Разработка метода стохастической оптимизации для задач машинного обучения с большими данными

Родоманов А. О.

научный руководитель: Ветров Д. П.

17 апреля 2015 г.

Оптимизация в машинном обучении

- Многие задачи сводятся к минимизации эмпирического риска:

$$F(\mathbf{w}) := \sum_{i=1}^{N} F_i(\mathbf{w}) \to \min_{\mathbf{w} \in \mathbb{R}^D}$$

- Например, логистическая регрессия с ℓ_2-регуляризатором:

$$F_i(\mathbf{w}) := \ln(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)) + \frac{\lambda}{2N} \|\mathbf{w}\|_2^2$$

- Считаем, что число объектов N очень большое.
- Рассматриваем методы, где сложность итерации не зависит от N.

Стохастические методы оптимизации

- Рассматриваем задачу минимизации $F(\mathbf{w}) := \sum_{i=1}^{N} F_i(\mathbf{w})$.
- Один из популярных методов — метод стохастического градиента:
  - Выбрать случайный номер $i \in \{1, 2, \ldots, N\}$ и вычислить

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \nabla \mathbf{F}_i(\mathbf{w}_k)$$

  - Сублинейная скорость сходимости: $O(1/k)$
  - Необходимость тонкого выбора параметров (например, α_k)
- Более эффективным методом является метод SAG:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \sum_{i=1}^{N} \mathbf{g}_i^k$$

  - Каждый раз меняется одна случайно выбранная компонента $\mathbf{g}_i$.
  - Память: $\mathbf{g}_i^k := \nabla \mathbf{F}_i(\mathbf{v}_i^k)$, где $\mathbf{v}_i^k$ — последняя точка, где вычислялась F_i.
  - Линейная скорость сходимости: $O(\rho^k)$, где $\rho \in (0, 1)$
  - Автоматическая процедура выбора параметров
- Цель работы: разработать метод, имеющий
  - суперлинейную скорость сходимости
  - не требующий ручной настройки параметров

Предлагаемый метод SO2

- Минимизируемая функция: $F(\mathbf{w}) := \sum_{i=1}^{N} F_i(\mathbf{w})$.
- Квадратичная модель для F_i с центром в точке $\mathbf{v}_i^k$:

$$Q_i^k(\mathbf{w}) := F_i(\mathbf{v}_i^k) + \nabla \mathbf{F}_i(\mathbf{v}_i^k)^\top (\mathbf{w} - \mathbf{v}_i^k) + \frac{1}{2}(\mathbf{w} - \mathbf{v}_i^k)^\top \nabla^2 \mathbf{F}_i(\mathbf{v}_i^k)(\mathbf{w} - \mathbf{v}_i^k)$$

- Квадратичная модель полной функции F: $Q^k(\mathbf{w}) := \sum_{i=1}^{N} Q_i^k(\mathbf{w})$.
- Гессиан полной модели Q^k:

$$\mathbf{H}_k = \sum_{i=1}^{N} \nabla^2 \mathbf{F}_i(\mathbf{v}_i^k)$$

- Итерация метода:
  - Шаг Ньютона для полной модели Q^k:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \mathbf{H}_k^{-1} \nabla \mathbf{Q}^k(\mathbf{w}_k),$$

  - Обновление одной компоненты модели: $\mathbf{v}_i^{k+1} := \mathbf{w}_{k+1}$.
- Параметр α_k задает длину шага и обычно равен единице.

Эффективное обновление для линейных моделей

- Матрицу $\mathbf{H}_k$ можно хранить и обновлять в итерациях:
$$\mathbf{H}_{k+1} = \mathbf{H}_k + \left(\nabla^2 \mathbf{F}_i(\mathbf{v}_i^{k+1}) - \nabla^2 \mathbf{F}_i(\mathbf{v}_i^k)\right)$$
- Чтобы не вычислять заново $\nabla^2 \mathbf{F}_i(\mathbf{v}_i^k)$, ее нужно хранить.
- Но тогда требуется слишком много памяти: $O(ND^2)$.
- Для линейных моделей $F_i(\mathbf{w}) := \phi_i(\mathbf{x}_i^\top \mathbf{w})$.
- Гессиан $\nabla^2 \mathbf{F}_i(\mathbf{v}_i^k) = \phi_i''(\mathbf{x}_i^\top \mathbf{v}_i^k)\mathbf{x}_i\mathbf{x}_i^\top$: одноранговая матрица.
- Объекты обучающей выборки $\mathbf{x}_i$ уже и так хранятся в памяти.
- Значит, $\nabla^2 \mathbf{F}_i(\mathbf{v}_i^k)$ можно хранить неявно: храним $\sigma_i^k := \phi_i''(\mathbf{x}_i^\top \mathbf{v}_i^k)$.
- Суммарный объем памяти: $O(N + D^2)$.
- Аналогично можно хранить и обновлять градиент $\nabla \mathbf{Q}^k(\mathbf{w}_k)$.
- Не решаем систему на каждой итерации: обновляем $\mathbf{B}_k := \mathbf{H}_k^{-1}$.
- Итоговая сложность обновления модели за итерацию: $O(D^2)$.

Теорема о суперлинейной сходимости

Теорема (локальная скорость сходимости)

Пусть

- *функции F_i являются дважды непрерывно дифференцируемыми*
- *гессианы $\nabla^2 \mathbf{F}_i$ удовлетворяют условию Липшица:*
$$\left\|\nabla^2 \mathbf{F}_i(\mathbf{w}) - \nabla^2 \mathbf{F}_i(\mathbf{u})\right\| \leq B \left\|\mathbf{w} - \mathbf{u}\right\|, \qquad \forall\, \mathbf{w}, \mathbf{u} \in \mathbb{R}^D$$
- *Точка $\mathbf{w}_*$ есть лок. минимум функции F, причем $\nabla^2 \mathbf{F}(\mathbf{w}_*) \succ \mathbf{0}$*
- *Начальная точка метода $\mathbf{w}_0$ находится достаточно близко к $\mathbf{w}_*$:*
$$\left\|\mathbf{w}_0 - \mathbf{w}_*\right\| \leq (N^2 RB)^{-1}, \qquad R := \left\|\left(\nabla^2 \mathbf{F}(\mathbf{w}_*)\right)^{-1}\right\|$$

Тогда $\{\mathbf{w}_k\}_{k=0}^{\infty}$, построенная методом SO2, сходится к $\mathbf{w}_*$ суперлинейно:
$$\lim_{k\to\infty} \frac{\left\|\mathbf{w}_{k+1} - \mathbf{w}_*\right\|}{\left\|\mathbf{w}_k - \mathbf{w}_*\right\|} = 0$$
Более того, посл. каждых N-х точек сходится к точке $\mathbf{w}_*$ квадратично:
$$\left\|\mathbf{w}_{k+N} - \mathbf{w}_*\right\| \leq (N^2 RB) \left\|\mathbf{w}_k - \mathbf{w}_*\right\|^2$$
для всех k, начиная с некоторого номера.

Набросок доказательства

- Пусть $r_k := \|\mathbf{w}_k - \mathbf{w}_*\|$.
- Справедлива следующая оценка:
$$r_k \leq (NRB)(r_{k-1}^2 + r_{k-2}^2 + \cdots + r_{k-N}^2)$$
- Посл. $\{r_k\}$ является монотонно убывающей (с нек. номера).
- Отсюда N-шаговая квадратичная скорость сходимости:
$$r_k \leq (N^2 RB) r_{k-N}^2$$
- Далее строится мажорирующая посл. $\{a_k\}$ для отношения r_{k+1}/r_k:
$$\frac{r_{k+1}}{r_k} \leq a_k,$$
где $a_k \to 0$ при $k \to \infty$.
- Из этой оценки следует суперлинейная скорость сходимости:
$$\frac{r_{k+1}}{r_k} \xrightarrow[k\to\infty]{} 0.$$

Сложность итерации и требуемая память

Минимизируемая функция: $F(\mathbf{w}) := \sum_{i=1}^{N} \phi_i(\mathbf{x}_i^\top \mathbf{w})$.

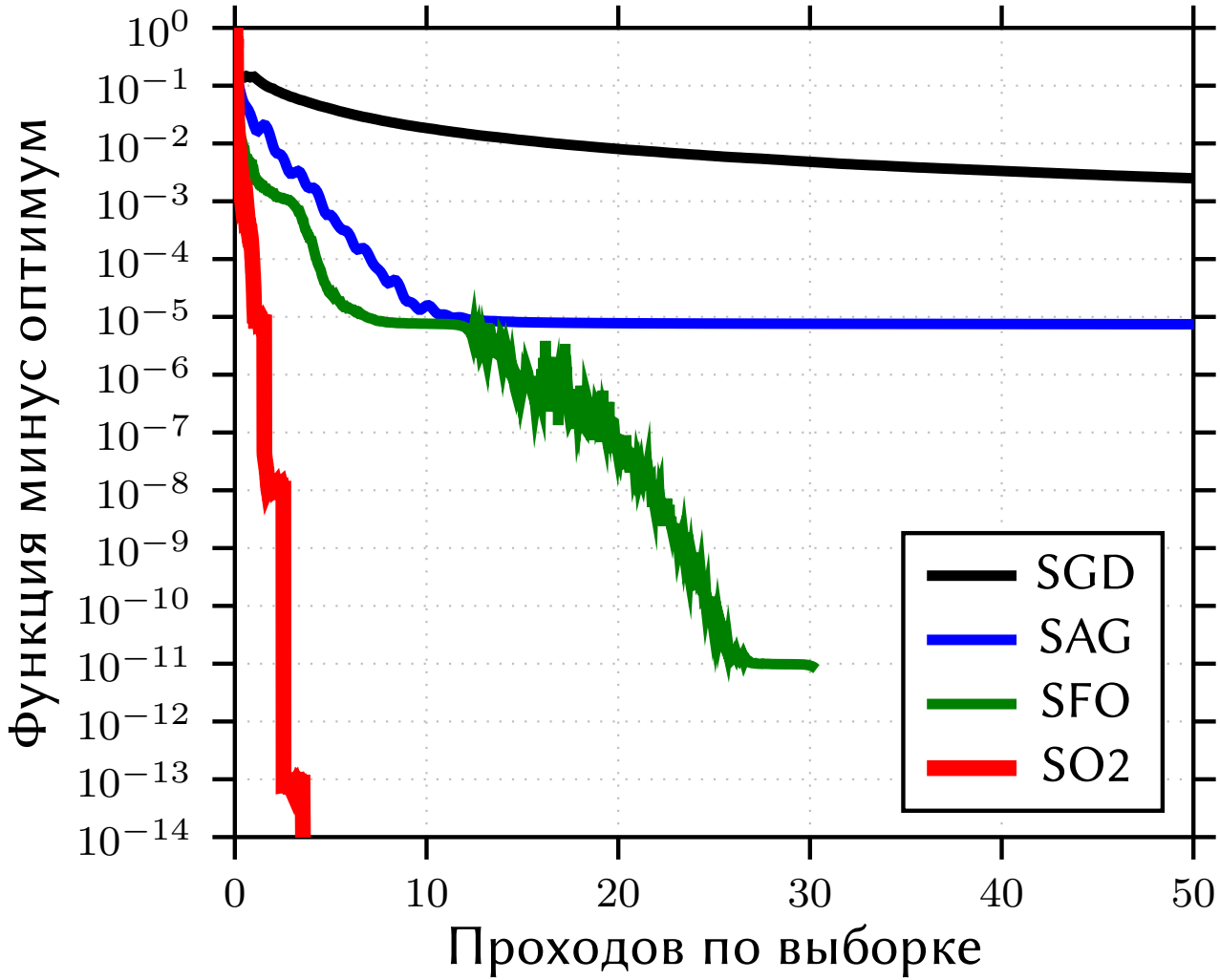
Метод	**Сл. итерации**	**Память**	**Ск. сходимости**	
			По итерац.	**По эпохам**
SGD	$O(C+D)$	$O(D)$	Сублин.	Сублин.
SAG	$O(C+D)$	$O(N+D)$	Линейная	Линейная
SO2	$O(C+D^2)$	$O(N+D^2)$	Суперлин.	Квадратич.

Обозначения:

- N: число объектов;
- D: число признаков;
- C: стоимость вычисления функции ϕ_i;
- По эпохам означает каждую N-ую итерацию.

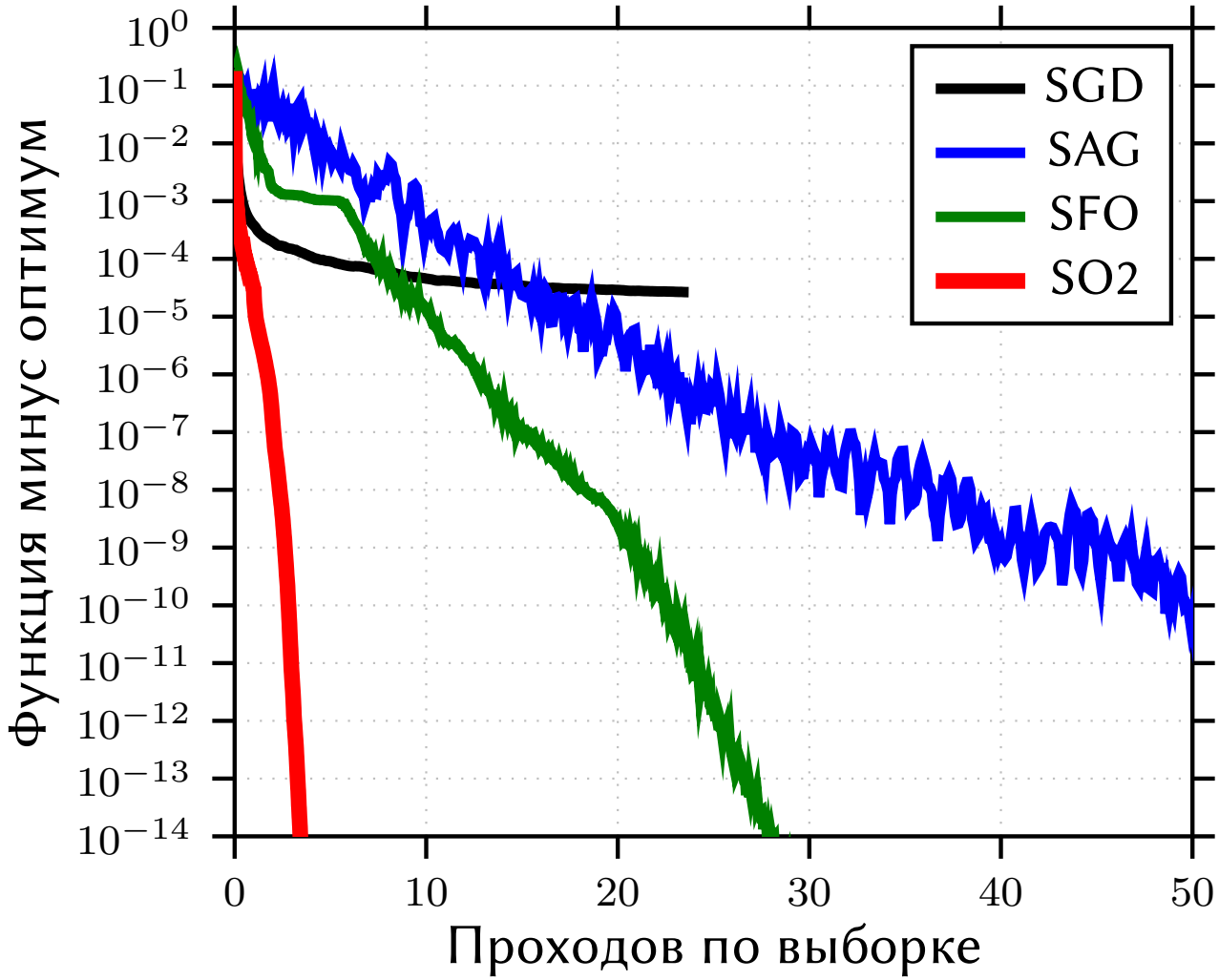
Эксперименты #1

- Оптимизируется логистическая регрессия с ℓ_2-регуляризатором.
- Набор данных quantum ($N = 50\,000$, $D = 78$):



Эксперименты #2

- Оптимизируется логистическая регрессия с ℓ_2-регуляризатором.
- Набор данных covtype ($N = 581\,012$, $D = 54$):



Положения, выносимые на защиту

На защиту выносится:

- Метод оптимизации SO2.
- Теорема о локальной скорости сходимости метода SO2.
- Экспериментальное сравнение метода SO2 с другими методами.