Оптимизация гиперпараметров градиентными методами

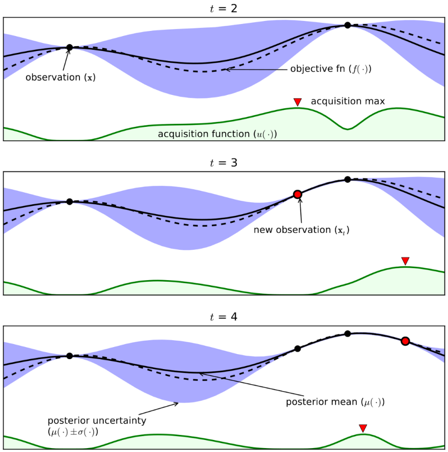
Бахтеев Олег

МФТИ

04.04.2018

Градиентные методы: зачем?

- Гиперпараметры — параметры распределения параметров модели.
- Основные методы оптимизации не позволяют проводить оптимизацию большого количества гиперпараметров (>10).
- Решение проблемы — использование градиентного спуска для гиперпараметров.



Shahriari et. al, 2016. Пример работы гауссового процесса.

Постановка задачи

Задана дифференцируемая по параметрам модель, приближающая зависимую переменную y:

$$f : \mathbb{R}^n \to \mathbb{Y}, \quad \mathbf{w} \in \mathbb{R}^u.$$

Функция f задает правдоподобие выборки $\log p(\mathbf{y}|\mathbf{X}, f)$.
Пусть также задано априорное распределение параметров:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}),$$

где $\mathbf{A}^{-1} = \text{diag}[\alpha_1, \dots, \alpha_u]^{-1}$ — матрица коварияций диагонального вида, определеяемая *гиперпараметрами* $[\alpha_1, \dots, \alpha_u]$.

Кросс-валидация

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

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

Запустим k оптимизаций модели, каждую на своей части выборки. Положим $\boldsymbol{\theta} = [\mathbf{w}_1, \dots, \mathbf{w}_k]$, где $\mathbf{w}_1, \dots, \mathbf{w}_k$ — параметры модели при оптимизации k.
Пусть L — функция потерь:

$$L(\boldsymbol{\theta}, \mathbf{h}) = -\frac{1}{k}\sum_{q=1}^{k}\bigl(\frac{k}{k-1}\log p(\mathbf{y} \setminus \mathbf{y}_q | \mathbf{X} \setminus \mathbf{X}_q, \mathbf{w}_q) + \log p(\mathbf{w}_q | \mathbf{A})\bigr). \quad (1)$$

Пусть Q — функция качества модели:

$$Q(\boldsymbol{\theta}, \mathbf{h}) = \frac{1}{k}\sum_{q=1}^{k} k \log p(\mathbf{y}_q | \mathbf{X}_q, \mathbf{w}_q).$$

Формальная постановка задачи

Задана дифференцируемая по параметрам модель, приближающая зависимую переменную y:

$$f : \mathbb{R}^n \to \mathbb{Y}, \quad \mathbf{w} \in \mathbb{R}^u.$$

Пусть $\boldsymbol{\theta} \in \mathbb{R}^s$ — множество всех оптимизируемых параметров.
$L(\boldsymbol{\theta}, \mathbf{h})$ — дифференцируемая функция потерь по которой производится оптимизация функции f.
$Q(\boldsymbol{\theta}, \mathbf{h})$ — дифференцируемая функция определяющая итоговое качество модели f и приближающая интеграл.
Требуется найти параметры $\hat{\boldsymbol{\theta}}$ и гиперпараметры $\hat{\mathbf{h}}$ модели, доставляющие минимум следующему функционалу:

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h} \in \mathbb{R}^h} Q(\hat{\boldsymbol{\theta}}(\mathbf{h}), \mathbf{h}),$$

$$\hat{\boldsymbol{\theta}}(\mathbf{h}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^s} L(\boldsymbol{\theta}, \mathbf{h}).$$

Байесовый подход к сложности

Правдоподобие модели ("Evidence"):

$$p(\mathbf{y}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{A})d\mathbf{w}.$$

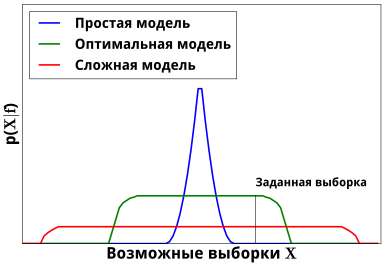
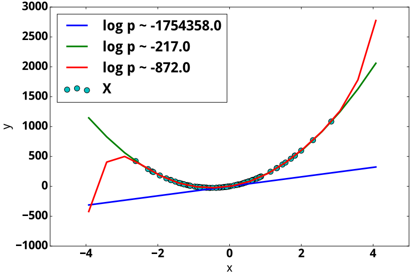


Схема выбора модели по правдоподобию



Пример: полиномы

Вариационная нижняя оценка

Пусть задано непрерывное распределение q. Тогда

$$\log\, p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \int_{\mathbf{w}} q(\mathbf{w})\log\, \frac{p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{A})}{q(\mathbf{w})} d\mathbf{w} + \mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{y},\mathbf{X},\mathbf{A})) \geq$$

$$\geq \int_{\mathbf{w}} q(\mathbf{w})\log\, \frac{p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{A})}{q(\mathbf{w})} d\mathbf{w} =$$

$$= -\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{A})) + \int_{\mathbf{w}} q(\mathbf{w})\log\, p(\mathbf{y}|\mathbf{X},\mathbf{w},\mathbf{A}) d\mathbf{w},$$

где

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{A})) = -\int_{\mathbf{w}} q(\mathbf{w})\log\, \frac{p(\mathbf{w}|\mathbf{A})}{q(\mathbf{w})} d\mathbf{w}.$$

Evidence: нормальное распределение

“Обычная” функция потерь:

$$L = \sum_{\mathbf{x},\mathbf{y}\in\mathfrak{D}} -\log p(\mathbf{y}|\mathbf{x},\mathbf{w}) + \lambda||\mathbf{w}||_2^2.$$

Вариационный вывод при $p(\mathbf{w}|\mathbf{f}) \sim \mathcal{N}(\mathbf{0},\mathbf{1})$:

$$L = \sum_{\mathbf{x},\mathbf{y}\in\mathfrak{D}} \log\, p(\mathbf{y}|\mathbf{x},\hat{\mathbf{w}}) + \frac{1}{2}\big(\mathrm{tr}(\mathbf{A}_q^{-1}) + \boldsymbol{\mu}_q^{\mathsf{T}}\mathbf{A}^{-1}\boldsymbol{\mu}_q - \ln|\mathbf{A}_q^{-1}|\big),$$

$$\hat{\mathbf{w}} \sim q = \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}).$$

Вариационная оценка: оптимизация гиперпараметров

Пусть $L = -Q$:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{A}) \geq \sum_{\mathbf{x},y} \log p(y|\mathbf{x}, \hat{\mathbf{w}}) - D_{\text{KL}}\bigl(q(\mathbf{w})||p(\mathbf{w}|\mathbf{A})\bigr) = -L(\boldsymbol{\theta}, \mathbf{h}) = Q((\boldsymbol{\theta}, \mathbf{h}), \quad (2)$$

где q — нормальное распределение с диагональной матрицей коварияций:

$$q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}), \quad (3)$$

$$D_{\text{KL}}\bigl(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})\bigr) = \frac{1}{2}\bigl(\text{Tr}[\mathbf{A}\mathbf{A}_q^{-1}] + (\boldsymbol{\mu} - \boldsymbol{\mu}_q)^{\mathsf{T}}\mathbf{A}(\boldsymbol{\mu} - \boldsymbol{\mu}_q) - u + \ln|\mathbf{A}^{-1}| - \ln|\mathbf{A}_q^{-1}|\bigr).$$

В качестве оптимизируемых параметров $\boldsymbol{\theta}$ выступают параметры распределения q:

$$\boldsymbol{\theta} = [\alpha_1, \dots, \alpha_u, \mu_1, \dots, \mu_u].$$

Формальная постановка задачи: градиентная оптимизация

Определение

Оператором T назовем оператор стохастического градиентного спуска, производящий η шагов оптимизации:

$$\hat{\boldsymbol{\theta}} = T \circ T \circ \cdots \circ T(\boldsymbol{\theta}_0, \mathbf{h}) = T^{\eta}(\boldsymbol{\theta}_0, \mathbf{h}), \tag{4}$$

где

$$T(\boldsymbol{\theta}, \mathbf{h}) = \boldsymbol{\theta} - \gamma \nabla L(\boldsymbol{\theta}, \mathbf{h})|_{\hat{\mathfrak{D}}},$$

γ — длина шага градиентного спуска, $\boldsymbol{\theta}_0$ — начальное значение параметров $\boldsymbol{\theta}$, $\hat{\mathfrak{D}}$ — случайная подвыборка исходной выборки $\mathfrak{D}$.

Перепишем итоговую задачу оптимизации:

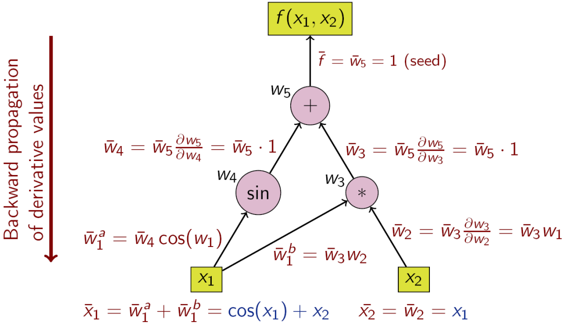
$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h} \in \mathbb{R}^{h}} Q(T^{\eta}(\boldsymbol{\theta}_0, \mathbf{h})),$$

где $\boldsymbol{\theta}_0$ — начальное значение параметров $\boldsymbol{\theta}$.

RMAD, Maclaurin et. al, 2015

1. Провести η шагов оптимизации: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{h})$.
2. Положим $\hat{\nabla}\mathbf{h} = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h})$.
3. Положим $d\mathbf{v} = \mathbf{0}$.
4. Для $\tau = \eta \dots 1$ повторить:
5. $\boldsymbol{\theta}^{\tau-1} = \boldsymbol{\theta}^{\tau} - \gamma \mathbf{v}^{\tau}$.
6. $\mathbf{v}^{\tau-1} = \mathbf{v}^{\tau} + \gamma \hat{\nabla}_{\boldsymbol{\theta}}$.
7. $d\mathbf{v} = \gamma \hat{\nabla}_{\boldsymbol{\theta}}$.
8. $\hat{\nabla}\mathbf{h} = \hat{\nabla}\mathbf{h} - d\mathbf{v}\nabla_{\mathbf{h}}\nabla_{\boldsymbol{\theta}} Q$.
9. $\hat{\nabla}\boldsymbol{\theta} = \hat{\nabla}\boldsymbol{\theta} - d\mathbf{v}\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} Q$.

Алгоритм RMAD основывается на Reverse-mode differentiation.



DrMAD

Алгоритм DrMad — упрощенный RMAD. Вводится предположение о линейности траектории обновления параметров $\boldsymbol{\theta}$.

1. Провести η шагов оптимизации: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{h})$.
2. Положим $\hat{\nabla}\mathbf{h} = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h})$.
3. Положим $d\mathbf{v} = \mathbf{0}$.
4. Для $\tau = \eta \dots 1$ повторить:
5. $\boldsymbol{\theta}^{\tau-1} = \boldsymbol{\theta}^{\tau} - \gamma \mathbf{v}^{\tau}$.
6. $\mathbf{v}^{\tau-1} = \mathbf{v}^{\tau} + \gamma \hat{\nabla}_{\boldsymbol{\theta}}$.
7. $d\mathbf{v} = \gamma \hat{\nabla}_{\boldsymbol{\theta}}$.
8. $\hat{\nabla}\mathbf{h} = \hat{\nabla}\mathbf{h} - d\mathbf{v}\nabla_{\mathbf{h}}\nabla_{\boldsymbol{\theta}} Q$.
9. $\hat{\nabla}\boldsymbol{\theta} = \hat{\nabla}\boldsymbol{\theta} - d\mathbf{v}\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} Q$.

1. Провести η шагов оптимизации: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{h})$.
2. Положим $\hat{\nabla}\mathbf{h} = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h})$.
3. Положим $d\mathbf{v} = \mathbf{0}$.
4. Для $\tau = \eta \dots 1$ повторить:
5. $\boldsymbol{\theta}^{\tau-1} = \boldsymbol{\theta}_0 + \frac{\tau-1}{\eta}\boldsymbol{\theta}^{\eta}$.
6.
7. $d\mathbf{v} = \gamma \hat{\nabla}_{\boldsymbol{\theta}}$.
8. $\hat{\nabla}\mathbf{h} = \hat{\nabla}\mathbf{h} - d\mathbf{v}\nabla_{\mathbf{h}}\nabla_{\boldsymbol{\theta}} Q$.
9. $\hat{\nabla}\boldsymbol{\theta} = \hat{\nabla}\boldsymbol{\theta} - d\mathbf{v}\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}} Q$.

Закрытая форма оптимизации параметров

Утверждение (Pedregosa, 2016).

Пусть L — дифференцируемая функция, такая что все стационарные точки L являются локальными минимумами. Пусть также гессиан $\mathbf{H}$ функции потерь L является обратимым в каждой стационарной точке.

Тогда

$$\nabla_{\mathbf{A}^{-1}} Q(T(\boldsymbol{\theta}_0), \mathbf{A}^{-1}) = \nabla_{\mathbf{A}^{-1}} Q(\boldsymbol{\theta}^\eta, \mathbf{A}^{-1}) - \nabla_{\mathbf{A}^{-1}} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^\eta, \mathbf{A}^{-1})^{\mathsf{T}} \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^\eta, \mathbf{A}^{-1}).$$

Доказательство

1. Т.к. точка $\boldsymbol{\theta}^\eta$ стационарна, то $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^\eta, \mathbf{A}^{-1}) = 0$.
2. Продифференцируем выражение по $\mathbf{A}^{-1}$:

$$\nabla_{\mathbf{A}^{-1}} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^\eta, \mathbf{A}^{-1} + \mathbf{H} \nabla_{\mathbf{A}^{-1}} T(\boldsymbol{\theta}_0).$$

3. По цепному правилу:

$$\nabla_{\mathbf{A}^{-1}} Q(T(\boldsymbol{\theta}_0), \mathbf{A}^{-1}) = \nabla_{\mathbf{A}^{-1}} Q(\boldsymbol{\theta}^\eta, \mathbf{A}^{-1}) + \nabla_{\mathbf{A}^{-1}} T(\boldsymbol{\theta}_0)^{\mathsf{T}} \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^\eta, \mathbf{A}^{-1}).$$

4. Подставим в выражение 3 выражение 2 и получим искомое.

Жадная оптимизация гиперпараметров

На каждом шаге оптимизации параметров $\boldsymbol{\theta}$:

$$\mathbf{h}' = \mathbf{h} - \gamma_{\mathbf{h}} \nabla_{\mathbf{h}} Q\big(T(\boldsymbol{\theta}, \mathbf{h}), \mathbf{h}\big) = \mathbf{h} - \gamma_{\mathbf{h}} \nabla_{\mathbf{h}} Q\big(\boldsymbol{\theta} - \gamma \nabla L(\boldsymbol{\theta}, \mathbf{h}), \mathbf{h}\big),$$

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

- Можно рассматривать как упрощение алгоритма RMAD, использующее только один элемент истории обнолвения параметров.
- Является приближением к решению закрытой формы в случае $\mathbf{H} \sim \mathbf{I}$.

HOAG

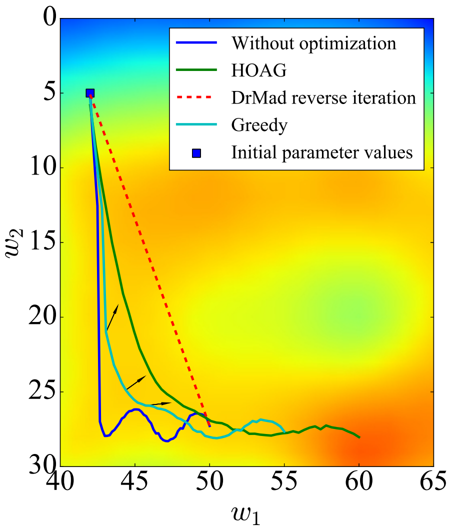
Численное приближение закрытой формы:

$$\nabla_{\mathbf{A}^{-1}} Q(\theta^\eta, \mathbf{A}^{-1}) - \nabla_{\mathbf{A}^{-1}} \nabla_{\boldsymbol{\theta}} L(\theta^\eta, \mathbf{A}^{-1})^{\mathsf{T}} \mathbf{H}^{-1} \nabla_{\boldsymbol{\theta}} Q(\theta^\eta, \mathbf{A}^{-1}).$$

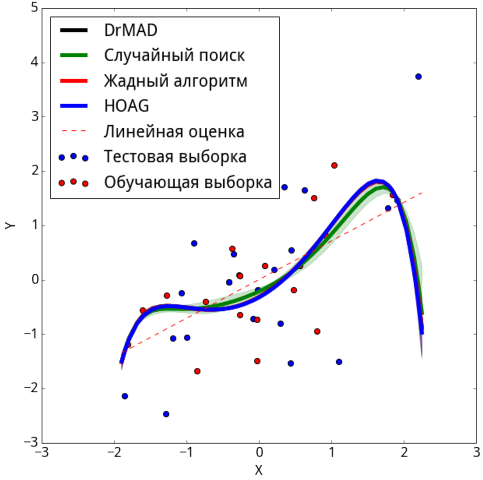
1. Провести η шагов оптимизации: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \mathbf{h})$.
2. Решить линейную систему для вектора $\boldsymbol{\lambda}$: $\mathbf{H}(\boldsymbol{\theta})\boldsymbol{\lambda} = \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \mathbf{h})$.
3. Приближенное значение градиентов гиперпараметра вычисляется как: $\hat{\nabla}_{\mathbf{h}} Q = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h}) - \nabla_{\boldsymbol{\theta}, \mathbf{h}} L(\boldsymbol{\theta}, \mathbf{h})^T \boldsymbol{\lambda}$.

Итоговое правило обновления:

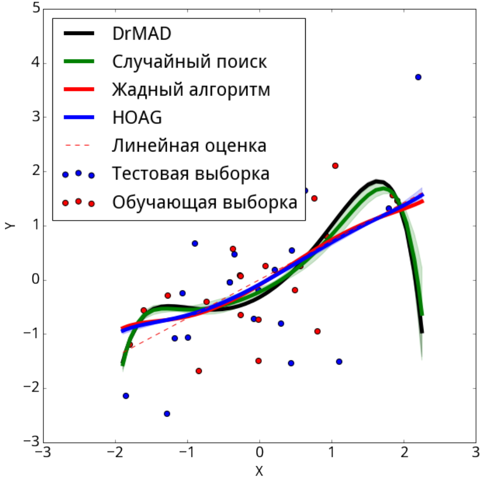
$$\mathbf{h}' = \mathbf{h} - \gamma_{\mathbf{h}} \hat{\nabla}_{\mathbf{h}} Q.$$


Without optimization
HOAG
DrMad reverse iteration
Greedy
Initial parameter values
w_2
w_1
0
5
10
15
20
25
30
40
45
50
55
60
65

Эксперименты: полиномы

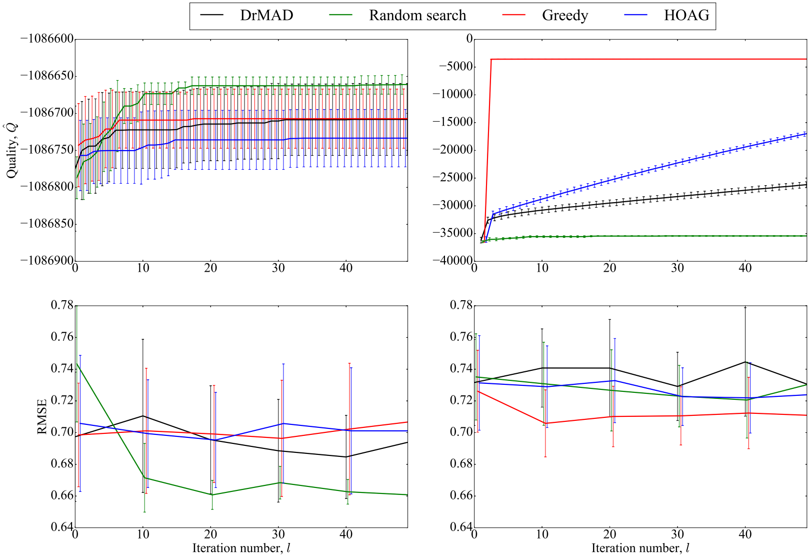


Кросс-валидация

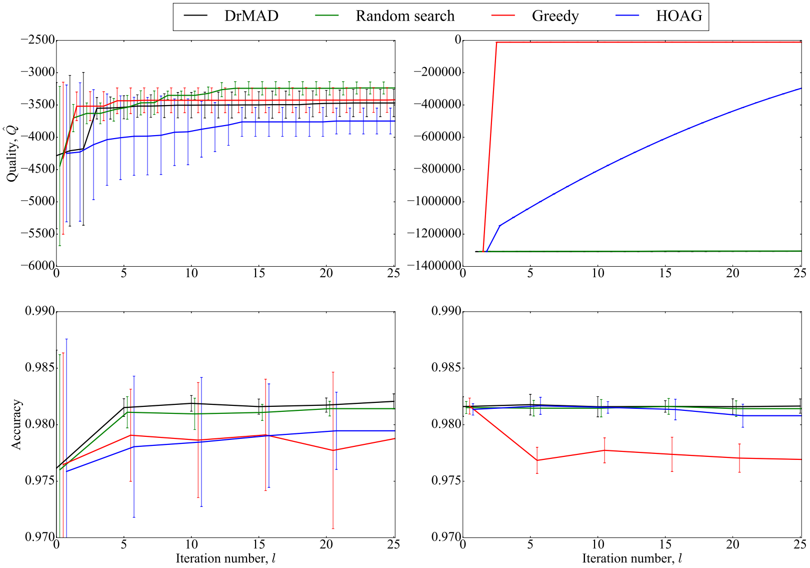


Evidence

Эксперименты: WISDM

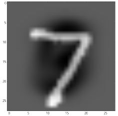


Эксперименты: MNIST

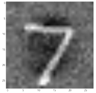


Эксперименты: MNIST

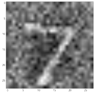
Добавление гауссового шума $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$:



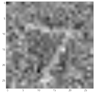
Без шума



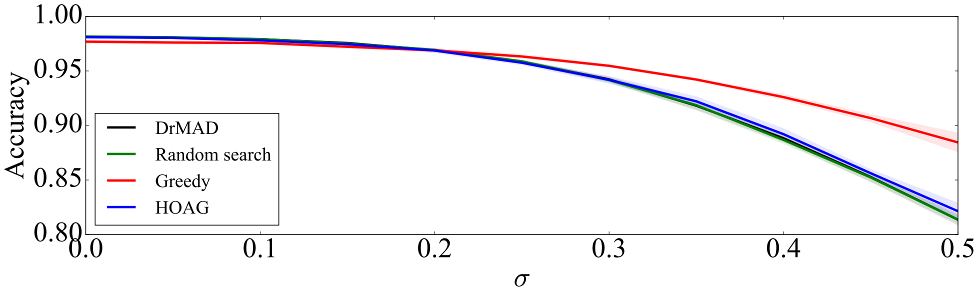
$\sigma = 0.1$



$\sigma = 0.25$



$\sigma = 0.5$



Используемые материалы

1. David J. C. MacKay, Information Theory, Inference & Learning Algorithms, 2003
2. Christopher Bishop, Pattern Recognition and Machine Learning, 2006
3. Dougal Maclaurin et. al, Gradient-based Hyperparameter Optimization through Reversible Learning, 2015
4. Jelena Luketina et. al, Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters, 2016
5. Jie Fu et. al, DrMAD: Distilling Reverse-Mode Automatic Differentiation for Optimizing Hyperparameters of Deep Neural Networks, 2016
6. Fabian Pedregosa, Hyperparameter optimization with approximate gradient, 2016
7. Bobak Shahriari et. al, Taking the Human Out of the Loop: A Review of Bayesian Optimization, 2016