Задача 24: Выбор алгоритма оптимизации нейронной сети

Пилькевич Антон
группа Б05-811
МФТИ

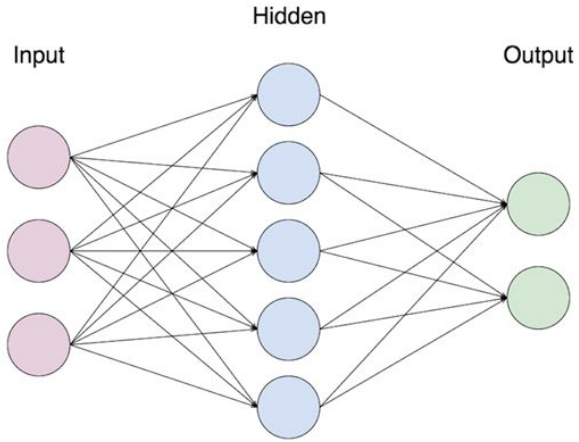
Цель:

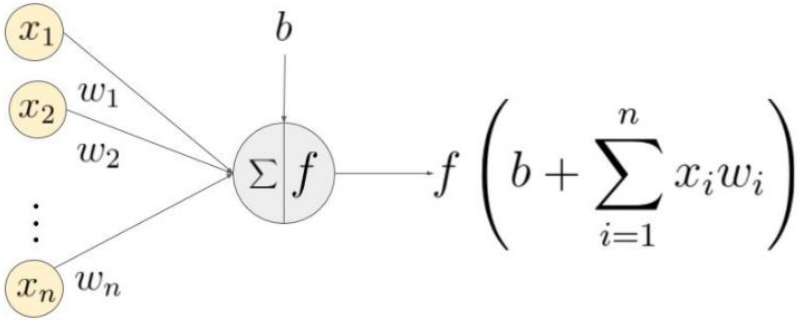
- Сравнить SGD, Nesterov Momentum, Adam.
- Сделать выводы о сценариях использования.

Критерии качества:

- Скорость сходимости (speed of convergence).
- Обобщение (generalization).

Общие сведения:





SGD (Stochastic gradient descent)

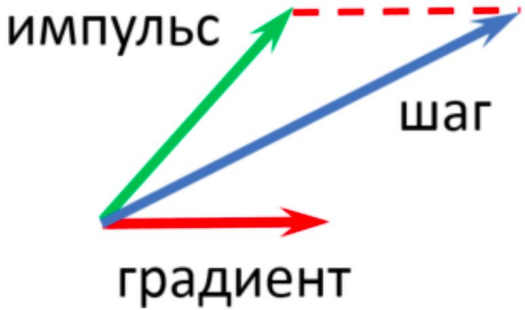
$$x_{t+1} = x_t - \alpha \nabla f(x_t)$$

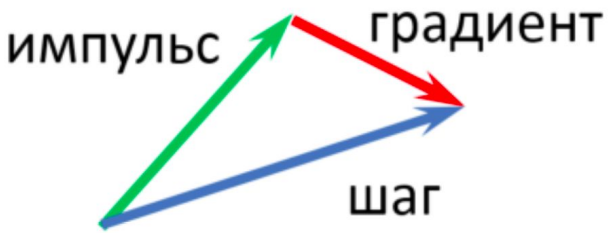
Добавим момент:

$$v_{t+1} = \rho v_t + \nabla f(x_t)$$
$$x_{t+1} = x_t - \alpha v_{t+1}$$

Nesterov Momentum

$$v_{t+1} = \rho v_t - \alpha \nabla f(x_t + \rho v_t)$$
$$x_{t+1} = x_t + v_{t+1}$$





Adagrad(adaptive grad)
RMSProp(root mean square propagation)

Добавили память о предыдущих шагах

$$\text{cache}_{t+1} = \text{cache}_t + (\nabla f(x_t))^2$$

$$x_{t+1} = x_t - \alpha \frac{\nabla f(x_t)}{\text{cache}_{t+1} + \varepsilon}$$

Добавили экспоненциальное скользящее среднее градиента

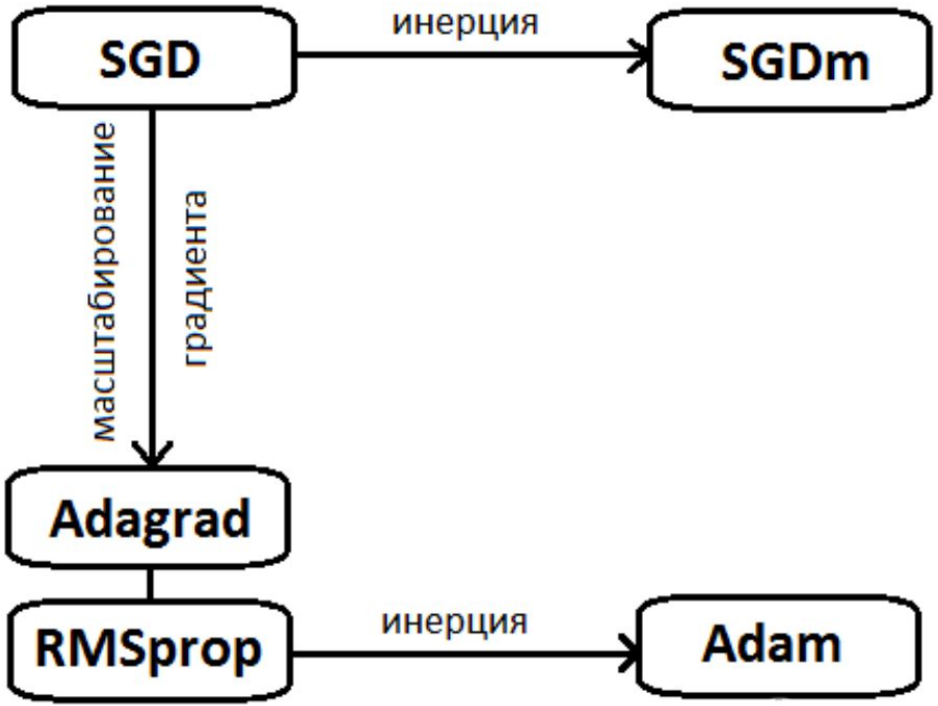
$$\text{cache}_{t+1} = \beta \text{cache}_t + (1-\beta)(\nabla f(x_t))^2$$

$$x_{t+1} = x_t - \alpha \frac{\nabla f(x_t)}{\text{cache}_{t+1} + \varepsilon}$$

Adam(adaptive momentum)

$$v_{t+1} = \gamma v_t + (1-\gamma)\nabla f(x_t)$$
$$\text{cache}_{t+1} = \beta \text{cache}_t + (1-\beta)(\nabla f(x_t))^2$$
$$x_{t+1} = x_t - \alpha \frac{v_{t+1}}{\text{cache}_{t+1} + \varepsilon}$$

Связь между методами

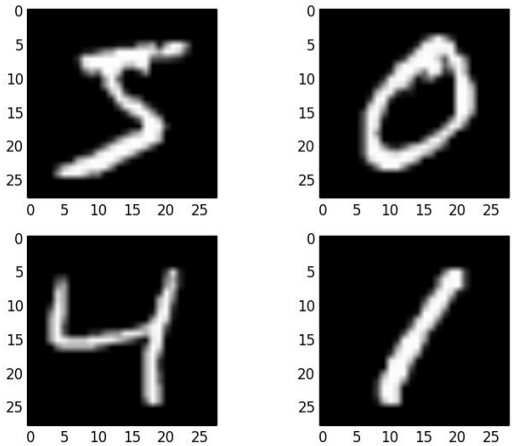


Используемая модель:

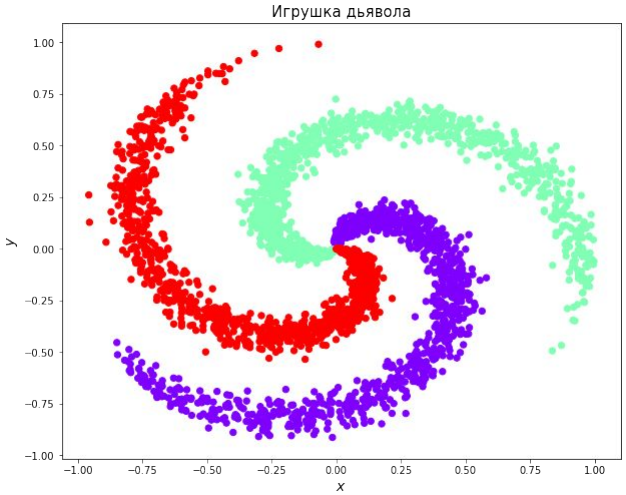
- Полносвязная сеть: состоит из 2-х скрытых слоёв по 100 нейронов.
- Рассмотрим модели с батч-нормализацией и без неё.
- В качестве функции активации используется ReLU.
- Функция потерь: кросс-энтропия(log loss).

Набор данных:

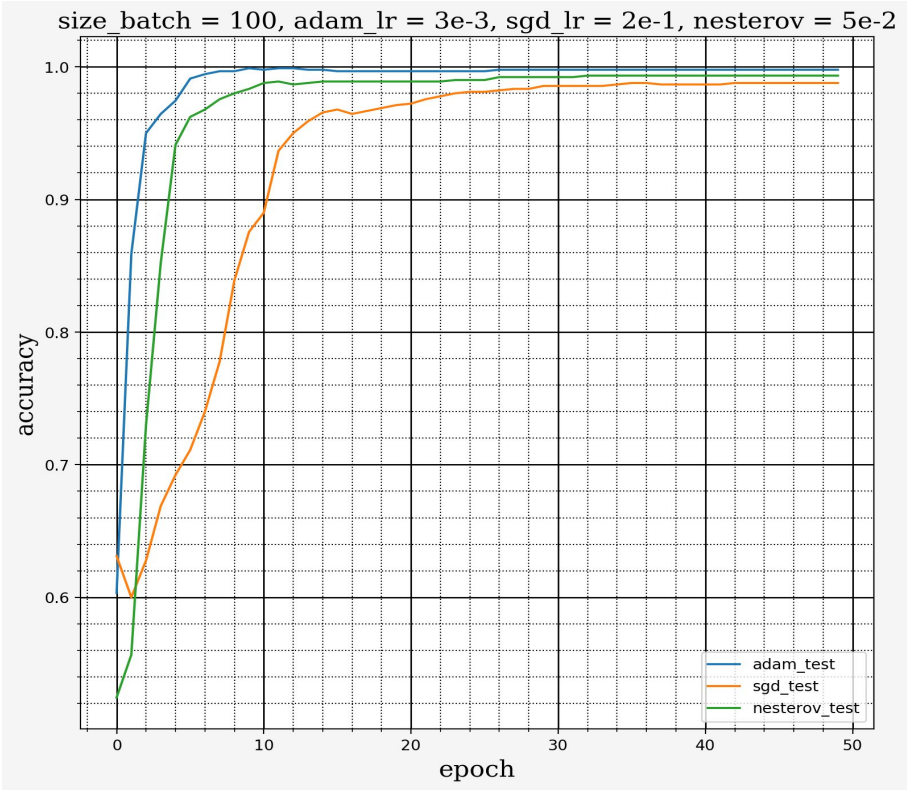
- MNIST.
- Тестовая выборка: 10000.
- Тренировочная: 60000.

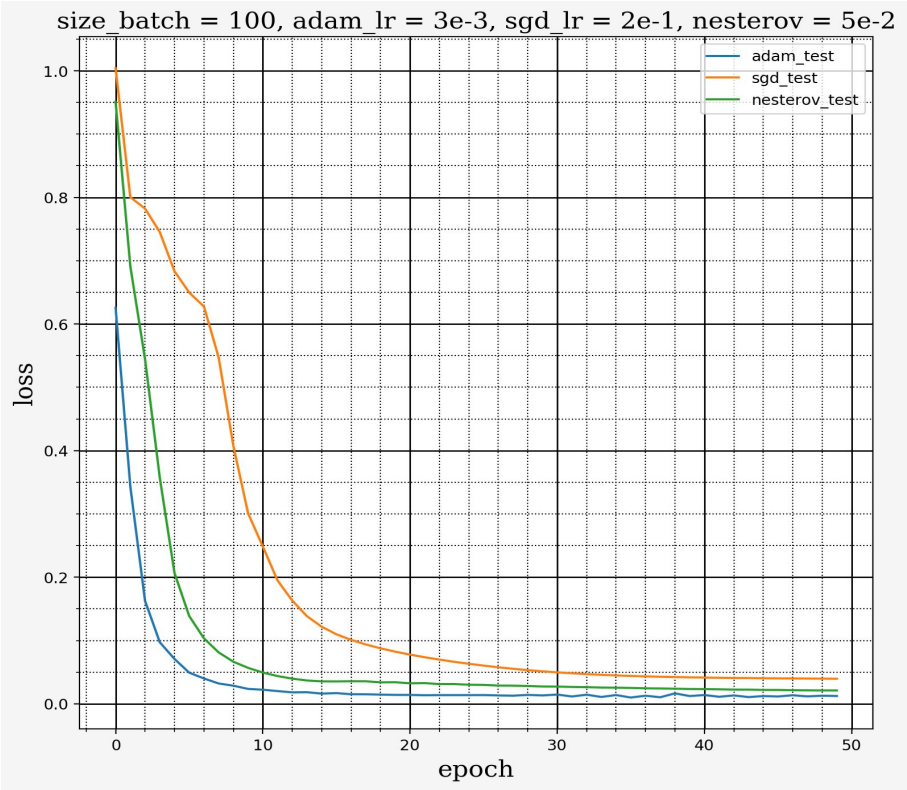


- Синтетическая выборка
- Тестовая выборка: 900
- Тренировочная: 2100

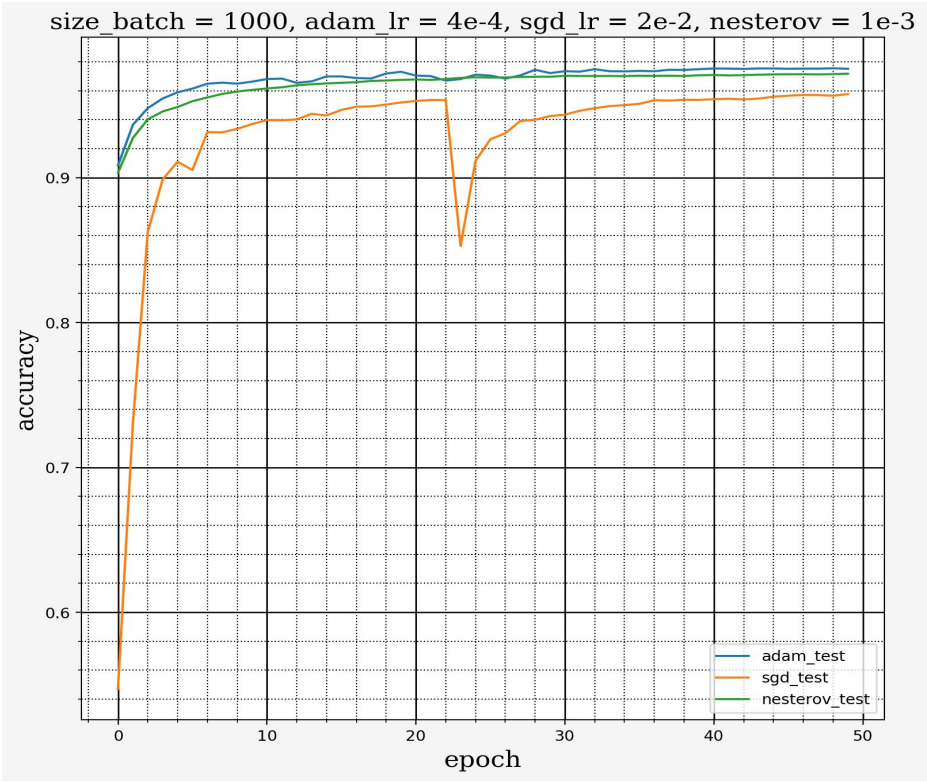


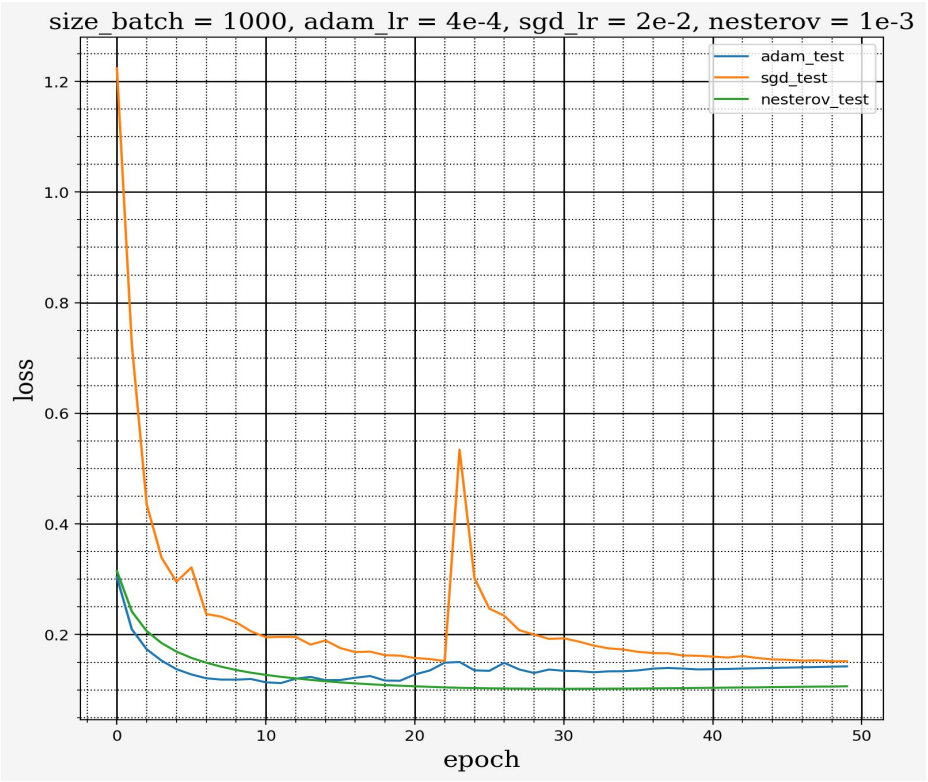
Синтетическая выборка без батчнормы



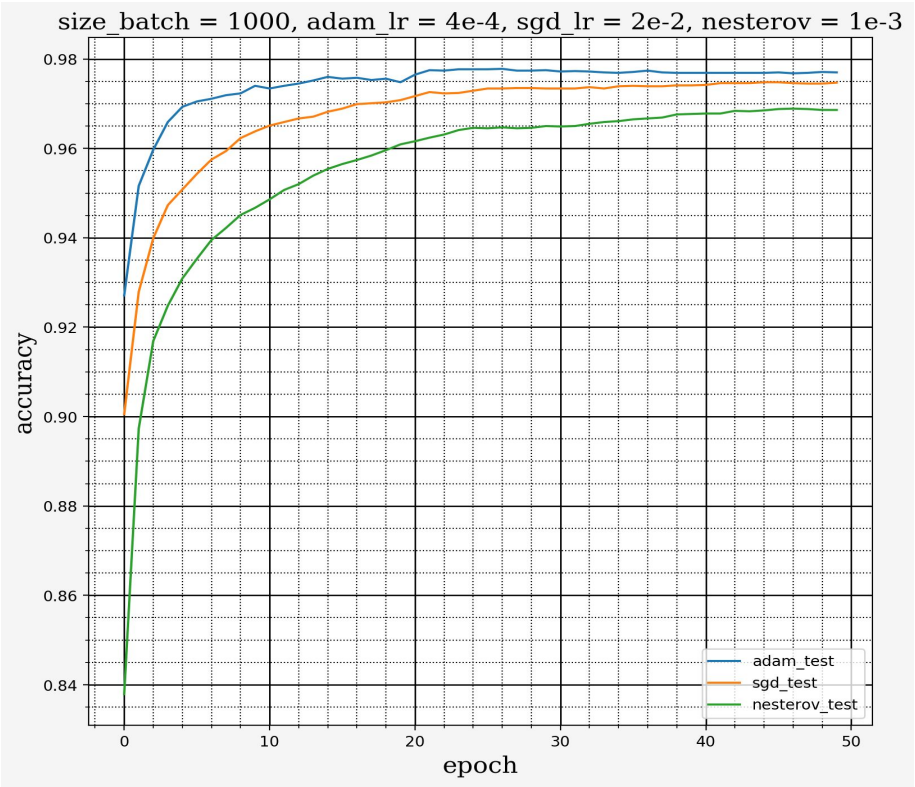


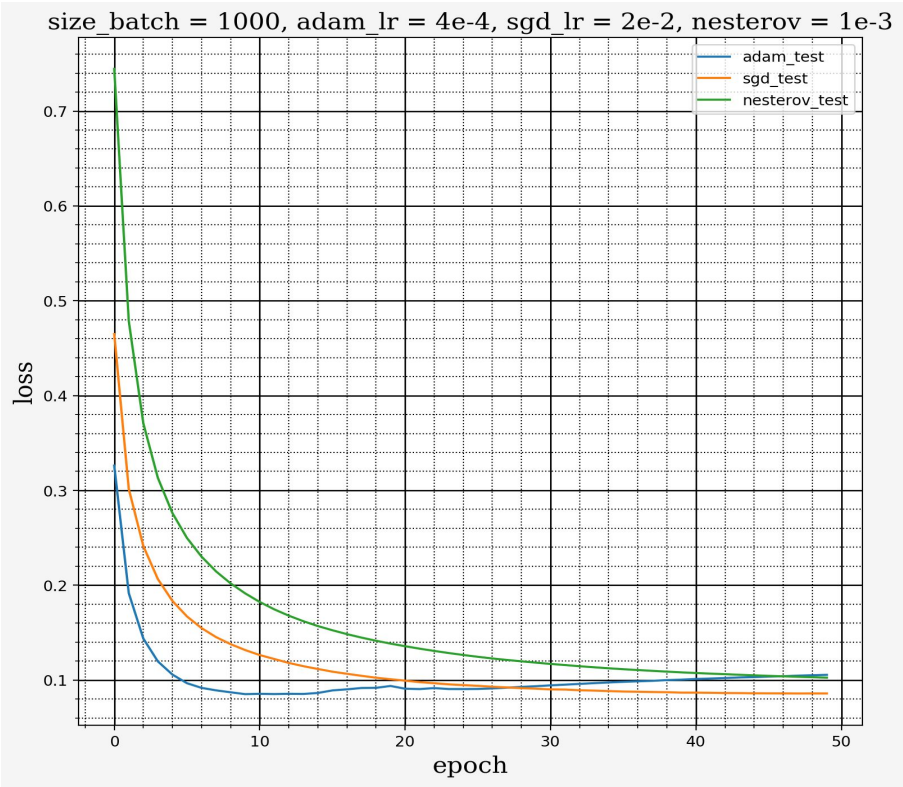
MNIST без батчнормы



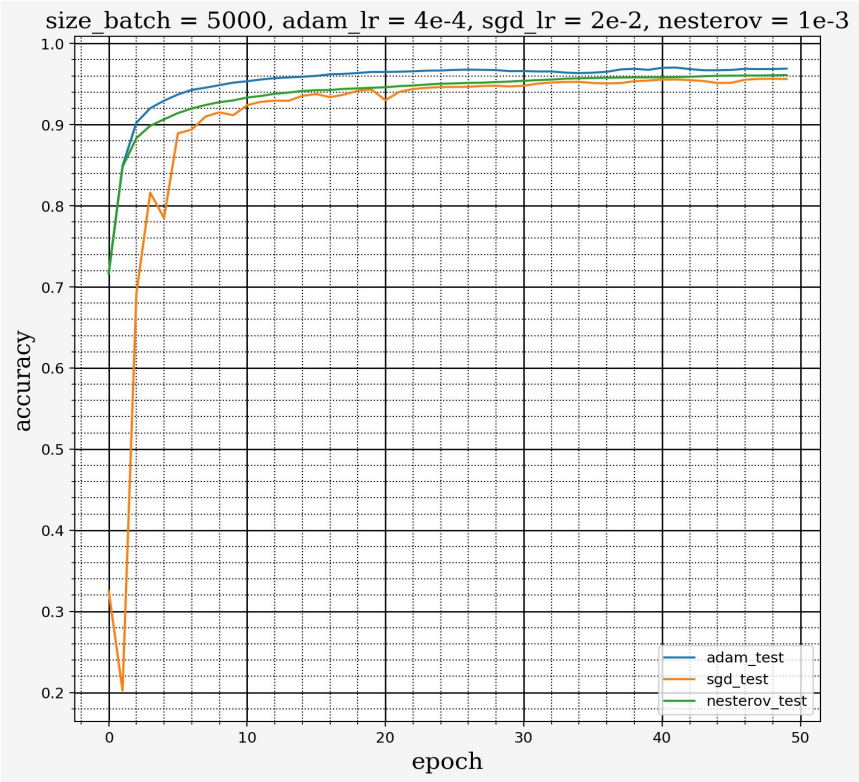


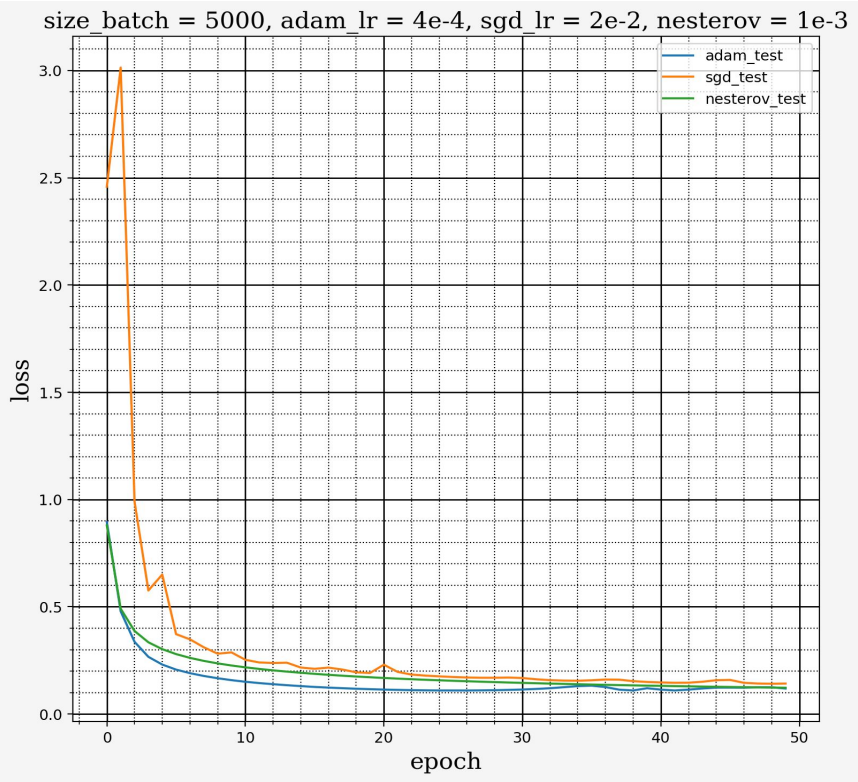
С батчнормой





Без батчнормы





Результаты научных статей и исследований

- ADAM в некоторых задачах не сходится к оптимуму.
- У адаптивных методов хуже показатель обобщения, чем у SGDm.
- На поздних этапах ADAM может сходится хуже, чем SGD.
- Для выпуклых функций использование импульса превосходят SGD.
- NAG для традиционно сложных задач превосходит SGD и SGDm.

Выводы:

- Для простых моделей оптимальнее использовать различные модификации SGD.
- В начале скорость сходимости Adam значительно выше.
- Есть задачи, где Adam будет уступать SGD и NAG.
- У SGD и NAG лучше показатель обобщения.
- Регуляризация даёт значительный прирост SGD.

Список литературы

- Лекция DLS https://drive.google.com/file/d/1_JmGiFfVv1frDQqSHNfCAMIwDUg4Y6hF/view
- Stanford lecture http://cs231n.stanford.edu/slides/2017/cs231n_2017_lecture7.pdf
- Дипломная работа Чабенко В. Д. http://www.machinelearning.ru/wiki/images/a/a0/2016_417_ChabanenkoVD.pdf
- Лекция Воронцова К. В. http://www.machinelearning.ru/wiki/images/3/38/Voron-ML-NeuralNets1-2018-slides.pdf
- Adam — latest trends in deep learning optimization. https://towardsdatascience.com/adam-latest-trends-in-deep-learning-optimization-6be9a291375c
- An overview of gradient descent optimization algorithms https://arxiv.org/pdf/1609.04747.pdf
- Adam: a method for stohastic optimization https://arxiv.org/pdf/1412.6980.pdf
- Improving Generalization Performance by Switching from Adam to SGD https://arxiv.org/pdf/1712.07628.pdf
- On the importance of initialization and momentum in deep learning http://proceedings.mlr.press/v28/sutskever13.pdf