Байесовское мультимоделирование: оптимизация гиперпараметров

Московский Физико-Технический Институт

2021

Выбор модели: связанный байесовский вывод

Первый уровень: выбираем оптимальные параметры:

$$\mathbf{w} = \arg\max \frac{p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathfrak{D}|\mathbf{h})},$$

Второй уровень: выбираем модель, доставляющую максимум обоснованности модели.
Обоснованность модели ("Evidence"):

$$p(\mathfrak{D}|\mathbf{h}) = \int_{\mathbf{w}} p(\mathfrak{D}|\mathbf{w})p(\mathbf{w}|\mathbf{h})d\mathbf{w}.$$

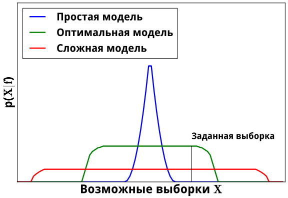
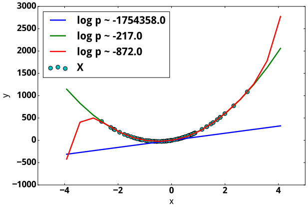


Схема выбора модели



Пример: полиномы

Что такое гиперпараметры

Определение

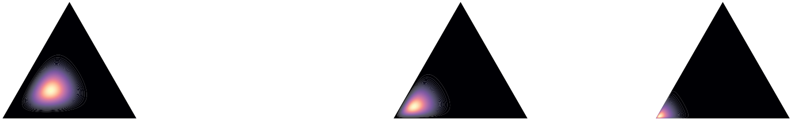
Априорным распределением $p(\boldsymbol{w}|\boldsymbol{h})$ параметров модели назовем вероятностное распределение, соответствующее предположениям о распределении параметров модели.

Определение

Гиперпараметрами $\boldsymbol{h} \in \mathbb{H}$ модели назовем параметры априорного распределения (параметры распределения параметров модели).

Дискретные распределения: релаксация

$\bar{\boldsymbol{\alpha}} = [1, 1, 1], t = 0.9$ $t = 1.0$ $t = 10.0$



$\bar{\boldsymbol{\alpha}} = [0.5, 0.25, 0.25], t = 30.0$ $[0.75, 0.125, 0.125]$ $[0.9, 0.05, 0.05]$

Аппроксимация Лапласа

Нелинейный случай с m объектами и n признаками: $\mathbf{y} \sim \mathcal{N}(\mathbf{f}(\mathbf{X}, \mathbf{w}), \beta^{-1}), \mathbf{w} \sim \mathcal{N}(0, \mathbf{A}^{-1})$.
Запишем интеграл:

$$p(\mathfrak{D}|\mathbf{h}) = p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \beta) = \frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} \int_{\mathbf{w}} \exp(-S(\mathbf{w}))d\mathbf{w}.$$

Разложим S в ряд Тейлора:

$$S(\mathbf{w}) \approx S(\hat{\mathbf{w}}) + \frac{1}{2}\Delta \mathbf{w}^{\mathsf{T}} \mathbf{H} \Delta \mathbf{w}$$

Интеграл приводится к виду:

$$\frac{\sqrt{\beta \cdot |\mathbf{A}|}}{\sqrt{(2\pi)^{m+n}}} S(\hat{\mathbf{w}}) \int_{\mathbf{w}} \exp(-\frac{1}{2}\Delta \mathbf{w}^{\mathsf{T}} \mathbf{H} \Delta \mathbf{w}) d\mathbf{w}$$

Выражение под интегралом соответствует плотности ненормированного нормального распределения.

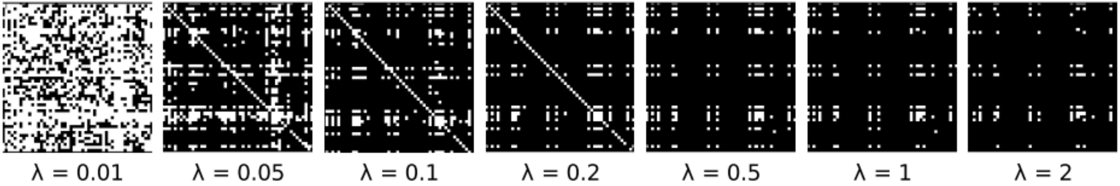
Graves, 2011

Априорное распределение: $p(\mathbf{w}|\sigma) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma \mathbf{I})$.
Вариационное распределение: $q(\mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \sigma_q \mathbf{I})$.
Жадная оптимизация гиперпараметров:

$$\boldsymbol{\mu} = \hat{\mathsf{E}}\mathbf{w}, \quad \sigma = \hat{\mathsf{D}}\mathbf{w}.$$

Прунинг параметра w_i определяется относительной плотностью:

$$\lambda = \frac{q(0)}{q(\boldsymbol{\mu}_{i,q})} = \exp(-\frac{\mu_i^2}{2\sigma_i^2}).$$



Постановка задачи

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

$$\boldsymbol{h}^* = \arg\max_{\boldsymbol{h} \in \mathbb{H}} Q(\boldsymbol{\theta}^*(\boldsymbol{h}), \boldsymbol{h}),$$

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

Байесовский вывод

Пусть $\boldsymbol{\theta} = [\mathbf{w}]^{\mathsf{T}}$.
Первый уровень:

$$\boldsymbol{\theta}^* = \arg\max\bigl(-L(\boldsymbol{\theta}, \mathbf{h})\bigr) = p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \mathbf{h}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{h})}{p(\mathbf{y}|\mathbf{X}, \mathbf{h})}.$$

Второй уровень:

$$p(\mathbf{h}|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{X}, \mathbf{h})p(\mathbf{h}),$$

Полагая распределение параметров $p(\mathbf{h})$ равномерным на некоторой большой окрестности, получим задачу оптимизации гиперпараметров:

$$Q(\boldsymbol{\theta}, \mathbf{h}) = p(\mathbf{y}|\mathbf{X}, \mathbf{h}) = \int_{\mathbf{w} \in \mathbb{R}^u} p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{h}) \to \max_{\mathbf{h} \in \mathbb{H}}.$$

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, 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:

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

Пусть Q — функция качества модели:

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

Вариационная нижняя оценка

Пусть $L = -Q$:

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{A}) \geq \sum_{\boldsymbol{x}, y} \log p(y|\boldsymbol{x}, \hat{\boldsymbol{w}}) - D_{\mathrm{KL}}\big(q(\boldsymbol{w})||p(\boldsymbol{w}|\boldsymbol{A})\big) = -L(\boldsymbol{\theta}, \boldsymbol{A}^{-1}) = Q(\boldsymbol{\theta}, \boldsymbol{A}^{-1}),$$

где q — нормальное распределение с диагональной матрицей коварияций:

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

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

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

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

Evidence vs Кросс-валидация

Оценка Evidece:

$$\log\, p(\mathfrak{D}|\boldsymbol{f}) = \log\, p(\mathfrak{D}_1|\boldsymbol{f}) + \log\, p(\mathfrak{D}_2|\mathfrak{D}_1, \boldsymbol{f}) + \cdots + \log\, p(\mathfrak{D}_n|\mathfrak{D}_1, \ldots, \mathfrak{D}_{n-1}, \boldsymbol{f}).$$

Оценка leave-one-out:

$$\mathrm{LOU} = \mathrm{E}\log\, p(\mathfrak{D}_n|\mathfrak{D}_1, \ldots, \mathfrak{D}_{n-1}, \boldsymbol{f}).$$

Кросс-валидация использует среднее значение последнего члена $p(\mathfrak{D}_n|\mathfrak{D}_1, \ldots, \mathfrak{D}_{n-1}, \boldsymbol{f})$ для оценки сложности.
Evidence учитывает **полную** сложность описания заданной выборки, определяющую предсказательную способность модели с самого начала.

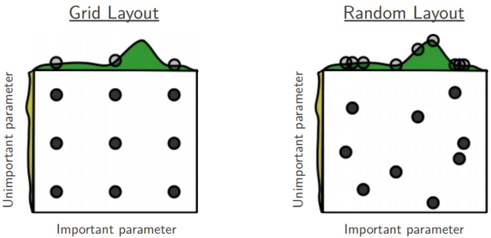
Базовые методы оптимизации гиперпараметров

Варианты:

- Поиск по решетке;
- Случайный поиск.

Оба метода страдают от проклятия размерности.

Случайный поиск может быть более эффективным, если пространство гиперпараметров вырождено.



Bergstra et al., 2012

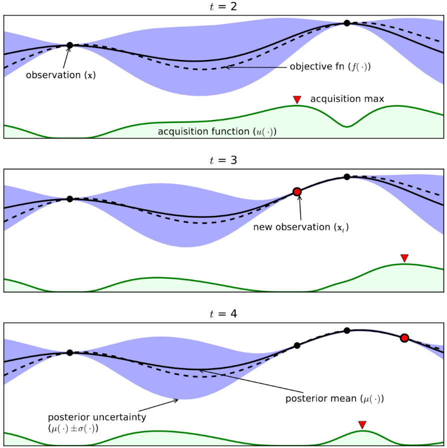
Гауссовый процесс

Идея:
Будем моделировать $Q(\boldsymbol{\theta}(\boldsymbol{h})^*, \boldsymbol{h})$ гауссовым процессом, зависящим от $\boldsymbol{h}$.

Плюсы:

- Гибкость модели.
- Дешевле, чем обучения модели.

Минусы: кубическая сложность по количеству гиперпараметров.



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

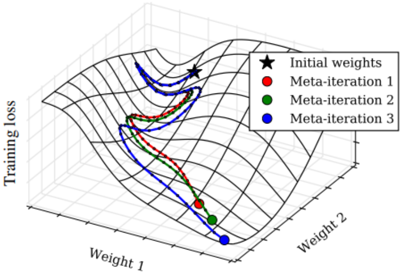
Градиентные методы

Идея: Будем производить оптимизацию вдоль всей траектории оптимизации параметров.

Плюсы:

- Оптимизация гиперпараметров будет учитывать оптимизацию параметров.
- Сложность меняется незначительно от количества гиперпараметров.

Минусы: вычилсительно дорого.



Maclaurin et. al, 2015. Пример работы.

Формальная постановка задачи: градиентная оптимизация

Определение

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

$$\hat{\boldsymbol{\theta}} = T \circ T \circ \cdots \circ T(\boldsymbol{\theta}_0, \boldsymbol{h}) = T^\eta(\boldsymbol{\theta}_0, \boldsymbol{h}), \tag{2}$$

где

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

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:

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

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

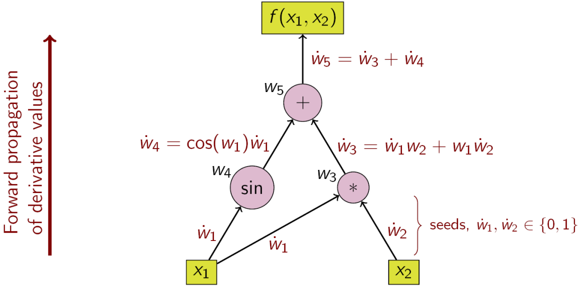
Forward-mode differentiation

Идея дифференцирования: применение формулы:

$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial w_{n-1}} \frac{\partial w_{n-1}}{\partial x} = \frac{\partial y}{\partial w_{n-1}} \left(\frac{\partial w_{n-1}}{\partial w_{n-2}} \frac{\partial w_{n-2}}{x} \partial x \right) = \dots .$$

Пример (wiki):

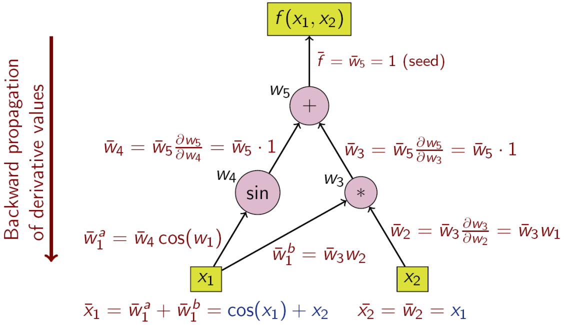
$$x_1 x_2 + \sin(x_1)$$



Reverse-mode differentiation

Идея дифференцирования: применение формулы:

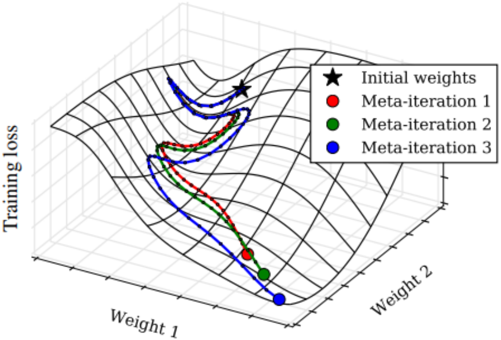
$$\frac{\partial y}{\partial x} = \frac{\partial y}{\partial w_1}\frac{\partial w_1}{\partial x} = \left(\frac{\partial y}{\partial w_2}\frac{\partial w_2}{\partial w_1}\right)\frac{\partial w_1}{\partial x} = \dots.$$



RMAD, Maclaurin et. al, 2015

Алгоритм RMAD основывается на Reverse-mode differentiation.

1. Провести η шагов оптимизации с моментом γ: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \boldsymbol{h})$.
2. Положим $\hat{\nabla}\boldsymbol{h} = \nabla_{\boldsymbol{h}} Q(\boldsymbol{\theta}, \boldsymbol{h})$.
3. Положим $d\boldsymbol{v} = 0$.
4. Для $\tau = \eta \dots 1$ повторить:
5. Вычислить $\boldsymbol{\theta}^{\tau-1}$.
6. Вычислить градиент на шаге $\tau - 1$, используя RMD.



DrMAD

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

1. Провести η шагов оптимизации с моментом γ: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \boldsymbol{h})$.
2. Положим $\hat{\nabla} \boldsymbol{h} = \nabla_{\boldsymbol{h}} Q(\boldsymbol{\theta}, \boldsymbol{h})$.
3. Положим $d\boldsymbol{v} = 0$.
4. Для $\tau = \eta \dots 1$ повторить:
5. Вычислить $\boldsymbol{\theta}^{\tau-1}$.
6. Вычислить градиент на шаге $\tau - 1$, используя RMD.

1. Провести η шагов оптимизации с моментом γ: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \boldsymbol{h})$.
2. Положим $\hat{\nabla} \boldsymbol{h} = \nabla_{\boldsymbol{h}} Q(\boldsymbol{\theta}, \boldsymbol{h})$.
3. Положим $d\boldsymbol{v} = 0$.
4. Для $\tau = \eta \dots 1$ повторить:
5. $\boldsymbol{\theta}^{\tau-1} = \boldsymbol{\theta}_0 + \frac{\tau-1}{\eta}\boldsymbol{\theta}^{\eta}$.
6. Вычислить градиент на шаге $\tau - 1$, используя RMD.

Аналитическая формула оптимизации параметров

Утверждение (Pedregosa, 2016)

Пусть L — дифференцируемая функция, такая что все стационарные точки L являются глобальными минимумами. Пусть также гессиан $\boldsymbol{H}^{-1}$ функции потерь L является обратимым в каждой стационарной точке.
Тогда

$$\nabla_{\boldsymbol{h}} Q(T(\theta_0, \boldsymbol{h}), \boldsymbol{h}) = \nabla_{\boldsymbol{h}} Q(\boldsymbol{\theta}^{\eta}, \boldsymbol{h}) - \nabla_{\boldsymbol{h}} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^{\eta}, \boldsymbol{h})^{\mathsf{T}} \boldsymbol{H}^{-1} \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}^{\eta}, \boldsymbol{h}).$$

Жадная оптимизация гиперпараметров

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

$$\boldsymbol{h}' = \boldsymbol{h} - \beta_{\boldsymbol{h}} \nabla_{\boldsymbol{h}} Q\big(T(\boldsymbol{\theta}, \boldsymbol{h}), \boldsymbol{h}\big) = \boldsymbol{h} - \beta_{\boldsymbol{h}} \nabla_{\boldsymbol{h}} Q\big(\boldsymbol{\theta} - \beta \nabla L(\boldsymbol{\theta}, \boldsymbol{h}), \boldsymbol{h})\big),$$

где $\beta_{\boldsymbol{h}}$ — длина шага оптимизации гиперпараметров.

- Можно рассматривать как упрощение алгоритма RMAD, использующее только один элемент истории обновления параметров.
- Является приближением к решению аналитической формуле в случае $\boldsymbol{H}^{-1} \sim \boldsymbol{I}$.
- Сложность: $O(|\boldsymbol{\theta}| \cdot |\boldsymbol{h}|)$
- Можно упростить, используя формулу конечных приращений, см. DARTS, $O(|\boldsymbol{\theta}| + |\boldsymbol{h}|)$

HOAG

Численное приближение аналитической формулы:

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

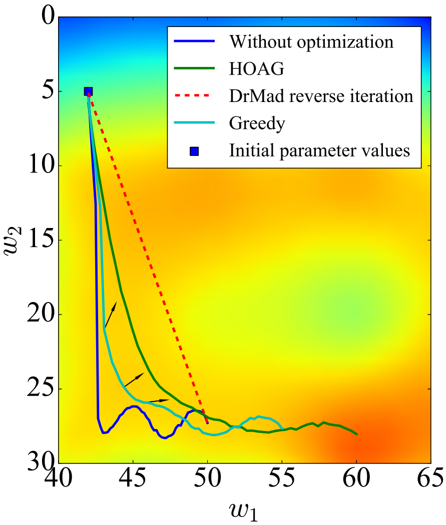
1. Провести η шагов оптимизации: $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0, \boldsymbol{h})$.
2. Решить линейную систему для вектора $\boldsymbol{\lambda}$: $\boldsymbol{H}(\theta)\boldsymbol{\lambda} = \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{h})$.
3. Приближенное значение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

Итоговое правило обновления:

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

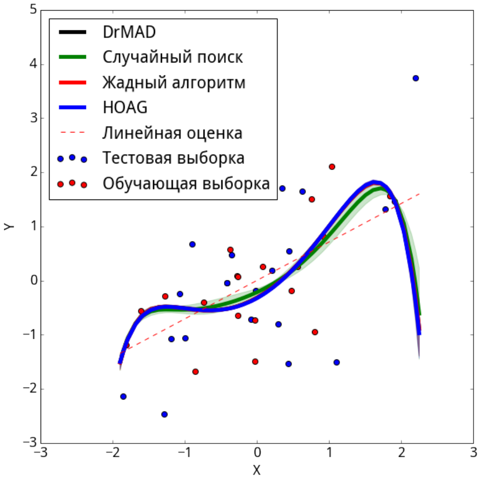
Сравнение алгоритмов

Алгоритм	+	-
Random search	Легко реализовать	Проклятие размерности
Жадная оптимизация	Оптимизация проводится внутри цикла оптимизации параметров. Легко реализовать	Жадность, неоптимальность.
HOAG	Быстрая сходимость.	Качество результатов зависит от решения линейного уравнения $\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{\lambda} = \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{h})$.
DrMAD	Учитывает особенности оператора оптимизации. Можно использовать для оптимизации мета-параметров.	Неустойчив при больших значениях длины градиентного шага $\gamma_{\boldsymbol{h}}$. Качество оптимизации зависит от кривизны траектории обновления параметров.

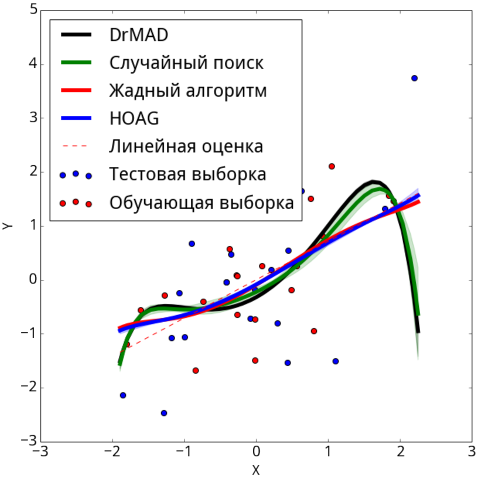


Without optimization
HOAG
DrMad reverse iteration
Greedy
Initial parameter values
w_2
w_1
0
5
10
15
20
25
30
40
45
50
55
60
65

Эксперименты: полиномы

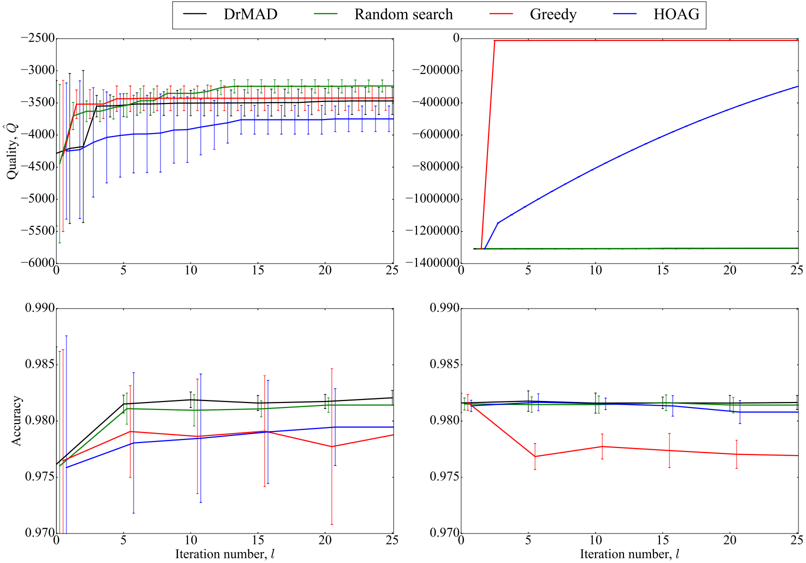


Кросс-валидация



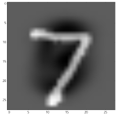
Evidence

Эксперименты: MNIST

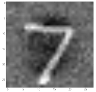


Эксперименты: MNIST

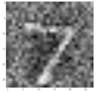
Добавление гауссового шума $\mathcal{N}(0, \sigma^2 \boldsymbol{I})$:



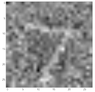
Без шума



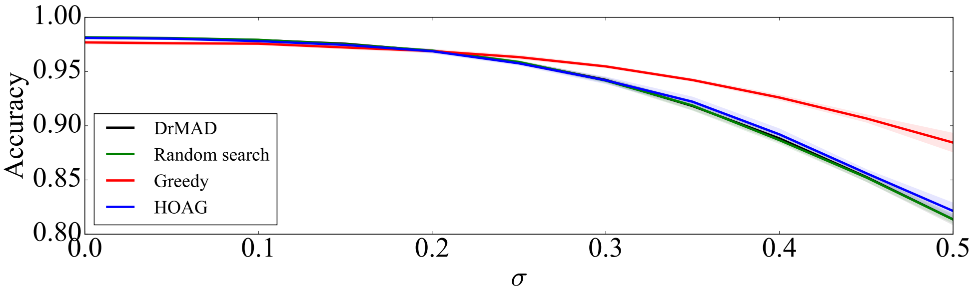
$\sigma = 0.1$



$\sigma = 0.25$



$\sigma = 0.5$



Литература и прочие ресурсы

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – T. 128. – №. 9.
- Bakhteev O. Y., Strijov V. V. Comprehensive analysis of gradient-based hyperparameter optimization algorithms //Annals of Operations Research. – 2020. – T. 289. – №. 1. – C. 51-65.
- Graves A. Practical variational inference for neural networks //Advances in neural information processing systems. – 2011. – T. 24.
- Bergstra et al., Random Search for Hyper-Parameter Optimization, 2012
- Dougal Maclaurin et. al, Gradient-based Hyperparameter Optimization through Reversible Learning, 2015
- Jelena Luketina et. al, Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters, 2016
- Jie Fu et. al, DrMAD: Distilling Reverse-Mode Automatic Differentiation for Optimizing Hyperparameters of Deep Neural Networks, 2016
- Fabian Pedregosa, Hyperparameter optimization with approximate gradient, 2016
- Bobak Shahriari et. al, Taking the Human Out of the Loop: A Review of Bayesian Optimization, 2016
- Liu H., Simonyan K., Yang Y. Darts: Differentiable architecture search //arXiv preprint arXiv:1806.09055. – 2018.