Стохастическая оптимизация

Натуральный градиент

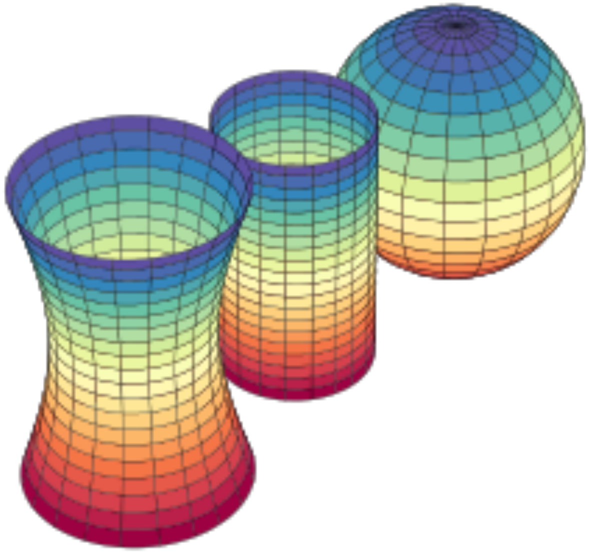
- Метрика в евклидовом пространстве:

$$\| d\boldsymbol{w} \|^2 = \sum_{i=1}^{n} (dw_i)^2,$$

- Метрика в римановом пространстве:

$$\| d\boldsymbol{w} \|^2 = \sum_{i,j} g_{ij}(\boldsymbol{w}) dw_i dw_j.$$

- квадратичная форма с матрицей G



Натуральный градиент

Формула расстояния:

$$dist(\theta, \theta + \delta\theta) = \|\delta\theta\| = \sqrt{\delta\theta^T G(\theta)\delta\theta}, \qquad G \in S^n_{++}$$

Оптимизационная задача (минимизируем функцию L по шагу $\delta\theta$ фиксированной длины):

$$\begin{cases} L(\theta + \delta\theta) \approx L(\theta) + \nabla L(\theta)^T\, \delta\theta \ \to \min\limits_{\delta\theta} \\ \|\delta\theta\|^2 = \varepsilon = const \end{cases}$$

Решение:

$$\delta\theta = -cG^{-1}\nabla L(\theta), \qquad c = const$$

Получили направление наискорейшего спуска: $-G^{-1}\nabla L(\theta)$,
$\widetilde{\nabla} L(\theta) = G^{-1}\nabla L(\theta)$ – натуральный градиент.

Натуральный градиент

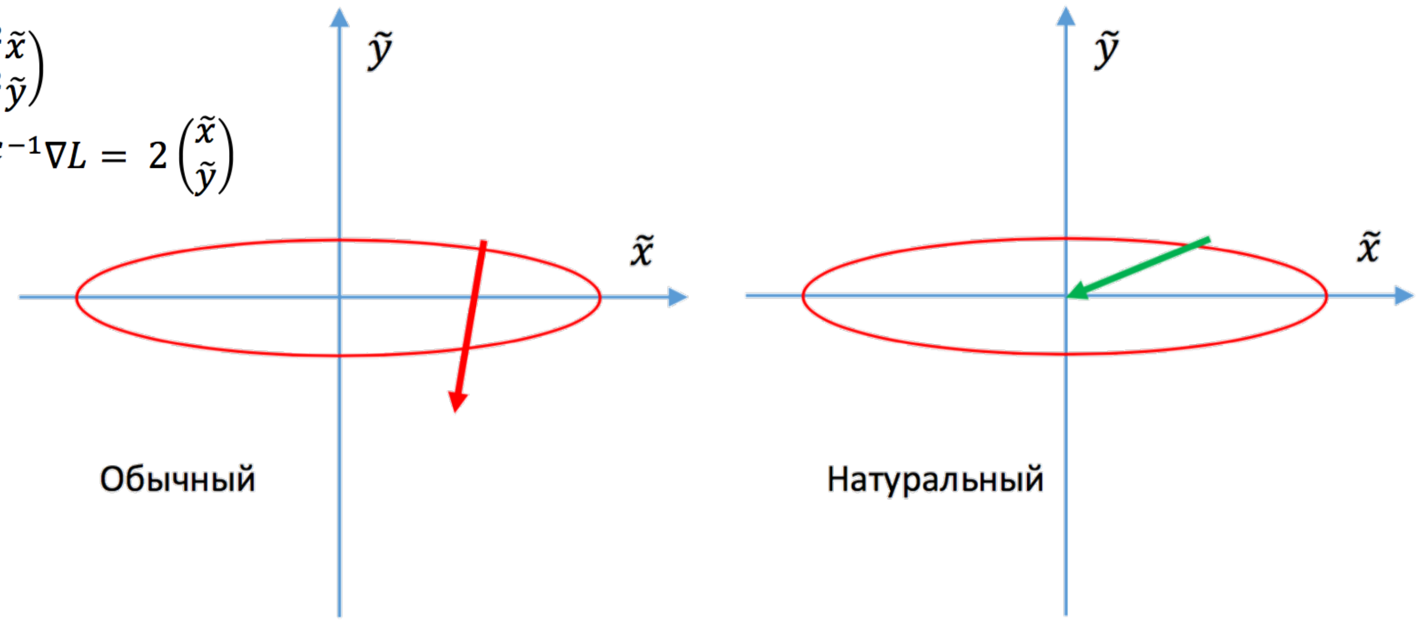
Простой пример. Плохая обусловленность:

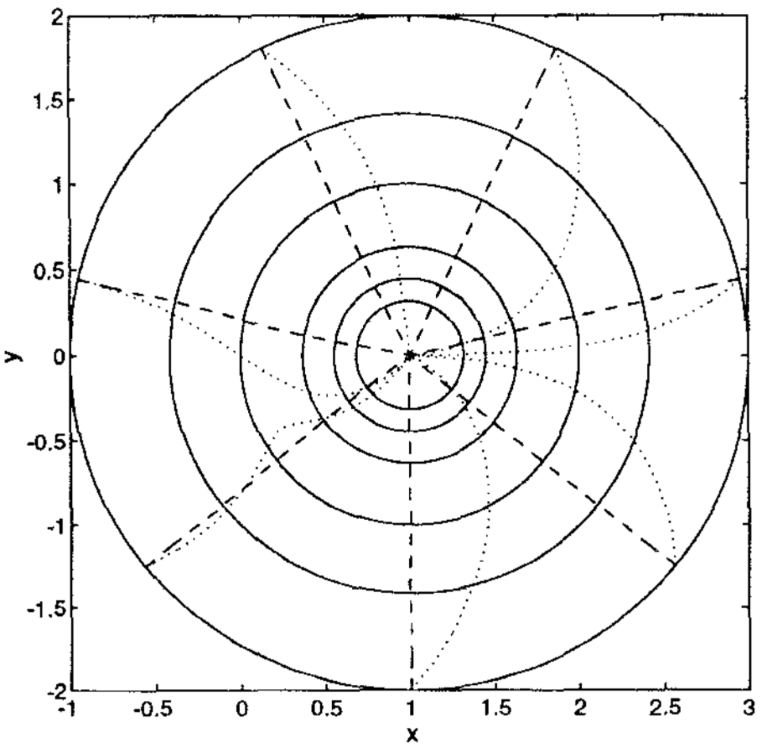
"Сжатые" координаты: $\begin{cases} \tilde{x} = \frac{1}{a} x \\ \tilde{y} = \frac{1}{b} y \end{cases}$, матрица $G = \begin{pmatrix} a^2 & 0 \\ 0 & b^2 \end{pmatrix}$.

Оптимизируем квадратичную функцию $L(\tilde{x}, \tilde{y}) = a^2\tilde{x}^2 + b^2\tilde{y}^2 = x^2 + y^2$.

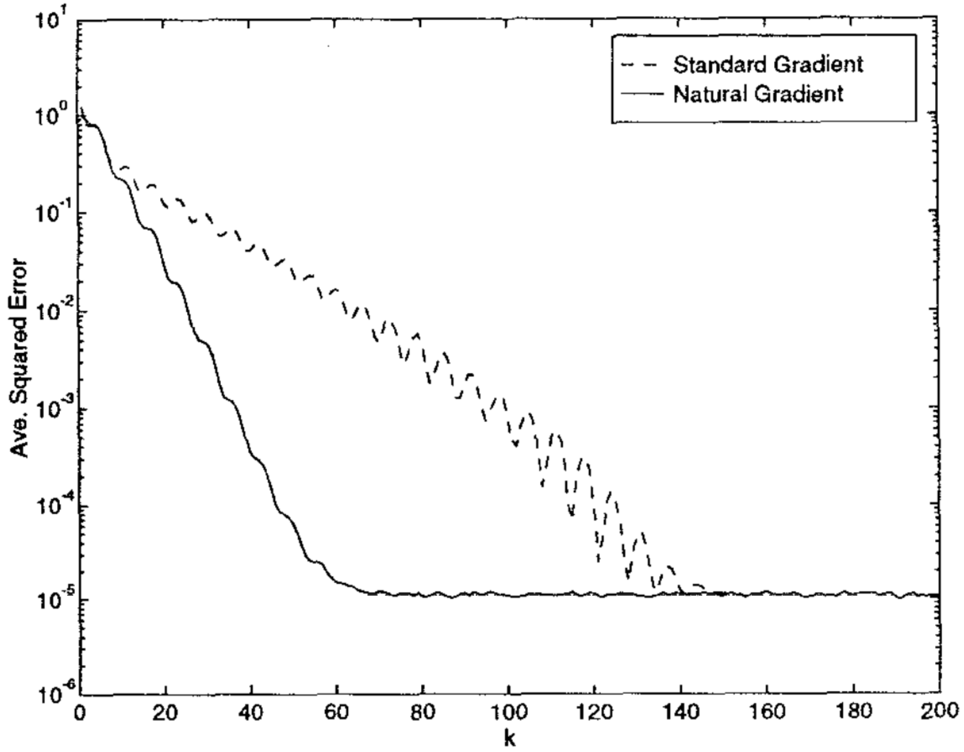
Обычный градиент: $\nabla L = 2\begin{pmatrix} a^2\tilde{x} \\ b^2\tilde{y} \end{pmatrix}$

Натуральный градиент: $\widetilde{\nabla} L = G^{-1}\nabla L = 2\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix}$

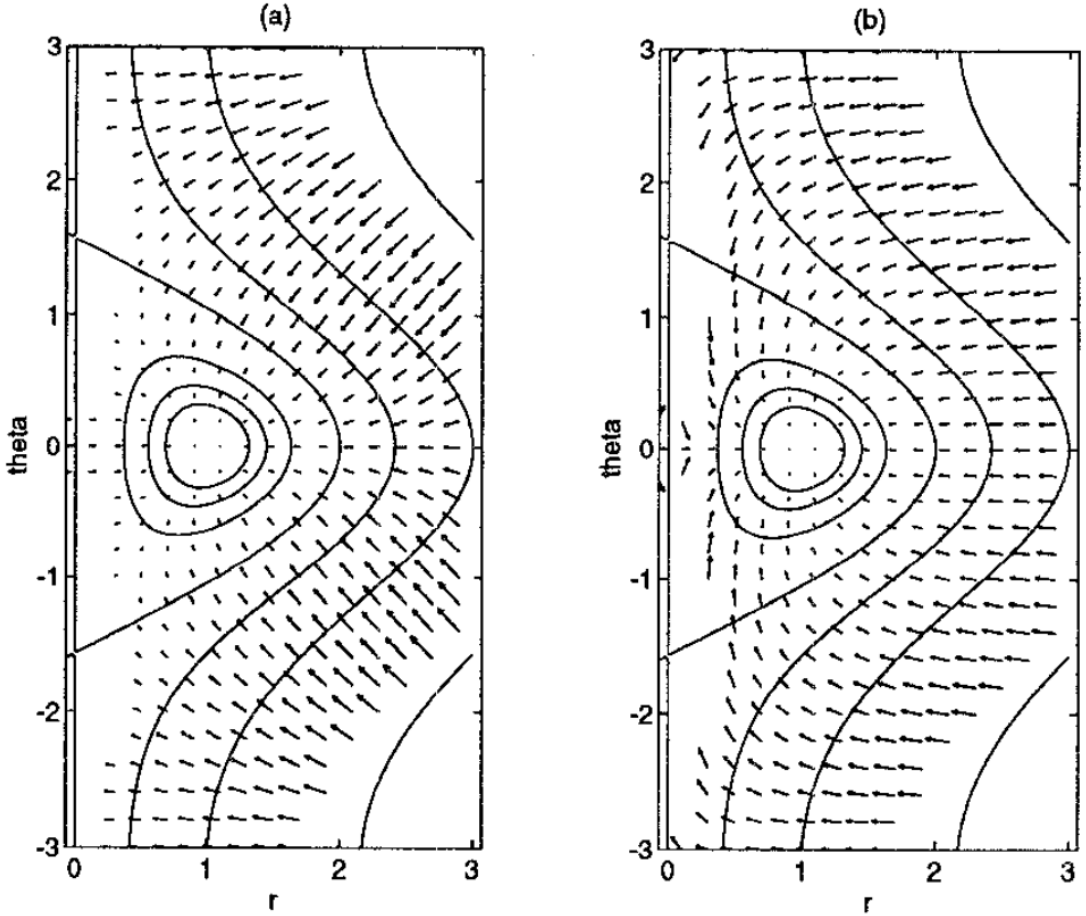




y
x



Standard Gradient
Natural Gradient
Ave. Squared Error
k


(a)
(b)
theta
r
3
2
1
0
-1
-2
-3
0
1
2
3

Натуральный градиент

- Рассмотрим семейство параметрических распределений F, задающее соответствие между параметром θ и функцией плотности $p_\theta = F(\theta)(z)$.
- Тогда расстояние между двумя векторами параметров можно задать в римановом пространстве с помощью KL-дивергенции:

$$D(\theta, \theta + d\theta) = KL(p_\theta || p_{\theta + d\theta})$$

Натуральный градиент

- При малом d можно задать метрику

$$D(\theta, \theta + d\theta) = KL(p_\theta || p_{\theta + d\theta})$$

- Как квадратичную форму с матрицей F – информационная матрица Фишера

$$\mathbf{F}_\theta = \mathbb{E}_\mathbf{z}\left[(\nabla \log p_\theta(\mathbf{z}))^T (\nabla \log p_\theta(\mathbf{z}))\right]$$

$$< \mathbf{u}, \mathbf{v} >_\theta = \mathbf{u}\mathbf{F}_\theta\mathbf{v}$$

Натуральный градиент

- Задача в терминах KL-дивергенции:

$$\arg\min_{\Delta\theta} \mathcal{L}(\theta + \Delta\theta)$$
$$\text{s. t. } KL(p_\theta || p_{\theta+\Delta\theta}) = const.$$

- Разложим дивергенцию в ряд Тейлора:

$$\begin{aligned}
KL(p_\theta \parallel p_{\theta+\Delta\theta}) &\approx (\mathbb{E}_{\mathbf{z}}[\log p_\theta] - \mathbb{E}_{\mathbf{z}}[\log p_\theta]) \\
&\quad - \mathbb{E}_{\mathbf{z}}[\nabla \log p_\theta(\mathbf{z})]\,\Delta\theta - \frac{1}{2}\Delta\theta^T \mathbb{E}_{\mathbf{z}}\left[\nabla^2 \log p_\theta\right]\Delta\theta \\
&= \frac{1}{2}\Delta\theta^T \mathbb{E}_{\mathbf{z}}\left[-\nabla^2 \log p_\theta(\mathbf{z})\right]\Delta\theta \\
&= \frac{1}{2}\Delta\theta^T \mathbf{F} \Delta\theta
\end{aligned}$$

Натуральный градиент

- Запишем лагранжиан: $\mathcal{L}(\theta) + \nabla\mathcal{L}(\theta)\Delta\theta + \frac{1}{2}\lambda\Delta\theta^T\mathbf{F}\Delta\theta$
- Приравняв к нулю получаем результат, аналогичный предыдущему

$$\nabla_N \mathcal{L}(\theta) \stackrel{\text{def}}{=} \nabla\mathcal{L}(\theta)\mathbb{E}_{\mathbf{z}}\left[\left(\nabla \log p_\theta(\mathbf{z})\right)^T \left(\nabla \log p_\theta(\mathbf{z})\right)\right]^{-1}$$
$$\stackrel{\text{def}}{=} \nabla\mathcal{L}(\theta)\mathbf{F}^{-1}.$$

Натуральный градиент

- Адаптация для нейросетей:

$$\arg\min_{\Delta\theta} \mathcal{L}(\theta + \Delta\theta)$$
$$\text{s. t. } \mathbb{E}_{\mathbf{x}\sim\tilde{q}(\mathbf{x})} \left[KL(p_\theta(\mathbf{t}|\mathbf{x})||p_{\theta+\Delta\theta}(\mathbf{t}|\mathbf{x}))\right] = const.$$

- Можно вывести формулы для специфических функций активации:

$$\mathbf{F}_{linear} = \beta^2 \mathbb{E}_{\mathbf{x}\sim\tilde{q}} \left[\frac{\partial \mathbf{y}}{\partial \theta}^T \frac{\partial \mathbf{y}}{\partial \theta} \right] = \beta^2 \mathbb{E}_{\mathbf{x}\sim\tilde{q}} \left[\mathbf{J}_{\mathbf{y}}^T \mathbf{J}_{\mathbf{y}} \right]$$

$$\mathbf{F}_{sigmoid} = \mathbb{E}_{\mathbf{x}\sim\tilde{q}} \left[\mathbf{J}_{\mathbf{y}}^T \operatorname{diag}\left(\frac{1}{\mathbf{y}(1-\mathbf{y})} \right) \mathbf{J}_{\mathbf{y}} \right]$$

SGD, Momentum, Adagrad

- Обычный SGD:

$$\theta_{t+1} = \theta_t - \eta \nabla \mathcal{L}(\theta_t)$$

- Momentum:

$$v_{t+1} = \mu v_t - \eta \nabla \mathcal{L}(\theta_t)$$
$$\theta_{t+1} = \theta_t + v_{t+1}$$

- Adagrad:

$$g_{t+1} = g_t + \nabla \mathcal{L}(\theta_t)^2$$
$$\theta_{t+1} = \theta_t - \frac{\eta \nabla \mathcal{L}(\theta_t)}{\sqrt{g_{t+1}} + \epsilon}$$

RMSProp, Adadelta

- RMSProp:

$$g_{t+1} = \gamma g_t + (1-\gamma)\nabla\mathcal{L}(\theta_t)^2$$
$$\theta_{t+1} = \theta_t - \frac{\eta\nabla\mathcal{L}(\theta_t)}{\sqrt{g_{t+1}} + \epsilon}$$

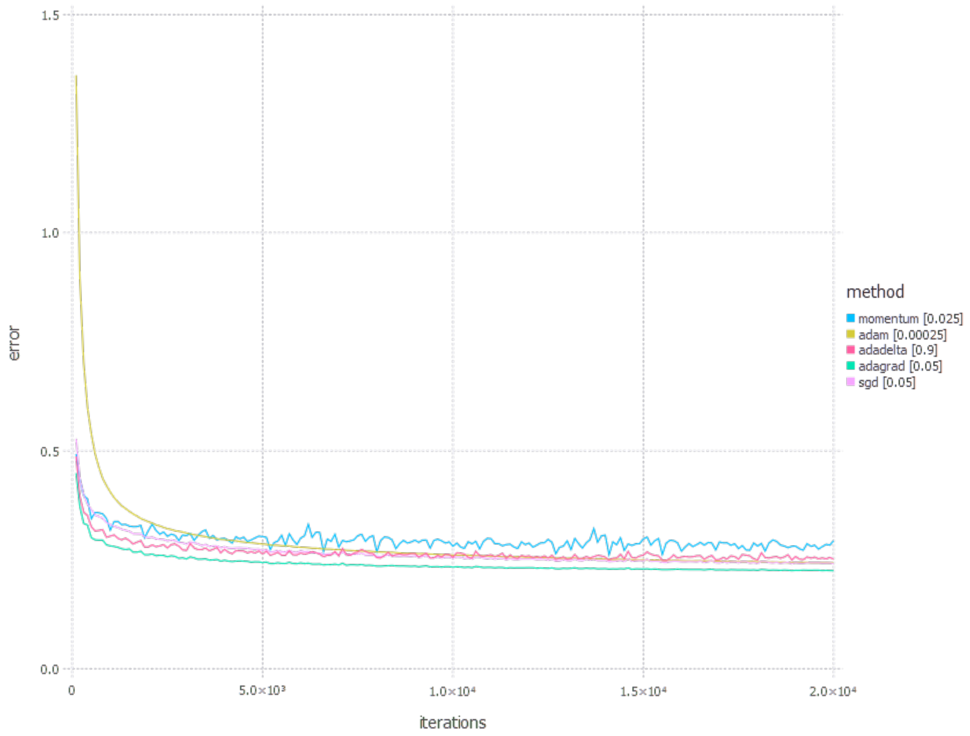
- Adadelta:

$$g_{t+1} = \gamma g_t + (1-\gamma)\nabla\mathcal{L}(\theta_t)^2$$
$$v_{t+1} = -\frac{\sqrt{x_t + \epsilon}\,\nabla\mathcal{L}(\theta_t)}{\sqrt{g_{t+1} + \epsilon}}$$
$$x_{t+1} = \gamma x_t + (1-\gamma)v_{t+1}^2$$
$$\theta_{t+1} = \theta_t + v_{t+1}$$

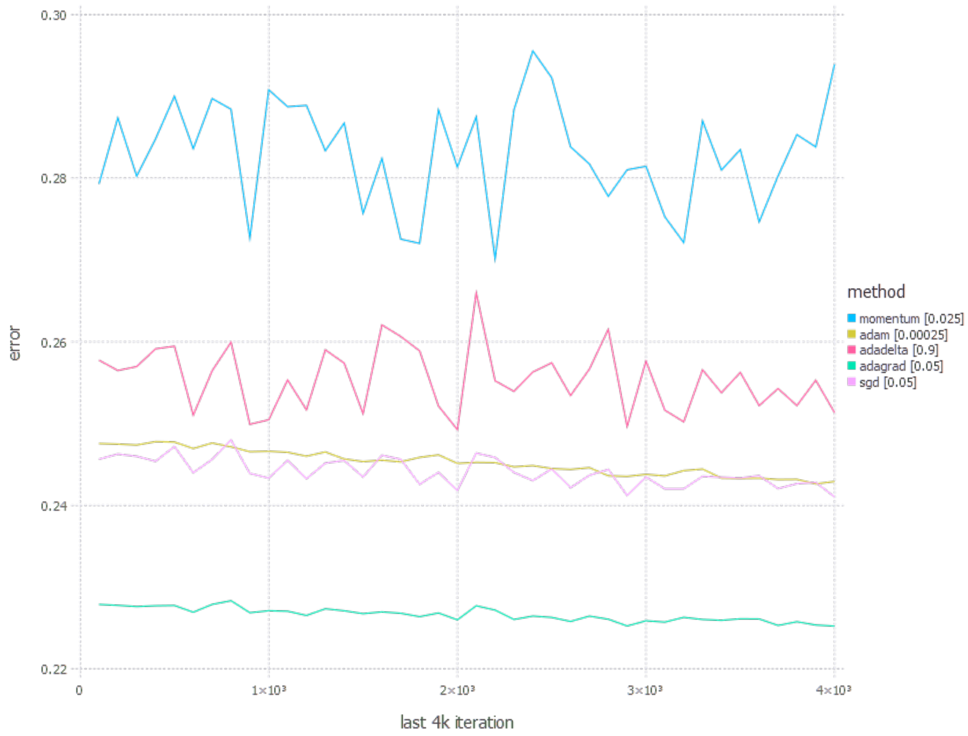
Adam

$$
\begin{aligned}
m_{t+1} &= \gamma_1 m_t + (1-\gamma_1)\nabla\mathcal{L}(\theta_t) \\
g_{t+1} &= \gamma_2 g_t + (1-\gamma_2)\nabla\mathcal{L}(\theta_t)^2 \\
\hat{m}_{t+1} &= \frac{m_{t+1}}{1-\gamma_1^{t+1}} \\
\hat{g}_{t+1} &= \frac{g_{t+1}}{1-\gamma_2^{t+1}} \\
\theta_{t+1} &= \theta_t - \frac{\eta \hat{m}_{t+1}}{\sqrt{\hat{g}_{t+1}}+\epsilon}
\end{aligned}
$$

Сравнение

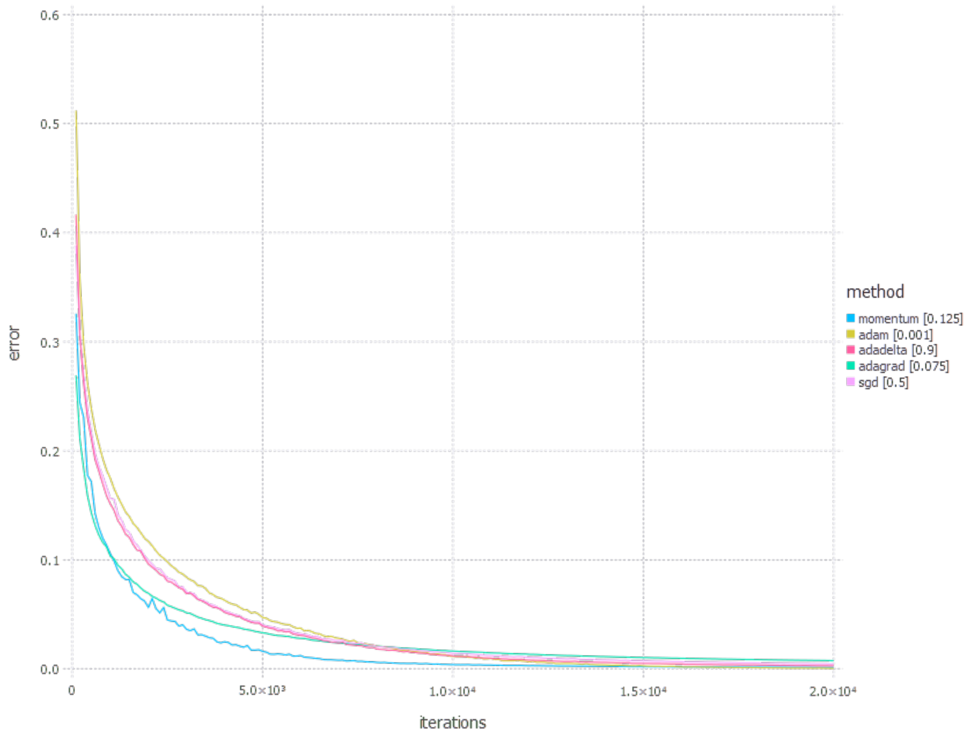


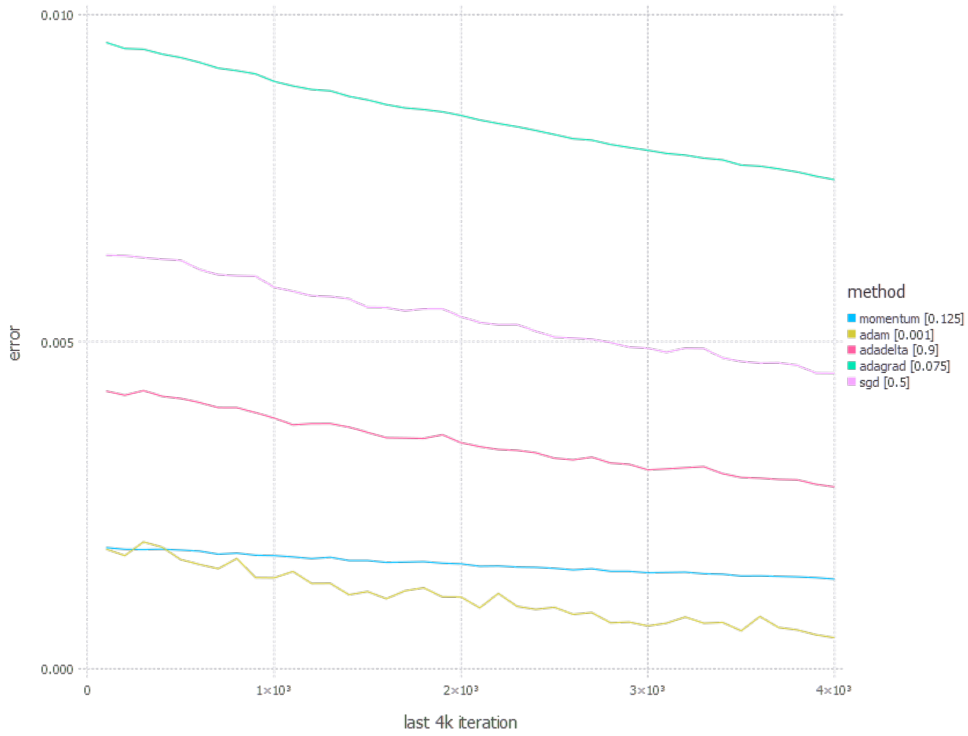
1.5
1.0
0.5
0.0
error
0
5.0×10³
1.0×10⁴
1.5×10⁴
2.0×10⁴
iterations
method
momentum [0.025]
adam [0.00025]
adadelta [0.9]
adagrad [0.05]
sgd [0.05]



0.30
0.28
0.26
0.24
0.22
error
0
1×10³
2×10³
3×10³
4×10³
last 4k iteration
method
momentum [0.025]
adam [0.00025]
adadelta [0.9]
adagrad [0.05]
sgd [0.05]

Сравнение


0.6
0.5
0.4
0.3
0.2
0.1
0.0
error
0
5.0×10³
1.0×10⁴
1.5×10⁴
2.0×10⁴
iterations
method
momentum [0.125]
adam [0.001]
adadelta [0.9]
adagrad [0.075]
sgd [0.5]


0.010
0.005
0.000
error
0
1×10³
2×10³
3×10³
4×10³
last 4k iteration
method
momentum [0.125]
adam [0.001]
adadelta [0.9]
adagrad [0.075]
sgd [0.5]