_{k+1}))$.

Для каждой модели Q_i в методе SFO напрямую хранятся $\mathbf{v}_i$, c_i и $\mathbf{g}_i$. Матрицы $\mathbf{H}_i$ напрямую не хранятся, поскольку это бы потребовало слишком много памяти. Вместо этого для каждой матрицы $\mathbf{H}_i$ хранится ее низкоранговая BFGS-аппроксимация, т. е. некоторая история из небольшого числа L (например, 10) векторов. При таком неявном хранении для всех $\mathbf{H}_i$ суммарно требуется $O(MLD)$ памяти.

В данной работе показывается, что для линейных моделей, т. е. когда $f_j(\mathbf{x}) = r_j(\mathbf{a}_j^\top \mathbf{x})$ для некоторой функции r_j и некоторого вектора $\mathbf{a}_j$, метод SFO можно ускорить. Будем считать, что число переменных D является небольшим и что по крайней мере одну матрицу размеров $D \times D$ возможно хранить в памяти (в этом случае в методе SFO не нужно использовать вспомогательное подпространство низкой размерности).

Для эффективного хранения матриц $\mathbf{H}_i$ вместо низкоранговой BFGS-аппроксимации предлагается использовать структуру самой задачи. В самом деле, т. к. $f_j(\mathbf{x}) = r_j(\mathbf{a}_j^\top \mathbf{x})$, то

$$\mathbf{H}_i = \sum_{j \in \mathcal{S}_i} r_j''(\mathbf{a}_j^\top \mathbf{x}) \mathbf{a}_j \mathbf{a}_j^\top. \tag{5}$$

Поскольку векторы $\mathbf{a}_j$ известны (хранятся в памяти), то для неявного хранения матриц $\mathbf{H}_i$ достаточно хранить в памяти лишь коэффициенты разложения $r_j''(\mathbf{a}_j^\top \mathbf{x})$. Суммарно для такого хранения всех $\mathbf{H}_i$ потребуется $O(N)$ памяти. Заметим, что равенство в формуле (5) является точным, в то время как в SFO в этом месте используется низкоранговая BFGS-аппроксимация.

Проведенные численные эксперименты на задаче обучения логистической регрессии показывают, что предложенная адаптация SFO на случай линейных моделей ускоряет сходимость метода. При этом уменьшается не только время выполнения отдельной итерации, но и общее их количество.

Литература

1. Sohl-Dickstein, J.; Poole, B.; Ganguli, S. (2014). "Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods". International Conference on Machine Learning (ICML).