Стохастические квазиньютоновские методы оптимизации в контексте глубоких нейронных сетей

А. Жогов М. Сысак Б. Усеинов

Московский физико-технический институт, Москва, Россия

26 мая 2020 г.

Постановка задачи

- Нейронные сети являются передовыми методами во множестве задач машинного обучения.
- Квазиньютоновские методы широко используются в выпуклой оптимизации.
- В глубинном обучении стохастические модификации квазиньютоновских методов используются редко из-за их неустойчивости.

Постановка задачи

- $\mathcal{X} = \{\mathbf{x}_i, y_i\}_1^n$ — выборка, $f(\mathbf{x}_i, \theta)$ — нейронная сеть, функция потерь:

$$\mathcal{L}(\mathbf{p}, y) = -\log p_y$$

- Задача оптимизации:

$$\min_\theta F_{\mathcal{X}}(\theta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x_i, \theta), y_i)$$

- Минибатч $\mathcal{X}' \subset \mathcal{X}$, $s = |\mathcal{X}'| \ll |\mathcal{X}|$
- Вспомогательная задача

$$\min_\theta F_{\mathcal{X}'}(\theta) \triangleq \frac{1}{s} \sum_{j=1}^{s} \mathcal{L}(f(x_{i_j}, \theta), y_{i_j})$$

- На каждой итерации генерируется минибатч $\mathcal{X}'_t \subset \mathcal{X}$
- Шаг стохастического градиентного спуска (SGD):

$$\theta_{k+1} = \theta_k - \eta \nabla F_{\mathcal{X}'_t}(\theta_k)$$

- Шаг SGD-Momentum:

$$\theta_{k+1} = \theta_k - \eta \nabla F_{\mathcal{X}'_t}(\theta_k) + \beta(\theta_k - \theta_{k+1})$$

Методы

Adam

- На каждой итерации генерируется минибатч $\mathcal{X}'_t \subset \mathcal{X}$
- Вычисляется градиент $g_{t+1} = \nabla F_{\mathcal{X}'_t}(\theta_t)$ и обновляются скользящие средние

$$m_{t+1} = \beta_1 m_t + (1 - \beta_1) g_{t+1} \quad v_{t+1} = \beta_2 v_t + (1 - \beta_2) g^2_{t+1}$$

- Поправка на смещение

$$\hat{m}_{t+1} = m_{t+1} / (1 - \beta_1^{t+1}) \quad \hat{v}_{t+1} = v_{t+1} / (1 - \beta_2^{t+1})$$

- Обновление целевой переменной

$$\theta_{t+1} = \theta_t - \eta \cdot \hat{m}_{t+1} / (\sqrt{\hat{v}_{t+1}} + \varepsilon)$$

Методы

Multi-Batch LBFGS

- На каждой итерации генерируется минибатч $\mathcal{X}'_k \subset \mathcal{X}$
- Требование пересечения минибатчей:

$$O_k = \mathcal{X}'_{k+1} \cap \mathcal{X}'_k \quad |O_k| = l < s = |\mathcal{X}'_k|$$

- Квазиньютоновское уравнение:

$$y_{k+1} = \nabla F_{O_k}(\theta_{k+1}) - \nabla F_{O_k}(\theta_k) \quad s_{k+1} = \theta_{k+1} - \theta_k$$

$$s_{k+1} = H_{k+1} y_{k+1}$$

- Вычисление оценки гессиана H_{k+1} с помощью процедуры two_loop_recursion
- Обновление целевой переменной

$$\theta_{k+1} = \theta_k - \eta H_k \nabla F_{\mathcal{X}'_k}(\theta_k)$$

Методы
SGD-BB

- Эпоха — $\left\lceil \frac{|\mathcal{X}|}{|\mathcal{X}'|} \right\rceil \triangleq T$ итераций
- k — номер эпохи, t — номер итерации, минибатч $\mathcal{X}'_{k,t}$
- Обновление переменной и оценки градиента:

$$\theta_{k,0} = \theta_{k-1,T} \quad \theta_{k,t+1} = \theta_{k,t} - \eta_k \nabla F_{\mathcal{X}'_{k,t}}(\theta_{k,t})$$

$$g_{k,t+1} = (1-\beta) g_{k,t} + \beta \nabla F_{\mathcal{X}'_{k,t}}(\theta_{k,t})$$

- Обновление размера шага:

$$y_k = g_{k,T} - g_{k-1,T} \quad s_k = T^{-1}(\theta_{k,T} - \theta_{k-1,T})$$

$$\eta_{k+1} = \frac{\|s_k\|_2^2}{|s_k^T y_k|}$$

Архитектура нейронной сети в эксперименте с MNIST

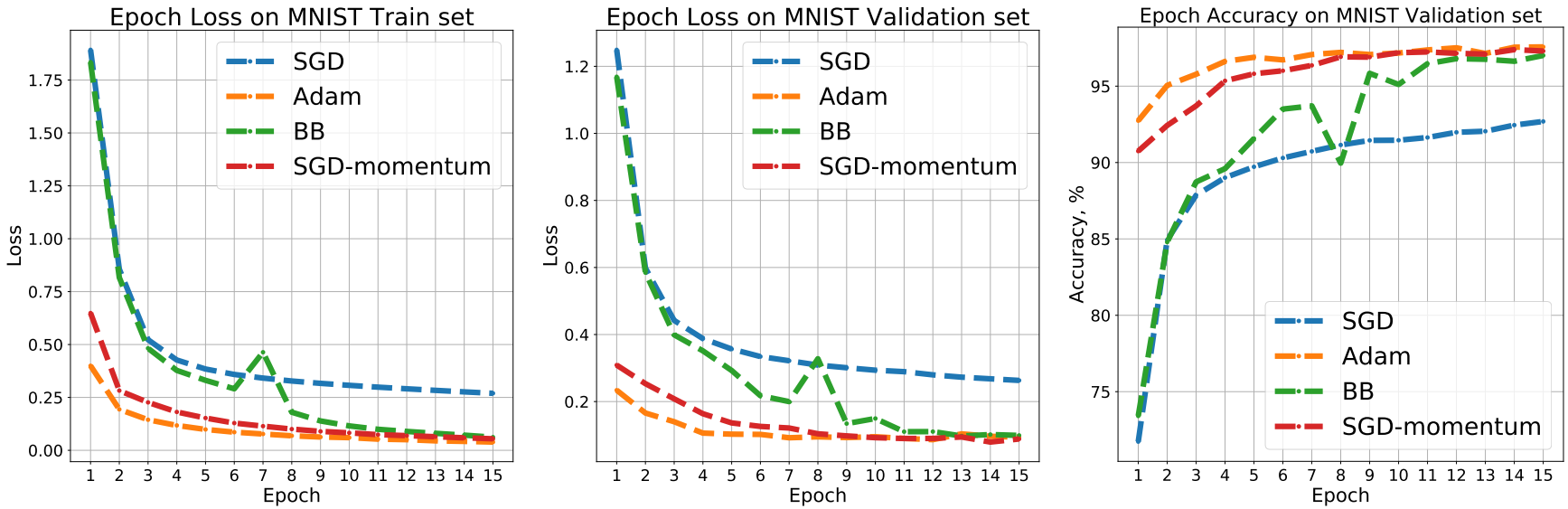
№	Слой
0: вход	28×28 изображение, 1×784
1:	Полносвязный, $(784, 128)$, ReLU
2:	Полносвязный, $(128, 64)$, ReLU
3: выход	Полносвязный, $(64, 10)$, LogSoftmax

Результаты

Эксперимент с MNIST

Время, затраченное на 15 эпох обучения

Метод	SGD	SGD-Momentum	Adam	SGD-BB
Время, мин	**2.81**	3.10	3.80	3.33



Графики сходимости методов

Архитектура нейронной сети в эксперименте с CIFAR-10

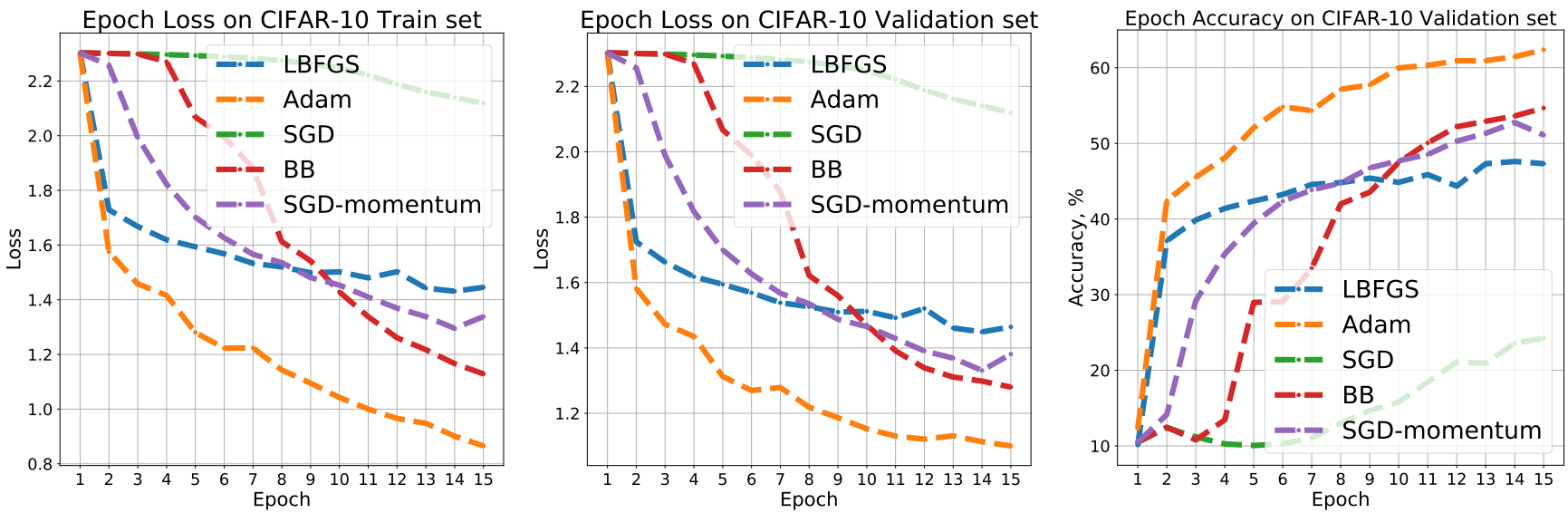
№	Слой
0: вход	32×32 изображение
1:	Сверточный, $(6,\ 5 \times 5,\ 1)$, ReLU
2:	Субдискретизирующий (макс.), $(2 \times 2, 1)$
3:	Сверточный, $(16,\ 5 \times 5,\ 1)$, ReLU
4:	Полносвязный, $(4096, 1000)$, ReLU
5: выход	Полносвязный, $(1000, 10)$, LogSoftmax

Результаты

Эксперимент с CIFAR-10

Время, затраченное на 15 эпох обучения

Метод	MB-LBFGS	Adam	SGD	SGD-Momentum	SGD-BB
Время, мин	6.37	5.35	5.19	5.40	**5.16**



Графики сходимости методов

Выводы

- Квазиньютоновские методы сопоставимы с Adam и SGD-Momentum по качеству решения.
- SGD требует больше итераций для сходимости.

- Все методы кроме MB-LBFGS затрачивают сопоставимое количество времени.
- MB-LBFGS работает на 20% дольше.