Градиентные методы оптимизации гиперпараметров моделей глубокого обучения

О. Ю. Бахтеев

Научный руководитель: д.ф.-м.н. В.В. Стрижов
Московский Физико-Технический Институт (Государственный Университет)

ММРО-2017

09.10.2017

Цель работы

Исследуются

Методы автоматического порождения и прореживания моделей глубокого обучения.

Требуется

Предложить алгоритм нахождения оптимальных значений гиперпараметров (параметров распределения параметров) модели.

Проблемы нахождения оптимальных значений гиперпараметров

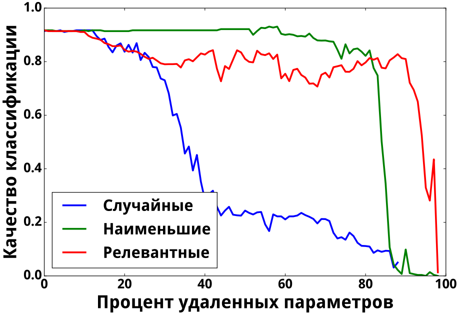
- Многоэкстремальность задачи оптимизации параметров модели,
- Вычислительная сложность оптимизации,
- Большое количество гиперпараметров.

Решение

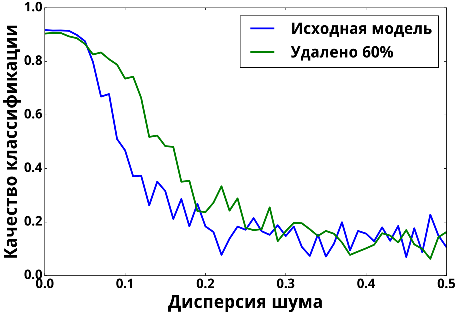
Предлагается оптимизировать параметры и гиперпараметры модели в единой процедуре с использованием градиентных методов. В качестве критерия оптимальности модели рассматривается нижняя оценка правдоподобия модели.

Проблемы обучения сетей

Правдоподобие моделей с избыточным количеством параметров не меняется при удалении параметров.

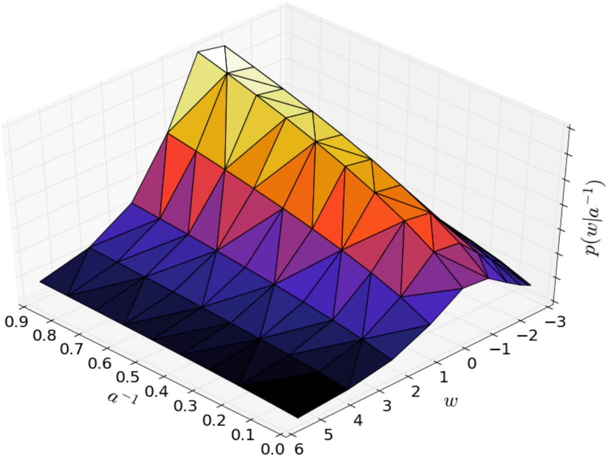


Избыточность параметров модели



Неустойчивость модели

Зависимость правдоподобия от гиперпараметров



Формальная постановка задачи

Задана дифференцируемая по параметрам модель $\mathbf{f}(\mathbf{w}, \mathbf{x})$, задающая правдоподобие выборки $p(\mathfrak{D}|\mathbf{w})$. Задано априорное распределение параметров модели:

$$p(\mathbf{w}|\mathbf{A}^{-1}) \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}) = \mathcal{N}(\mathbf{0}, \mathrm{diag}[\alpha_1, \dots, \alpha_n]^{\mathsf{T}}).$$

Пусть $\boldsymbol{\theta} \in \mathbb{R}^s$ — множество параметров, подлежащих оптимизации (соответствует параметрам модели $\mathbf{w}$),
L — оптимизируемая функция потерь,
Q — критерий качества модели.
Итоговая задача оптимизации:

$$\hat{\mathbf{A}}^{-1} = \underset{[\alpha_1, \dots, \alpha_n]^{\mathsf{T}} \in \mathbb{R}^n}{\arg\max} Q(\hat{\boldsymbol{\theta}}(\mathbf{A}^{-1}), \mathbf{A}^{-1}, \mathfrak{D}),$$

$$\hat{\boldsymbol{\theta}}(\mathbf{A}^{-1}) = \underset{\boldsymbol{\theta} \in \mathbb{R}^s}{\arg\min} L(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathfrak{D}).$$

L и Q: Кросс-валидация

Разобьем выборку $\mathfrak{D}$ на k равных частей:

$$\mathfrak{D} = \mathfrak{D}_1 \sqcup \dots \sqcup \mathfrak{D}_k.$$

Запустим k оптимизаций модели, r-я модель обучается на выборках $\mathfrak{D}^r = \mathfrak{D}_1, \dots, \mathfrak{D}_{r-1}, \mathfrak{D}_{r+1}, \dots, \mathfrak{D}_k$.
Положим $\boldsymbol{\theta} = [\mathbf{w}_1, \dots, \mathbf{w}_k]$ — параметры всех запусков модели.

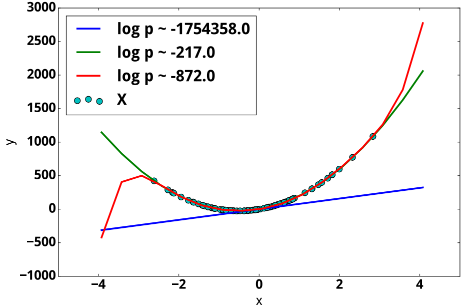
$$L(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathfrak{D}) = -\frac{1}{k}\sum_{r=1}^{k}\left(\frac{k}{k-1}\log p(\mathfrak{D}^r|\mathbf{w}_r) - \log p(\mathbf{w}_r|\mathbf{A}^{-1})\right).$$

$$Q(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathfrak{D}) = \frac{1}{k}\sum_{r=1}^{k} k \log p(\mathfrak{D}_r|\mathbf{w}_r).$$

Правдоподобие модели

Модель $\mathbf{f} \in \mathfrak{F}$ оптимальна, если достигается максимум **правдоподобия модели**:

$$p(\mathfrak{D}|\mathbf{A}^{-1}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{A}^{-1})d\mathbf{w}.$$



Пусть q — непрерывное распределение.

$$\log p(\mathfrak{D}|\mathbf{A}^{-1}) \geq \int q(\mathbf{w}) \log \frac{p(\mathfrak{D}, \mathbf{w}|\mathbf{A}^{-1})}{q(\mathbf{w})} d\mathbf{w} =$$

$$= \int q(\mathbf{w}) \log p(\mathfrak{D}|\mathbf{w}, \mathbf{A}^{-1}) d\mathbf{w} - \mathrm{D_{KL}},$$

где

$$\mathrm{D_{KL}} = -\int q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathbf{A}^{-1})}{q(\mathbf{w})} d\mathbf{w}.$$

Вариационная оценка на основе мультистарта

$$\log p(\mathfrak{D}|\mathbf{A}) \geq \mathsf{E}_{q(\mathbf{w})}[\log\ p(\mathfrak{D}, \mathbf{w}|\mathbf{A}^{-1})] - \mathsf{S}(q(\mathbf{w})),$$

S — энтропия.

Теорема [Бахтеев, 2016]. Пусть L — функция потерь, градиент которой — непрерывно-дифференцируемая функция с константой Липшица C. Пусть $\boldsymbol{\theta} = [\mathbf{w}^1, \dots, \mathbf{w}^k]$ — начальные приближения оптимизации модели. Пусть γ — шаг градиентного спуска, такой что:

- $\gamma < \frac{1}{C}$,
- $\gamma^{(-1)} > \max_{r \in \{1,\dots,k\}} \lambda_{\max}(\mathbf{H}(\mathbf{w}^r))$.

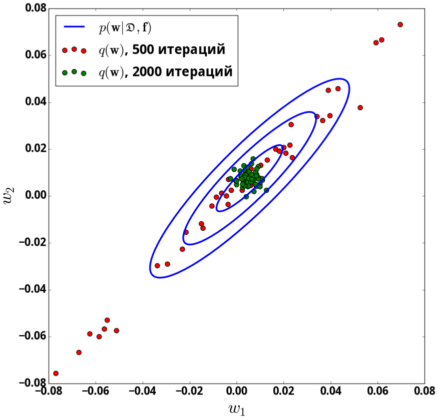
Тогда

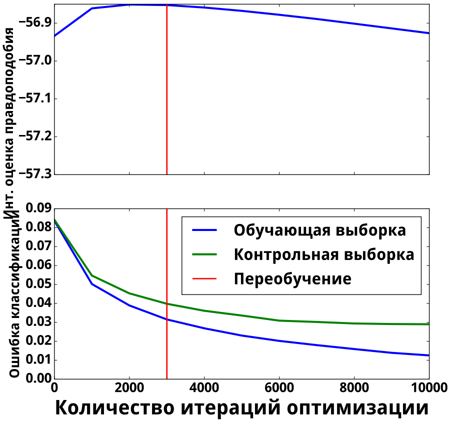
$$\mathsf{S}(q^\tau(\mathbf{w})) - \mathsf{S}(q^{\tau-1}(\mathbf{w})) \sim \frac{1}{k}\sum_{r=1}^{k}\big(\gamma\, Tr[\mathbf{H}(\mathbf{w}^r)] - \gamma^2\, Tr[\mathbf{H}(\mathbf{w}^r)\mathbf{H}(\mathbf{w}^r)]\big) + o_{\gamma\to 0}(1),$$

где $\mathbf{H}$ — гессиан функции потерь L, q^τ — распределение $q(\mathbf{w})$ в момент оптимизации τ.

Вариационная оценка с использованием градиентного спуска

Максимизация вариационной оценки эквивалентна минимизации $D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D}, \mathbf{A}^{-1}))$. Градиентный спуск не минимизирует $D_{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D}, \mathbf{A}^{-1}))$.





L и Q: Вариационная оценка

Пусть $q = \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}), \quad \boldsymbol{\theta} = [\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}]$.
Тогда вариационная оценка имеет вид:

$$\int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathfrak{D}, \mathbf{w}, \mathbf{A}^{-1}) d\mathbf{w} - D_{\mathrm{KL}}\bigl(q(\mathbf{w}) || p(\mathbf{w}|\mathbf{A}^{-1})\bigr) \simeq$$

$$\sum_{i=1}^{m} \log p(\mathbf{x}_i|\mathbf{w}_i) - D_{\mathrm{KL}}\bigl(q(\mathbf{w}) || p(\mathbf{w}|\mathbf{A}^{-1})\bigr) = -L(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathfrak{D}) = Q(\boldsymbol{\theta}, \mathbf{A}^{-1}, \mathfrak{D}),$$

где $\mathbf{w}_i \sim q$.
Дивергенция $D_{\mathrm{KL}}\bigl(q(\mathbf{w}) || p(\mathbf{w}|\mathbf{A}^{-1})\bigr)$ вычисляется аналитически:

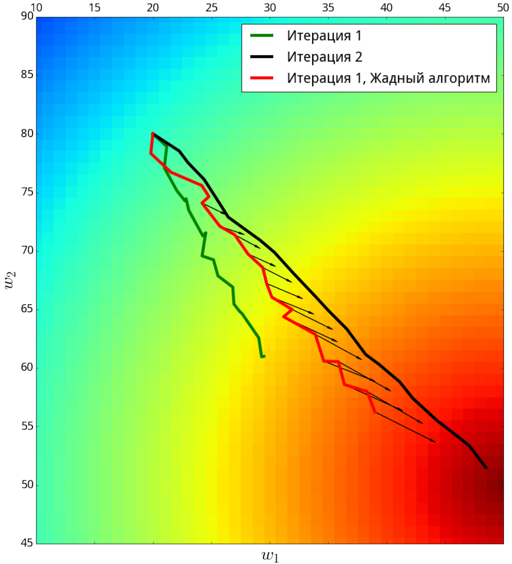
$$D_{\mathrm{KL}}\bigl(q(\mathbf{w}) || p(\mathbf{w}|\mathbf{A}^{-1})\bigr) = \frac{1}{2}\bigl(\mathrm{tr}(\mathbf{A}\mathbf{A}_q^{-1}) + \boldsymbol{\mu}_q^{\mathsf{T}}\mathbf{A}\boldsymbol{\mu}_q - n + \ln|\mathbf{A}^{-1}| - \ln|\mathbf{A}_q^{-1}|\bigr).$$

Общая схема алгоритма оптимизации

Вход: количество итераций оптимизации гиперпараметров ℓ, количество итераций оптимизации параметров τ, длина шага градиентного спуска γ.

1. Повторять в цикле от $1, \dots, \ell$:
   1. Инициализировать параметры $\boldsymbol{\theta}_0$.
   2. Провести оптимизацию параметров с использованием стохастического градиентного спуска:
   $$\hat{\boldsymbol{\theta}} = T \circ T \circ \dots \circ T(\boldsymbol{\theta}_0) = T^{\tau}(\boldsymbol{\theta}_0),$$
   где
   $$T(\boldsymbol{\theta}) = \boldsymbol{\theta} - \gamma \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{A}^{-1}, \hat{\mathfrak{D}}),$$
   $\hat{\mathfrak{D}}$ — случайная подвыборка $\mathfrak{D}$.
   3. Провести оптимизацию гиперпараметров: $Q(\hat{\boldsymbol{\theta}}(\mathbf{A}^{-1}), \mathbf{A}^{-1}, \mathfrak{D}) \to \max$.

HOAG и Жадный алгоритм



Жадный алгоритм

$$\mathbf{A}'^{-1} = \mathbf{A}^{-1} - \gamma_{\mathbf{A}}(\nabla_{\mathbf{A}^{-1}} Q(T(\boldsymbol{\theta}), \mathbf{A}^{-1}, \mathfrak{D})),$$

где $\gamma_{\mathbf{A}}$ — длина шага оптимизации гиперпараметров.

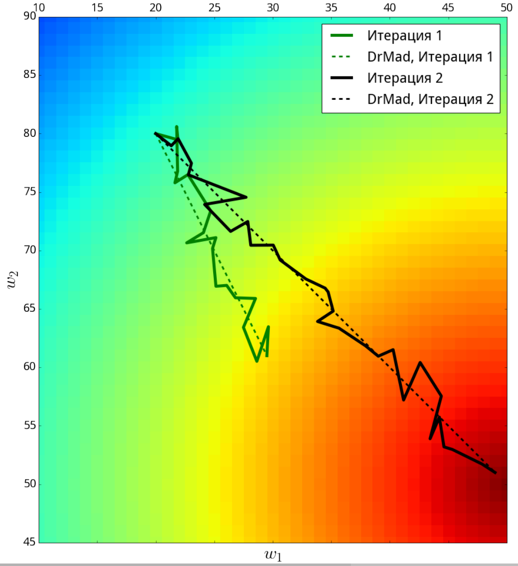
HOAG

$$\mathbf{A}'^{-1} = \mathbf{A}^{-1} - \gamma_{\mathbf{A}} \hat{\nabla}_{\mathbf{A}^{-1}} Q(T^{\tau}(\boldsymbol{\theta}_0), \mathbf{A}^{-1}, \mathfrak{D})),$$

где $\hat{\nabla}_{\mathbf{A}^{-1}}$ — численное приближение градиента:

$$\nabla_{\mathbf{A}^{-1}} Q(\hat{\boldsymbol{\theta}}) - \left(\frac{\partial^2 L}{\partial \boldsymbol{\theta} \partial \mathbf{A}^{-1}}\right)^{\mathsf{T}} \mathbf{H}(L(\boldsymbol{\theta}))^{-1} \nabla_{\boldsymbol{\theta}} Q.$$

Алгоритм DrMad



Рассматривается оптимизация функции $Q(T^\tau(\theta_0), \mathbf{A}^{-1}, \mathfrak{D}))$ по **всей** истории оптимизации параметров.

Вводятся предположения о линейности траектории оптимизации параметров. Градиент $\nabla_{\mathbf{A}}^{-1} Q$ аккумулируется по правилу:

$$\nabla_{\mathbf{A}^{-1}} Q = \nabla_{\mathbf{A}^{-1}} Q - \gamma \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^\tau) \nabla_{\mathbf{A}^{-1}} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^\tau).$$

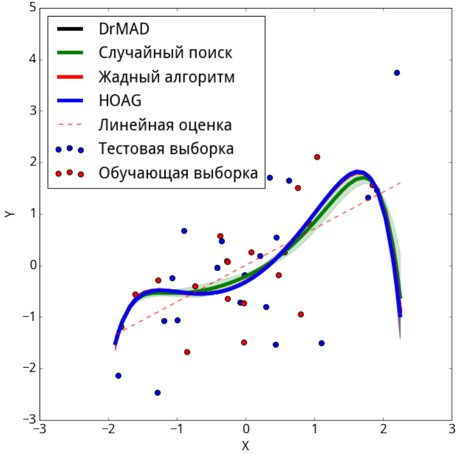
Вычислительный эксперимент

Цель эксперимента: анализ рассматриваемых алгоритмов и итоговых моделей.
Данные:

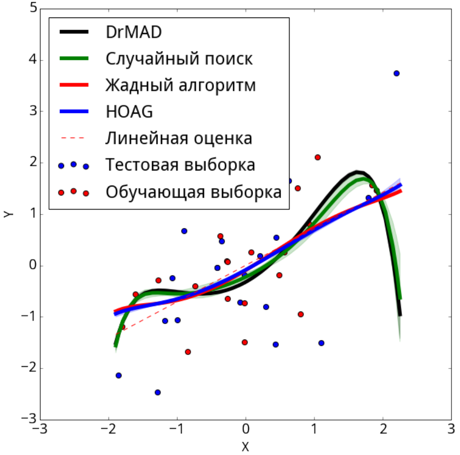
- Синтетические данные: 40 точек на плоскости. Модель: полином 12 степени.
- Набор записей акселерометра WISDM. Рассматривается задача регрессии с нейронной сетью с одним скрытым слоем (10 нейронов).
- Набор рукописных цифр MNIST. Рассматривается задача регрессии с нейронной сетью с одним скрытым слоем (300 нейронов).

В качестве Q и L рассматривается кросс-валидация ($k = 4$) и вариационная оценка.

Синтетические данные: результат

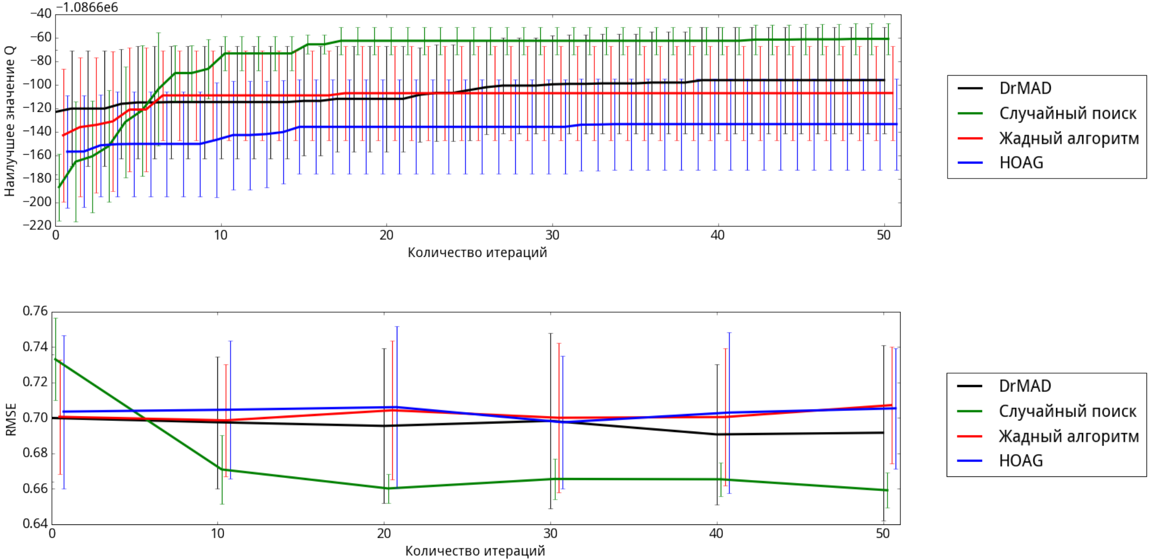


Кросс-Валидация

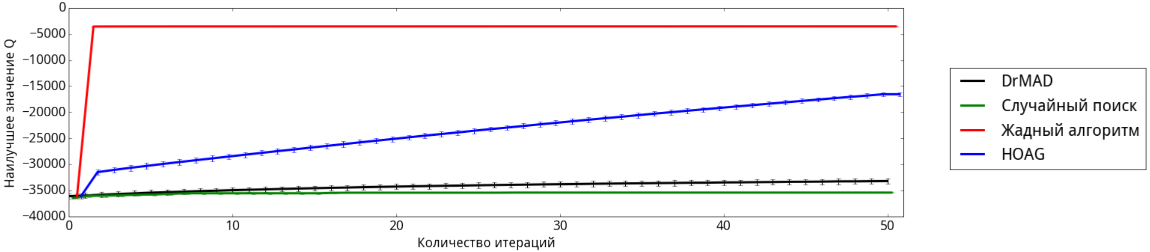


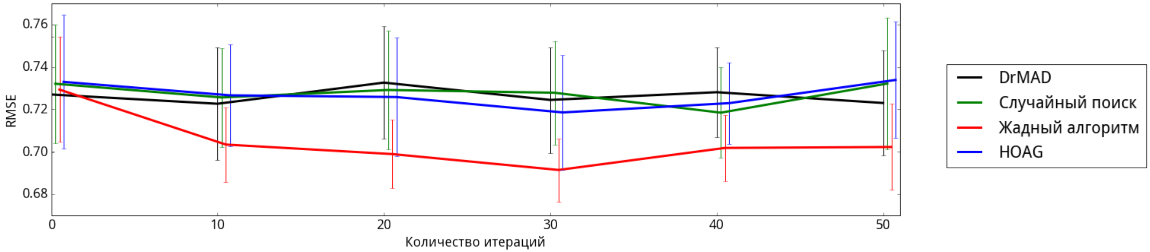
Вариационная оценка

WISDM: кросс-валидация

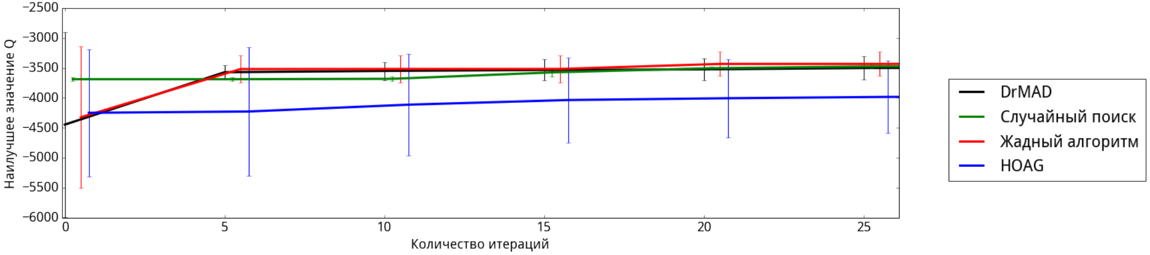


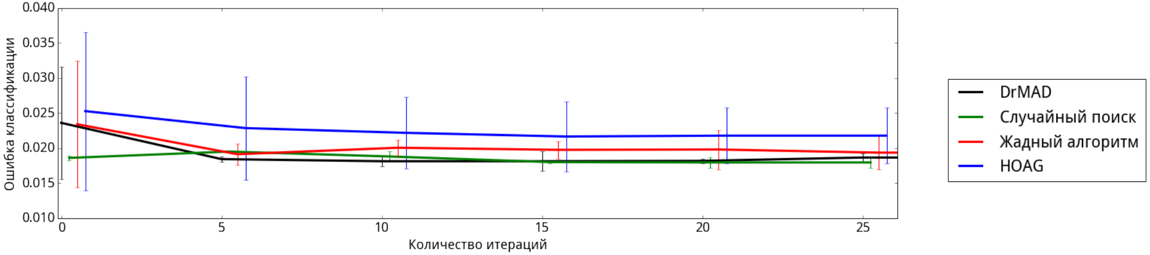
WISDM: вариационная оценка



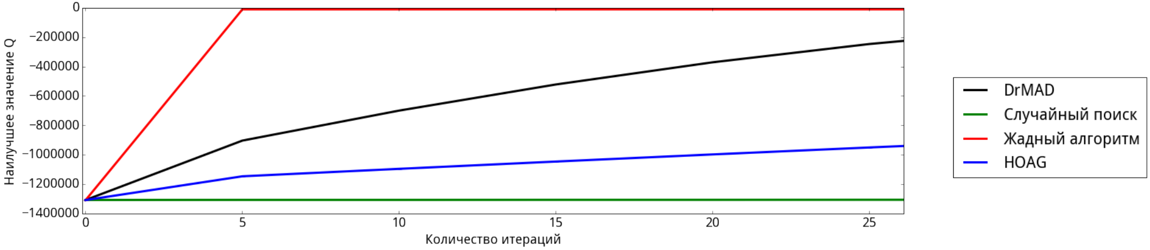


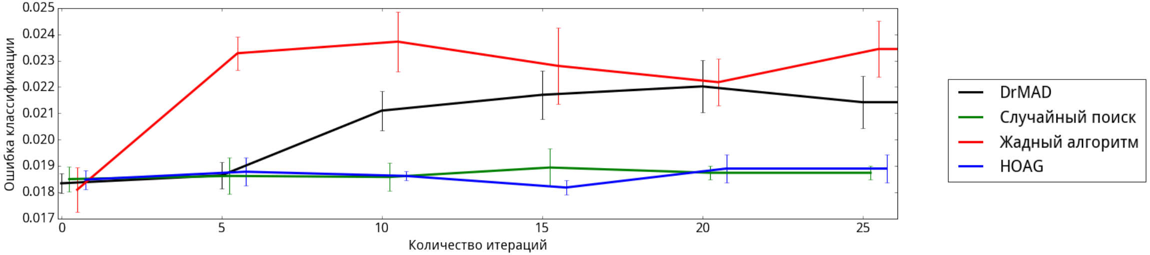
MNIST: кросс-валидация





MNIST: вариационная оценка

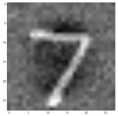




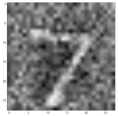
MNIST: добавление шума

Добавление гауссового шума $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$:

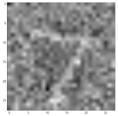
Без шума



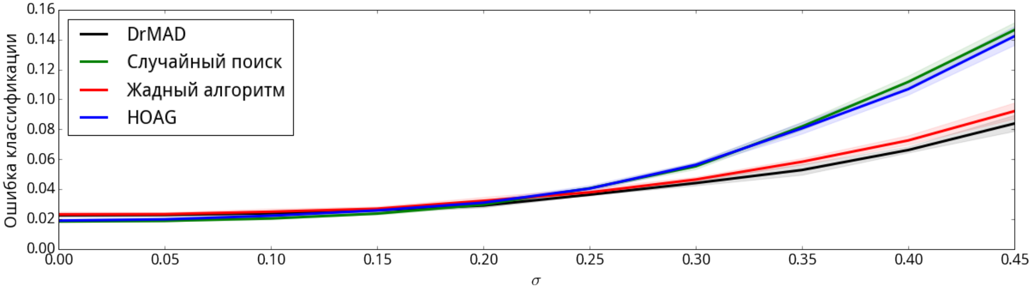
$\sigma = 0.1$



$\sigma = 0.25$



$\sigma = 0.5$



Заключение

- Предложен критерий оптимальной сложности модели глубокого обучения.
- Исследована зависимость интегральной оценки правдоподобия от устойчивости модели и возможности переобучения.
- Предложен алгоритм выбора субоптимальной модели классификации без использования кросс-валидации.
- Рассмотрены градиентные алгоритмы оптимизации гиперпараметров.
- Проведены эксперименты на ряде выборок в задачах классификации и регрессии.
- Наилучшие результаты показали алгоритмы жадной оптимизации.

Исследование основывается на следующих работах

- Graves A. Practical variational inference for neural networks //Advances in Neural Information Processing Systems. – 2011
- Maclaurin D., Duvenaud D., Adams R. Gradient-based hyperparameter optimization through reversible learning //International Conference on Machine Learning. – 2015
- Luketina J. et al. Scalable gradient-based tuning of continuous regularization hyperparameters //International Conference on Machine Learning. - 2016
- J. Fu et al., DrMAD: Distilling Reverse-Mode Automatic Differentiation for Optimizing Hyperparameters of Deep Neural Networks // IJCAI - 2016
- Pedregosa F. Hyperparameter optimization with approximate gradient //International Conference on Machine Learning. – 2016. –