Loss function optimization in the click prediction models

Maxim Hristolubov khristolyubov.me@phystech.edu
Daniil Merkulov daniil.merkulov@skoltech.ru

Project Proposal

Решается задача оптимизации логистической функции потерь с квадратичной регуляризацией, возникающей в задаче предсказания кликов пользователей. Особенностью задачи является факт наличия у данных двух типов признаков: плотных (сотни) и разряженных (десятки миллионов), для каждой группы свой коэффициент регуляризации. Предполагается использовать ускоренные адаптивные методы стохастического градиентного спуска.

1 Problem

Необходимо решить задачу минимазации:

$$F(w) = \frac{1}{m}\sum_{k=1}^{m} f_k(\boldsymbol{w}, (\boldsymbol{x}_k, y_k)) + g(w),$$

$$f_k(\boldsymbol{w}, (\boldsymbol{x}_k, y_k)) = ln(1 + \exp(-y_k \boldsymbol{w}^T \boldsymbol{x}_k)),$$

$$g(w) = \lambda_1 \sum_{i=1}^{n_1} w_i^2 + \lambda_2 \sum_{i=n_1+1}^{n_1+n_2} w_i^2,$$

where $\boldsymbol{y} \in \{1, -1\}^m$, $\boldsymbol{x}_k, \boldsymbol{w} \in R^n$, $\forall 1 \leq i \leq n_1 : x_{ki} \neq 0$ for almost all $k \leq m$, and $\forall n_1 + 1 \leq i \leq n_1 + n_2 : x_{ki} = 0$ for many $k \leq m$.

В связи с тем, что целевая функция представлена большим количеством слагаемых целесообразно начать решение со стохастического градиентного спуска, то есть брать градиент не от всех m слагаемых, а только от небольшой выборки. В связи с большой размерностью $\boldsymbol{w}$ имеет смысл применить методы покомпонентных (блочных) градиентных спусков, что тоже является разновидностью стохастического спуска. Наличие регуляризации наталкивает на мысль использования проксимального метода. Конечно, для эффективной сходимости градиентный спуск должен быть ускоренным.

2 Outcomes

Предлагается экспериментально на модельных данных сравнить эффективность нескольких методов и выбрать наилучший. Сконструировать некоторую(ые) суперпозицию(ии) этих методов и проверить численным экспериментом эффективность суперпозиции(ий). Для этой суперпозиции сделать теоретические оценки на сходимость метода (попытаться перенести оценки базовых методов). Выходом проекта будет код реализации базовых методов и их суперпозиций, графики их сходимости, а так же теоретические оценки сходимости суперпозиции методов.

3 Литературный обзор

В работе [1] предложено несколько методов ускорения адаптивного стохастического градиентного спуска, которые будет импользоваться как базовые способы решения задачи. В соответствии с [1], при выборе

шага в обычном постоянном стохастическом градиентном методе выполняется по формуле

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \frac{1}{2L_k}\nabla^r f(\boldsymbol{x}^k, \boldsymbol{\xi}^{k+1}),$$

где $L_k = L$ — константа Липшица градиента, а r — размер банча (кол-во f_k, которые берутся для вычисления стохастического градиента). Этот метод сделается адаптивным, если подбирать на каждой итерации более точным способом L_k. Его так же можно ускорить использовав стандартную идею ускорения: использовать усреднение по градиентам за несколько последних шагов и смотреть на градиент не в текущей точке, а на перед. Конкретный метод ускорения берется из [2].

В [3] рассматривается способ решения задачи минимизации эмпирического риска ускоренным градиентным методом, посредством параллельного вычисления градиента. В [4] доказывается теорема о сходимости бесконечного класса стохастических методов, определяемыми правилами выбора данных, используемых для формирования банчей, но не для ускоренных стохастических спусков. Общий подход, позволяющий ускорять почти произвольные неускоренные детерминированные и рандомизированные алгоритмы для гладких выпуклых задачах, предлагается в [5] (на основе Каталиста). В [6] предлагаются супербыстрый метод второго порядка (неточный метод третьего порядка, использующие только производные второго порядка), в том числе рассматривается задача, близкая к нашей. Никакие методы второго порядка на прямую применить в нашей задаче не получиться в силу большой размерности, но, возможно, можно как-то адаптировать при использовании блочных стохастических методов. В [7] [8] [9] рассматриваются покомпонентные (блочные) стохастические спуски.

4 Метрики качества

В качестве метрик эффективности работы методов стохастического спуска будет использоваться:

1) Минимальное значение функции $f(x_{end})$, до которого сошелся метод, пока скорость уменьшения значения функции не стала меньше некоторого заданного малого предела.

2) Минимальное значение отклонения $||x_{end} - x^*||_2$, до которого сошелся метод, пока скорость уменьшения значения функции не стал меньше некоторого заданного малого предела.

3) Скорость сходимости: $K = \sum\limits_{p=1}^{m} T_i$, где T_p — кол-во вызовов оракула для вычисления стохастического градиента (градиента функции f_k) на p компонентах (если используется метод покоординатного спуска) до достижения нижнего предела скорости сходимости.

4) Теоретические оценки скорости сходимости метода в условиях поставленной задачи.

5 Примерный план

- Реализовать адаптивный ускоренный непокомпонентный стохастический спуск, вкупе с проксимальным методом к 4 апреля.
- Потом разобраться в оболочке Каталист, реализовать/разобраться с покомпонентными (блочными) спусками и рассмотреть другие методы ускорений спуска.
- И составить некоторую суперпозицию этих методов - программа минимум.
- Сделать оценки на сходимость метода-суперпозиции - если успею программу минимум.

References

[1] Aleksandr Ogaltsov Darina Dvinskikh Pavel Dvurechensky Alexander Gasnikov Vladimir G. Spokoiny. Adaptive gradient descent for convex and non-convex stochastic optimization. 2020.

[2] Alexander Tyurin Alexander Ogaltsov. Heuristic adaptive fast gradient method in stochastic optimization tasks. 2019.

[3] Hadrien Hendrikx Lin Xiao Sebastien Bubeck Francis Bach Laurent Massoulie. Statistically preconditioned accelerated gradient method for distributed optimization. 2020.

[4] Robert M. Gower Nicolas Loizou Xun Qian Alibek Sailanbayev Egor Shulgin Peter Richtarik. Sgd: General analysis and improved rates. 2019.

[5] Anastasiya Ivanova Dmitry Pasechnyuk Dmitry Grishchenko Egor Shulgin Alexander Gasnikov. Adaptive catalyst for smooth convex optimization. 2020.

[6] Alexander Gasnikov Dmitry Kamzolov. Near-optimal hyperfast second-order method for convex optimization and its sliding. 2020.

[7] Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. 2010.

[8] Martin Takáč Peter Richtárik. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. 2011.

[9] Tong Zhang Lin Xiao. A proximal stochastic gradient method with progressive variance reduction. 2014.