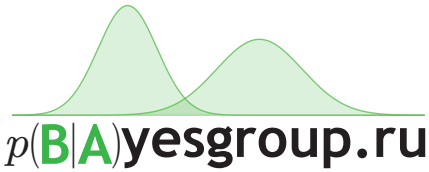
Инкрементальный метод Ньютона для больших сумм функций

Родоманов А.О.

Высшая школа экономики



Группа байесовских методов

29 октября 2016

Семинар «Стохастический анализ в задачах», НМУ, Москва

Введение

Рассматриваемая задача:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}.$$

Пример (Минимизация эмпирического риска):

- Имеются наблюдения a_i и их метки β_i.
- Цель: найти оптимальные параметры x^* параметрической модели.
- Линейная регрессия ($a_i \in \mathbb{R}^d,\ \beta_i \in \mathbb{R}$):

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} (\langle a_i, x \rangle - \beta_i)^2.$$

- Логистическая регрессия ($a_i \in \mathbb{R}^d,\ \beta_i \in \{-1, 1\}$):

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-\beta_i \langle a_i, x \rangle)).$$

Мотивация

Рассматриваем методы, которые на каждой итерации вычисляют лишь одну функцию f_i (вместо всей суммы):

- Методы стохастической оптимизации для $\min_x \mathbb{E}_\xi f(x;\xi)$:
  - **Примеры:** SGD [Robbins-Monro, 1951], oLBFGS [Schraudolph et al., 2007], AdaGrad [Duchi et al., 2011], SQN [Byrd et al., 2014], Adam [Kingma, 2014] и др.
  - **Итерация:** $x_{k+1} = x_k - \alpha_k B_k \nabla f_{i_k}(x_k)$.
  - **Скорость сходимости:** сублинейная, обычно $O(1/k)$.
- Специальные градиентные методы для $\min_x \left\{ \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$:
  - **Примеры:** SAG [Le Roux et al., 2012], SVRG [Johnson & Zhang, 2013], FINITO [Defazio et al., 2014b], SAGA [Defazio et al., 2014a], MISO [Mairal, 2015] и др.
  - **Основная идея:** уменьшение дисперсии со временем.
  - **Скорость сходимости:** линейная, $O(c^k)$.

Цель: метод с суперлинейной сходимостью.

Метод SAG [Le Roux et al., 2012]

Задача: $\min\limits_{x \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum\limits_{i=1}^{n} f_i(x) \right\}$.

Stochastic Average Gradient (SAG):

Выбрать $i_k \in \{1, \ldots, n\}$ случайно (равномерно)

Обновить $y_k^i = \begin{cases} \nabla f_i(x_k) & \text{если } i = i_k \\ y_{k-1}^i & \text{иначе} \end{cases}$

$g^k = g^{k-1} + \frac{1}{n}(y_k^{i_k} - y_{k-1}^{i_k})$

$x_{k+1} = x_k - \gamma g^k$

Идея метода NIM

Задача: $\min\limits_{x\in\mathbb{R}^d}\left\{\frac{1}{n}\sum\limits_{i=1}^{n} f_i(x)\right\}$.

Идея:

- Для каждой f_i рассмотрим ее квадратичную модель:
$$m_k^i(x) := f_i(v_k^i)+\langle\nabla f_i(v_k^i), x-v_k^i\rangle+\frac{1}{2}\langle\nabla^2 f_i(v_k^i)(x-v_k^i), x-v_k^i\rangle.$$
- Тогда модель для f равна $m_k(x) := \frac{1}{n}\sum\limits_{i=1}^{n} m_k^i(x)$.
- Выбрать x_{k+1} как минимум модели:
$$x_{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^d} m_k(x).$$
- Обновить только одну точку v_k^i: выбрать $i_k \in \{1,\dots,n\}$ и изменить
$$v_k^i = \begin{cases} x_k & \text{если } i = i_k, \\ v_{k-1}^i & \text{иначе.} \end{cases}$$

Обновление модели в методе NIM

Модель целевой функции:

$$m_k(x) = \frac{1}{n}\sum_{i=1}^{n}[f_i(v_k^i)+\langle\nabla f_i(v_k^i), x-v_k^i\rangle+\frac{1}{2}\langle\nabla^2 f_i(v_k^i)(x-v_k^i), x-v_k^i\rangle]$$

Заметим: m_k является квадратичной функцией,

$$m_k(x) = \frac{1}{2}\langle H_k x, x\rangle + \langle g_k - u_k, x\rangle + \text{const},$$

и полностью определяется следующими тремя величинами:

$$H_k := \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i),\ g_k := \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(v_k^i),\ u_k := \frac{1}{n}\sum_{i=1}^{n}\nabla^2 f_i(v_k^i)v_k^i.$$

Поскольку всегда обновляется только одна компонента, то

$$H_k = H_{k-1} + \frac{1}{n}\left[\nabla^2 f_i(v_k^i) - \nabla^2 f_i(v_{k-1}^i)\right],$$

$$g_k = g_{k-1} + \frac{1}{n}\left[\nabla f_i(v_k^i) - \nabla f_i(v_{k-1}^i)\right],$$

$$u_k = u_{k-1} + \frac{1}{n}\left[\nabla^2 f_i(v_k^i)v_k^i - \nabla^2 f_i(v_{k-1}^i)v_{k-1}^i\right].$$

Итоговая схема метода NIM

Задача: $\min\limits_{x\in\mathbb{R}^d}\left\{\frac{1}{n}\sum\limits_{i=1}^{n} f_i(x)\right\}$.

Инкрементальный метод Ньютона (NIM):

Взять $i_k = k \bmod n + 1$

Обновить $v_k^i = \begin{cases} x_k & \text{если } i = i_k \\ v_{k-1}^i & \text{иначе} \end{cases}$

$$H_k = H_{k-1} + \frac{1}{n}\left[\nabla^2 f_i(v_k^i) - \nabla^2 f_i(v_{k-1}^i)\right],$$

$$g_k = g_{k-1} + \frac{1}{n}\left[\nabla f_i(v_k^i) - \nabla f_i(v_{k-1}^i)\right],$$

$$u_k = u_{k-1} + \frac{1}{n}\left[\nabla^2 f_i(v_k^i)v_k^i - \nabla^2 f_i(v_{k-1}^i)v_{k-1}^i\right]$$

Вычислить $x_{k+1} = H_k^{-1}(u_k - g_k)$.

Суперлинейная сходимость метода NIM

Теорема: Пусть $\nabla^2 f_i$ удовлетворяют условию Липшица:

$$\|\nabla^2 f_i(x) - \nabla^2 f_i(y)\| \leq M\|x - y\|, \qquad \forall x, y \in \mathbb{R}^d.$$

Предположим, что x^* — невырожденная точка минимума:

$$\nabla^2 f(x^*) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 f_i(x^*) \succeq \mu I, \qquad \mu > 0,$$

и начальные точки $x_0, \ldots, x_{n-1}$ лежат достаточно близко к x^*:

$$\|x_i - x^*\| \leq \frac{\mu}{2M}.$$

Тогда последовательность $\{x_k\}$ сходится к x^* с R-суперлинейной скоростью, т.е. существует $\{z_k\}$, что

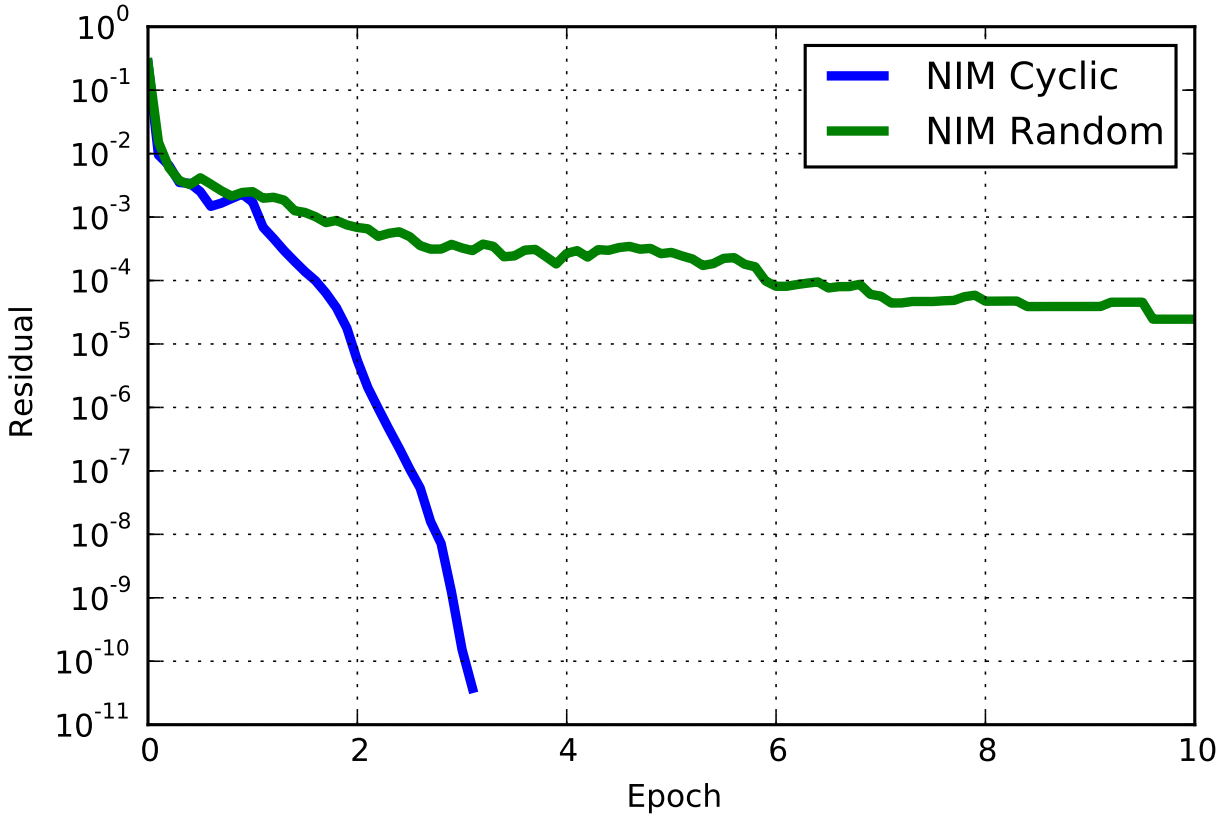
$$\|x_k - x^*\| \leq z_k, \qquad z_{k+1} \leq \left(1 - \frac{3}{4n}\right)^{2^{\lceil k/n \rceil - 1}} z_k.$$

Кроме этого, имеет место n-шаговая квадратичная сходимость:

$$z_{k+n} \leq \frac{M}{\mu} z_k^2.$$

Порядок выбора компонент

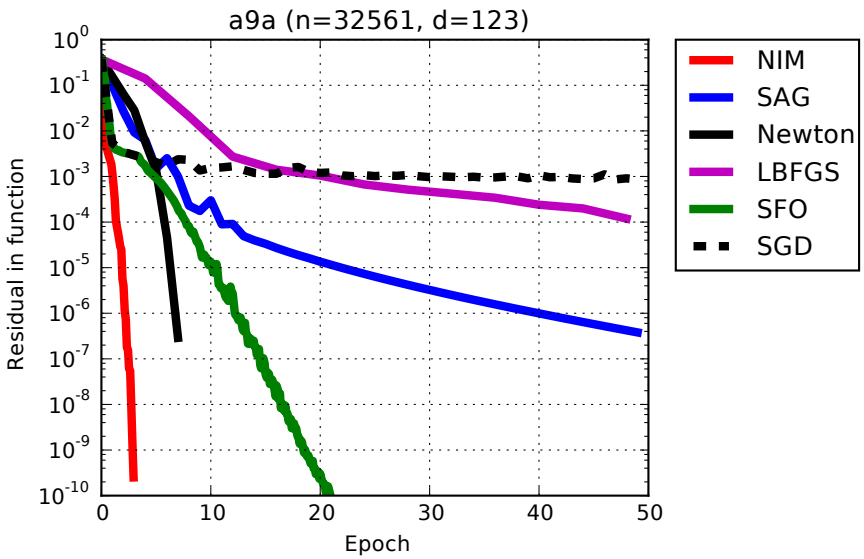
Сравнение циклического и случайного порядков:

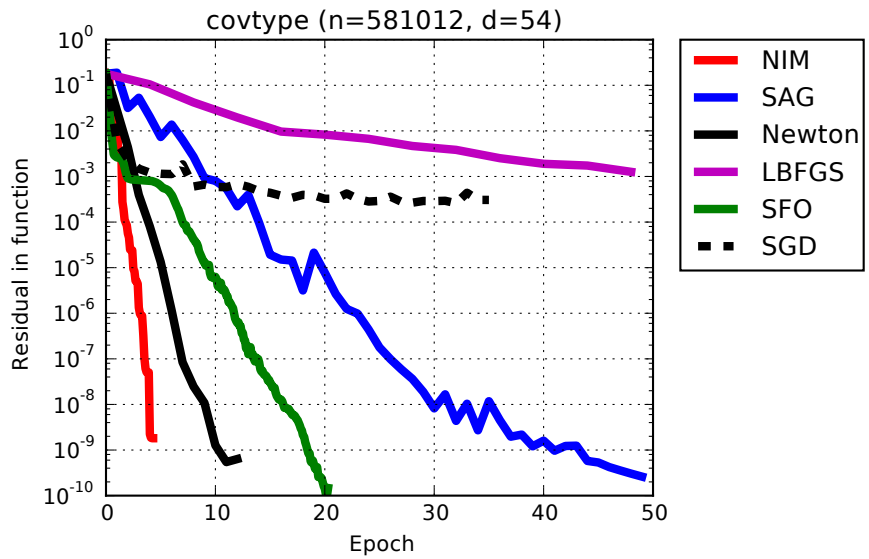


Экспериментальное сравнение – 1

Логистическая регрессия с $L2$-регуляризатором:

$$f(x) := \frac{1}{n}\sum_{i=1}^{n} \ln(1 + \exp(-\beta_i \langle a_i, x\rangle)) + \frac{\mu}{2}\|x\|^2.$$





Результаты экспериментов – 2

	a9a (n=32561, d=123)					*covtype (n=581012, d=54)*				
Res	NIM	SAG	Newton	LBFGS	SFO	NIM	SAG	Newton	LBFGS	SFO
10^{-1}	**.01s**	.01s	.31s	.05s	.03s	.19s	.33s	.84s	.54s	**.04s**
10^{-2}	**.02s**	.05s	.56s	.10s	.08s	.51s	.96s	1.78s	1.77s	**.25s**
10^{-3}	.12s	**.11s**	.73s	.18s	.57s	**.72s**	1.58s	2.39s	5.67s	1.02s
10^{-4}	**.15s**	.19s	.81s	.43s	.98s	**.86s**	2.45s	3.09s	10.73s	3.80s
10^{-5}	**.21s**	.36s	.90s	.76s	1.34s	**1.20s**	3.37s	3.99s	19.07s	5.23s
10^{-6}	**.24s**	.66s	.93s	1.11s	1.57s	**1.49s**	4.12s	4.57s	31.84s	6.81s
10^{-7}	**.28s**	1.04s	1.00s	1.45s	1.93s	**1.69s**	4.69s	5.13s	-	8.23s
10^{-8}	**.31s**	1.46s	1.04s	1.82s	2.18s	**1.92s**	5.90s	6.52s	-	9.86s
10^{-9}	**.32s**	1.90s	1.04s	2.26s	2.46s	**2.10s**	7.34s	7.64s	-	11.30s
10^{-10}	**.34s**	2.38s	1.04s	2.61s	2.81s	**2.12s**	9.97s	8.84s	-	12.44s

Результаты экспериментов – 3

	alpha (n=500000, d=500)				*mnist8m (n=8100000, d=784)*			
Res	NIM	SAG	Newton	LBFGS	NIM	SAG	Newton	LBFGS
10^{-1}	1.91s	**1.36s**	1.6m	4.01s	57.68s	**34.91s**	47.8m	1.1m
10^{-2}	13.37s	**6.72s**	2.6m	17.68s	**1.6m**	2.1m	1.4h	5.2m
10^{-3}	28.56s	**17.73s**	3.0m	37.70s	**3.2m**	3.9m	-	22.9m
10^{-4}	36.65s	**36.04s**	3.4m	58.35s	16.7m	**7.1m**	-	1.6h
10^{-5}	**46.66s**	1.0m	3.6m	1.4m	**26.7m**	1.0h	-	-
10^{-6}	**53.92s**	1.5m	4.0m	1.9m	**33.5m**	-	-	-
10^{-7}	**57.63s**	2.0m	4.0m	2.4m	**40.1m**	-	-	-
10^{-8}	**1.0m**	2.7m	4.1m	2.8m	**46.0m**	-	-	-
10^{-9}	**1.1m**	3.5m	4.3m	3.2m	**49.6m**	-	-	-
10^{-10}	**1.2m**	4.3m	4.7m	3.4m	**53.3m**	-	-	-