

# Оптимизация гиперпараметров моделей глубокого обучения градиентными методами

О. Ю. Бахтеев<sup>1</sup>, В. В. Стрижов<sup>2</sup>

**Аннотация:** Решается задача оптимизации гиперпараметров модели глубокого обучения. Для оптимизации гиперпараметров модели предлагаются алгоритмы, основанные на градиентном спуске. Так как сложность рассматриваемых алгоритмах сопоставима со сложностью оптимизации параметров модели, предлагается проводить оптимизацию параметров и гиперпараметров в единой процедуре. Для выбора адекватных значений гиперпараметров вводятся вероятностные предположения о распределении параметров. В качестве оптимизируемой функции выступает байесовское правдоподобие модели и кросс-валидация. Для получения оценки правдоподобия используются вариационные методы. Проводится вычислительный эксперимент на нескольких выборках.

**Ключевые слова:** градиентный спуск, стохастический поиск гиперпараметров, оценка гиперпараметров, выбор модели, глубокое обучение, классификация, регрессия.

## 1 Введение

В работе решается задача оптимизации гиперпараметров моделей глубокого обучения. Под *моделью* понимается суперпозиция функций, решающая задачу классификации или регрессии. Под *гиперпараметрами* модели понимается параметры распределения параметров модели.

Одна из проблем построения моделей глубокого обучения — большое число параметров модели [1], которое достигает нескольких миллионов, а оптимизация модели достигает десятков дней [2]. Задача выбора модели глубокого обучения включает в себя выбор стратегии построения модели, эффективной по вычислительным ресурсам. Проблема оптимизации параметров модели глубокого обучения является вычислительно сложной в силу невыпуклости оптимизируемой

---

<sup>1</sup>Московский физико-технический институт, bakhteev@phystech.edu

<sup>2</sup>Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН, strijov@ccas.ru

функции потерь. Поэтому задача поиска параметров оптимизации является важной, и нахождение оптимальных гиперпараметров сильно влияет на итоговое качество модели [?, 3].

В данной работе сравниваются градиентные методы оптимизации гиперпараметров. Основным достоинством подобных алгоритмов является их возможность одновременной оптимизации значительного количества гиперпараметров. В качестве базового алгоритма выступает выбор гиперпараметров модели с использованием случайного поиска. В работах [4–6] в качестве целевой функции потерь рассматривается потеря на валидационной подвыборке с  $L2$  регуляризацией. В данной работе рассматривается общая задача оптимизации гиперпараметров. Рассматриваемые алгоритмы и целевые функции потерь реализованы и представлены в качестве библиотеки для оптимизации гиперпараметров моделей [?]. Основным теоретическим вкладом данной работы является анализ рассматриваемых алгоритмов оптимизации гиперпараметров при использовании функции потерь общего вида, а также исследование качества и устойчивости итоговых моделей в случае использования кросс-валидации и вариационной оценки правдоподобия. В экспериментальной части в качестве критерия выбора модели выступают вариационная нижняя оценка правдоподобия модели и ошибка на валидационной части выборки. В отличие от [5], где также производится сравнение алгоритмов оптимизации гиперпараметров, в данной работе исследуется поведение алгоритмов на выборках большой мощности, таких как WISDM [7] и MNIST [8]. Численные эксперименты показывают, что при значительном количестве гиперпараметров, сопоставимым с количеством параметров модели, рассматриваемые алгоритмы предпочтительнее стохастических.

**Обзор литературы** В работах [9, 10] предлагаются стратегии выбора гиперпараметров модели, основанные на случайном выборе параметров. Другим методом, представленным в литературе [11, 12], является обучение вероятностных моделей для предсказания гиперпараметров. В работе [13] отмечается, что данный метод нахождения оптимальных гиперпараметров является неэффективным в случае, когда число гиперпараметров велико.

В работах [4–6, 13, 14] предлагаются методы оптимизации гиперпараметров, основанные на градиентных алгоритмах оптимизации: восстанавливается вся история изменения параметров в ходе оптимизации, в качестве функции для оптимизации гиперпараметров рассматривается функция потерь от конечного значения параметров, которое выражается через начальное значение параметров. Данная процедура является неэффективной по памяти, т.к. для хранения всей истории оптимизации параметров требуется большое количество памяти. В работе [6] предлагается жадный вариант градиентной оптимизации гиперпараметров. В работе [13] рассматривается оптимизация параметров с моментом, позволяющая эффективно хранить историю параметров в памяти. В работе [4] предлагается метод, рассматривающий траекторию оптимизации параметров как линейную, что также

Алгоритм	Тип алгоритма	Преимущества алгоритма	Недостатки алгоритма
Случайный поиск	стохастический	простота реализации	алгоритм неэффективен при большом количестве гиперпараметров (проклятие размерности)
Жадный алгоритм [6]	градиентный	Возможность одновременной оптимизации параметров и гиперпараметров	Жадность алгоритма
HOAG [5]	градиентный	Быстрая сходимость	Алгоритм требователен к настройкам параметров
DrMAD [4]	градиентный	Алгоритм учитывает алгоритм оптимизации параметров модели и его параметры	Алгоритм страдает от проблем неустойчивости градиентного спуска (градиентный взрыв и затухание); Алгоритм работает в очень жестких предположениях.

Table 1: Преимущества и недостатки рассматриваемых алгоритмов

позволяет эффективно хранить историю параметров. В работе [5] рассматривается аппроксимация градиента оптимизируемой функции.

Для решения заданной рассматриваемой задачи требуется выбрать критерий выбора модели [15, 16]. В качестве критерия выбора модели в ряде работ [15–19] выступает правдоподобие модели. В работах [17–19] рассматривается проблема выбора модели и оценки гиперпараметров в задачах регрессии. Одним из методов получения приближенного значения интеграла правдоподобия является вариационный метод получения нижней оценки интеграла [16]. В работе [20] рассматривается стохастическая версия вариационного метода. В работе [21] рассматривается взаимосвязь градиентных методов получения вариационной нижней оценки интеграла с методом Монте-Карло. Альтернативным критерием выбора модели является минимальная длина описания [22], являющаяся показателем статистической сложности модели и заданной выборки. В работе [23] рассматривается алгоритм получения вариационной нижней оценки правдоподобия для оптимизации гиперпараметров моделей глубокого обучения, проводится связь между правдоподобием модели и минимальной длиной описания.

Альтернативным методом выбора модели является выбор модели на основе скользящего контроля [17, 24]. Проблемой такого подхода является возможная высокая вычислительная сложность [25, 26]. В работах [27, 28] рассматривается проблема смещения оценок качества модели и гиперпараметров, получаемых при использовании  $k$ -fold метода скользящего контроля, при котором выборка делится на  $k$ -частей с обучением на  $k - 1$  части и валидацией результата на оставшейся части выборки.

## 2 Постановка задачи

Задана выборка

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i = 1, \dots, m, \quad (1)$$

состоящая из множества пар «объект-метка»,

$$\mathbf{x}_i \in \mathbf{X} \subset \mathbb{R}^n, \quad y_i \in \mathbf{y} \subset \mathbb{Y},$$

где  $\mathbf{X}$  — матрица объектов,  $\mathbf{y}$  — вектор меток зависимой переменной  $y$ . Метка  $y$  объекта  $\mathbf{x}$  принадлежит либо конечному множеству:  $y \in \mathbb{Y} = \{1, \dots, Z\}$  в случае задачи классификации, где  $Z$  — число классов, либо некоторому подмножеству вещественных чисел  $y \in \mathbb{R}$  в случае задачи регрессии.

Задана дифференцируемая по параметрам модель, приближающая зависимую переменную  $y$ :

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

Рассмотрим модель  $f$  и ее вероятностную интерпретацию для случая задач регрессии и классификации.

**Регрессия.** Положим, что зависимая переменная распределена нормально:

$$\mathbf{y} = \mathcal{N}(\mathbf{f}, \mathbf{I}), \quad (2)$$

где  $\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X})$ .

Определим правдоподобие выборки  $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ :

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = -\frac{m}{2} \log(2\pi) - \frac{1}{2}(\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X}))^T \mathbf{I}(\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})).$$

**Классификация.** В случае двухклассовой классификации положим, что зависимая переменная распределена биномиально:

$$\mathbf{y} = \mathcal{B}(1 - \mathbf{f}, \mathbf{f}), \quad (3)$$

где вектор-функция  $\mathbf{f}$  задает вероятность принадлежности объектов  $\mathbf{X}$  к первому классу. В случае многоклассовой классификации зависимая переменная распределена мультиномиально,  $r$ -я компонента  $\mathbf{f}$  задает вероятность принадлежности классу  $r$ . Тогда правдоподобие выборки задается как

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \sum_{\mathbf{x}, y \in \mathbf{X}, \mathbf{x}} \sum_{r=1}^Z [y = r] \log f_r(\mathbf{w}, \mathbf{x}),$$

где  $f_r$  —  $r$ -я компонента функции  $\mathbf{f}$ .

Для задач классификации (3) и регрессии (2) задано параметрическое априорное распределение  $p(\mathbf{w}|\mathbf{A})$  вида:

$$\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}), \quad (4)$$

где  $\mathbf{A}^{-1} = \text{diag}[\alpha_1, \dots, \alpha_u]^{-1}$  — матрица ковариаций диагонального вида. Гипотезы (2),(3),(4) не противоречат друг другу в силу неограниченности нормального распределения [29].

Задача оптимизации гиперпараметров зависит как от критерия выбора модели, так и от метода оптимизации параметров модели. Проиллюстрируем задачу оптимизации гиперпараметров *двусвязным байесовским выводом*. Для дальнейшей формализации задачи в общем виде введем переобозначение:

$$\boldsymbol{\theta} = \mathbf{w}, \quad \mathbf{h} = [\alpha_1, \dots, \alpha_u], \quad (5)$$

где  $\boldsymbol{\theta}$  — множество оптимизируемых параметров модели,  $\mathbf{h}$  — множество гиперпараметров модели.

На *первом уровне* байесовского вывода производится оптимизация параметров модели  $f$  по заданной выборке  $\mathcal{D}$ :

$$\hat{\boldsymbol{\theta}} = \arg \max(-L(\boldsymbol{\theta}, \mathbf{h})) = p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \mathbf{A}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{A})}{p(\mathbf{y}|\mathbf{X}, \mathbf{A})}. \quad (6)$$

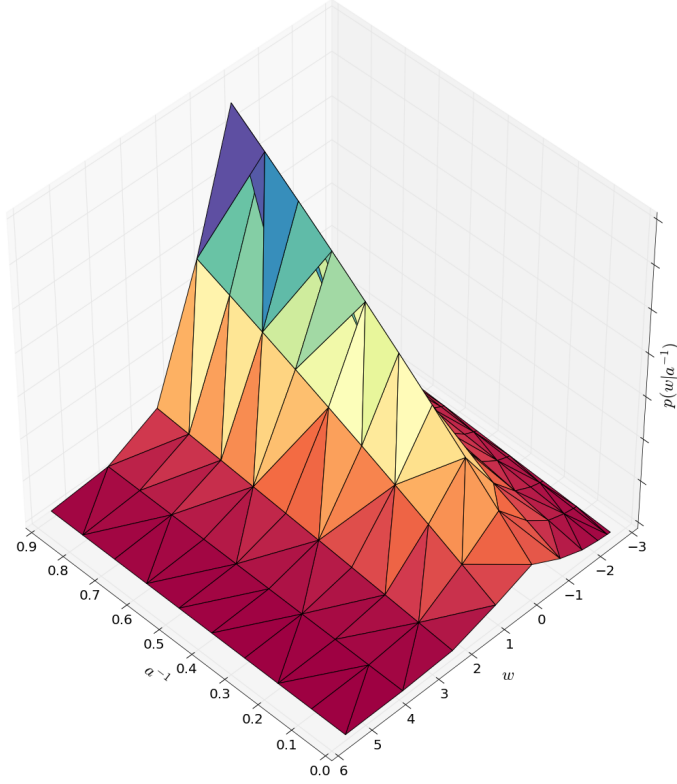


Figure 1: Зависимость правдоподобия модели от значения гиперпараметра  $\alpha$ . TODO: переделать

На *втором уровне* производится оптимизация апостериорного распределения гиперпараметров  $\mathbf{h}$ :

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

где знак « $\propto$ » означает равенство с точностью до нормирующего множителя.

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

$$Q(\boldsymbol{\theta}, \mathbf{h}) = p(\mathbf{y}|\mathbf{X}, \mathbf{A}) = \int_{\mathbf{w} \in \mathbb{R}^u} p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{A}) \rightarrow \max_{[\alpha_1, \dots, \alpha_u] \in \mathbb{R}^n}. \quad (7)$$

Сформулируем задачу оптимизации гиперпараметров в общем виде. Обозначим за  $\mathbf{h} \in \mathbb{R}^h$  вектор гиперпараметров модели (5). Обозначим за  $\boldsymbol{\theta} \in \mathbb{R}^s$  множество всех оптимизируемых параметров (5). Пусть задана дифференцируемая функция потерь  $L(\boldsymbol{\theta}, \mathbf{h})$ , по которой производится оптимизация функции  $f$  (6). Пусть также задана

дифференцируемая функция  $Q(\boldsymbol{\theta}, \mathbf{h})$ , определяющая итоговое качество модели  $f$  и приближающая интеграл (7).

Требуется найти параметры  $\hat{\boldsymbol{\theta}}$  и гиперпараметры  $\hat{\mathbf{h}}$  модели, доставляющие минимум следующему функционалу:

$$\hat{\mathbf{h}} = \arg \max_{\mathbf{h} \in \mathbb{R}^h} Q(\hat{\boldsymbol{\theta}}(\mathbf{h}), \mathbf{h}), \quad (8)$$

$$\hat{\boldsymbol{\theta}}(\mathbf{h}) = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^s} L(\boldsymbol{\theta}, \mathbf{h}). \quad (9)$$

Рассмотрим вид переменной  $\boldsymbol{\theta}$  и функций  $L, Q$  для различных методов выбора модели и оптимизации ее параметров.

**Базовый метод** Пусть оптимизация параметров и гиперпараметров производится по всей выборке  $\mathfrak{D}$  по одной и той же функции:

$$L(\boldsymbol{\theta}, \mathbf{h}) = Q(\boldsymbol{\theta}) = \log p(\mathbf{y}, \mathbf{w} | \mathbf{X}, \mathbf{A}) = \log p(\mathbf{y} | \mathbf{X}, \mathbf{w}) + \log p(\mathbf{w} | \mathbf{A})$$

Вспомогательная переменная  $\boldsymbol{\theta}$ , по которой производится оптимизация модели  $f$ , соответствует параметрам модели:

$$\boldsymbol{\theta} = \mathbf{w}.$$

**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  $k$ .

Положим функцию  $L$  равной среднему значению минус логарифма апостериорной вероятности по всем  $k - 1$  разбиениям  $\mathfrak{D}$ :

$$L(\boldsymbol{\theta}, \mathbf{h}) = -\frac{1}{k} \sum_{q=1}^k \left( \frac{k}{k-1} \log p(\mathbf{y} \setminus \mathbf{y}_q | \mathbf{X} \setminus \mathbf{X}_q, \mathbf{w}_q) + \log p(\mathbf{w}_q | \mathbf{A}) \right). \quad (10)$$

Положим функцию  $Q$  равной среднему значению правдоподобия выборки по частям выборки  $\mathfrak{D}_q$ , на которых не проходила оптимизация параметров:

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

**Вариационная оценка правдоподобия** Положим  $L = -Q$ , равной вариационной оценке правдоподобия модели:

$$\begin{aligned} \log p(\mathbf{y}|\mathbf{X}, \mathbf{A}) &\geq -D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{A})) + \int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{y}|\mathbf{X}, \mathbf{w}, \mathbf{A}) d\mathbf{w} \approx \\ &\approx \sum_{i=1}^m \log p(y_i|\mathbf{x}_i, \mathbf{w}_i) - D_{\text{KL}}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{A})) = -L(\boldsymbol{\theta}, \mathbf{h}) = Q(\boldsymbol{\theta}), \end{aligned} \quad (11)$$

где  $q$  — нормальное распределение с диагональной матрицей ковариаций:

$$q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q^{-1}), \quad (12)$$

где  $\mathbf{A}_q = \text{diag}[\alpha_1^q, \dots, \alpha_u^q]^{-1}$  — диагональная матрица ковариаций,  $\boldsymbol{\mu}_q$  — вектор средних компонент. Расстояние  $D_{\text{KL}}$  между двумя гауссовыми величинами задается как

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

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

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

### 3 Градиентные методы оптимизации гиперпараметров

Рассмотрим случай, когда оптимизация (9) параметров  $\boldsymbol{\theta}$  производится с использованием градиентных методов.

**Определение.** Назовем оператором оптимизации алгоритм  $T$  выбора вектора параметров  $\boldsymbol{\theta}'$  по параметрам предыдущего шага  $\boldsymbol{\theta}$ :

$$\boldsymbol{\theta}' = T(\boldsymbol{\theta}, \mathbf{h}).$$

Рассмотрим оператор градиентного спуска, производящий  $\eta$  шагов оптимизации:

$$\hat{\boldsymbol{\theta}} = T \circ T \circ \dots \circ T(\boldsymbol{\theta}_0, \mathbf{h}) = T^\eta(\boldsymbol{\theta}_0, \mathbf{h}), \quad (13)$$

где

$$T(\boldsymbol{\theta}, \mathbf{h}) = \boldsymbol{\theta} - \gamma \nabla L(\boldsymbol{\theta}, \mathbf{h}),$$

$\gamma$  — длина шага градиентного спуска,  $\boldsymbol{\theta}_0$  — начальное значение параметров  $\boldsymbol{\theta}$ . В данной работе в качестве оператора оптимизации параметров модели выступает стохастический градиентный спуск:

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



где  $\hat{\mathfrak{D}}$  — случайная подвыборка исходной выборки  $\mathfrak{D}$ .

Перепишем задачу оптимизации (8), (9) в следующем виде

$$\hat{\mathbf{h}} = \arg \max_{\mathbf{h} \in \mathbb{R}^h} Q(T^\eta(\boldsymbol{\theta}_0, \mathbf{h})), \quad (14)$$

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

Оптимизационную задачу (14) предлагается решать с использованием градиентного спуска. Вычисление градиента от функции  $Q(T^\eta(\boldsymbol{\theta}_0, \mathbf{h}))$  по гиперпараметрам  $\mathbf{h}$  является вычислительно сложным в силу внутренней процедуры оптимизации  $T(\boldsymbol{\theta}_0, \mathbf{h})$ . Общая схема оптимизации гиперпараметров представлена следующим образом:

1. От 1 до  $l$ :
2. Инициализировать параметры  $\boldsymbol{\theta}$  при условии гиперпараметров  $\mathbf{h}$ .
3. Приблизительно решить задачу оптимизации (14) и получить новый вектор параметров  $\mathbf{h}'$
4.  $\mathbf{h} = \mathbf{h}'$ .

где  $l$  — количество итераций оптимизации гиперпараметров. Рассмотрим методы приближенного решения данной задачи оптимизации.

**Жадный алгоритм** В качестве правила обновления вектора гиперпараметров  $\mathbf{h}$  на каждом шаге оптимизации (13) выступает градиентный спуск с учетом обновления параметров  $\boldsymbol{\theta}$  на данном шаге:

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

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

**Алгоритм НОАГ** Предлагается получить приближенные значения градиента гиперпараметров  $\nabla_{\mathbf{h}} Q(T^\eta(\boldsymbol{\theta}_0, \mathbf{h}))$  на основе следующей формулы:

$$\nabla_{\mathbf{h}} Q(T^\eta(\boldsymbol{\theta}_0, \mathbf{h})) = \nabla_{\mathbf{h}} Q(\boldsymbol{\theta}, \mathbf{h}) - (\nabla_{\boldsymbol{\theta}, \mathbf{h}}^2 L(\boldsymbol{\theta}, \mathbf{h}))^T \mathbf{H}(\boldsymbol{\theta})^{-1} \nabla_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \mathbf{h}),$$

где  $\mathbf{H}$  — гессиан функции  $L$  по параметрам  $\boldsymbol{\theta}$ .

Процедура получения приближенного значения градиента гиперпараметров  $\nabla_{\mathbf{h}} Q$  производится итеративно:

1. Провести  $\eta$  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:

$$\mathbf{h}' = \mathbf{h} - \gamma_{\mathbf{h}} \hat{\nabla}_{\mathbf{h}} Q. \quad (15)$$

В данной работе для приближенного решения шага 2 алгоритма НОАГ используется стохастический градиентный спуск в силу сложности вычисления гессиана  $\mathbf{H}(\boldsymbol{\theta})$ .

**Алгоритм DrMad** Для получения градиента от оптимизируемой функции  $Q$  как от функции от начальных параметров  $\boldsymbol{\theta}_0$  предлагается пошагово восстановить  $\eta$  шагов оптимизации  $T(\boldsymbol{\theta}_0)$  в обратном порядке аналогично методу обратного распространения ошибок. Для упрощения данной процедуры вводится предположение, что траектория изменения параметров  $\boldsymbol{\theta}$  линейна:

$$\boldsymbol{\theta}^\tau = \boldsymbol{\theta}_0 + \frac{\tau}{\eta} T(\boldsymbol{\theta}). \quad (16)$$

Алгоритм вычисления приближенного значения градиента  $\nabla \mathbf{h}$  является частным случаем алгоритма обратного распространения ошибки и представим в следующем виде:

1. Провести  $\eta$  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. Вычислить значения параметров  $\boldsymbol{\theta}^\tau$  (16).
6.  $d\mathbf{v} = \gamma \hat{\nabla}_{\boldsymbol{\theta}}$ .
7.  $\hat{\nabla} \mathbf{h} = \hat{\nabla} \mathbf{h} - d\mathbf{v} \nabla_{\mathbf{h}} \nabla_{\boldsymbol{\theta}} Q$ .
8.  $\hat{\nabla} \boldsymbol{\theta} = \hat{\nabla} \boldsymbol{\theta} - d\mathbf{v} \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} Q$ .

Итоговое правило обновления гиперпараметров аналогично (15). В работе [4] отмечается неустойчивость алгоритма при высоких значениях длины шага градиентного спуска  $\gamma$ . Поэтому вместо исходного правила (16) в данной работе первые 5% значений параметров не рассматриваются, а также учитывается только каждый  $\tau_k$  шаг оптимизации:

$$\boldsymbol{\theta}^\tau = \boldsymbol{\theta}_{\tau_0} + \frac{\tau}{\eta} T(\boldsymbol{\theta}), \quad \tau \in \{\tau_0, \dots, \eta\}, \tau \bmod \tau_k = 0, \quad (17)$$

где  $\tau_0 = [0.05 \cdot \eta]$ .

## 4 Вычислительный эксперимент

Для анализа рассматриваемых алгоритмов оптимизации гиперпараметров был проведен ряд вычислительных экспериментов на выборках MNIST [8], WISDM [7], а также на синтетических данных.

Рассматривались следующие критерии качества:

1. Наилучшее значение  $\hat{Q} = \max_{j \in \{1, \dots, l\}} Q^j$ .
2. Среднее число итераций алгоритма для сходимости. Под данным показателем понимается число шагов оптимизации гиперпараметров, при котором ошибка  $Q$  изменяется не более чем на 1% от своего наилучшего значения:

$$\arg \min_j : \frac{Q^j - Q^0}{\hat{Q} - Q^0} \geq 0.99,$$

где  $Q^0$  — значение функции  $Q$  до начала оптимизации гиперпараметров.

3. Внешний критерий качества моделей  $E$ :

$$E = \text{RMSE} = \left( \frac{1}{m} \sum_1^m (f(\mathbf{x}_i, \mathbf{w}) - y_i) \right)^{\frac{1}{2}}$$

в случае задачи регрессии,

$$E = \text{Accuracy} = 1 - \frac{1}{m} \sum_1^m [f(\mathbf{x}_i, \mathbf{w}) \neq y_i]$$

в случае задачи классификации.

4. Внешний критерий качества моделей  $E_\sigma$  при возмущении параметров модели:

$$E_\sigma = \text{RMSE}_\sigma = \left( \frac{1}{m} \sum_1^m (f(\mathbf{x}_i, \mathbf{w} + \boldsymbol{\varepsilon}) - y_i) \right)^{\frac{1}{2}}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{I}).$$

В качестве улучшаемого алгоритма рассматривался случайный поиск параметров с количеством итераций поиска, совпадающих с количеством итераций оптимизации гиперпараметров  $l$ :  $l = 50$  для синтетической выборки и выборки WISDM,  $l = 25$  для выборки MNIST. Рассматриваемые алгоритмы представлены в Табл. 2. Пример поведения траекторий параметров под действием алгоритмов приведен на Рис. ???. В качестве функций  $Q$  и  $L$  рассматривались функции кросс-валидации (10) с  $k = 4$  и вариационной оценки правдоподобия (11).

На всех выборках гиперпараметры инициализировались случайно из равномерного распределения:

$$\mathbf{h} \sim \mathcal{U}(a, b)^h,$$

Алгоритм	Тип алгоритма	Сложность работы одной итерации	Предположения для корректности
Случайный поиск	стохастический	$O(\eta s  \hat{\mathcal{D}} )$	-
Жадный алгоритм [6]	градиентный	$O(\eta(s + h)  \hat{\mathcal{D}} )$	$\mathbf{H}(\theta) = \mathbf{I}$
НОAG [5]	градиентный	$O(\eta s  \hat{\mathcal{D}}  + h^2  \hat{\mathcal{D}}  + o)$ , где $o$ — время решения уравнения в пункте 3	первые производные $Q$ и вторые производные $L$ — липшецевы; $\det \mathbf{H} \neq 0$ ;
DrMAD [4]	градиентный	$O(\eta s  \hat{\mathcal{D}} )$	Траектория оптимизации параметров $\theta = \theta_0, \dots, \theta_\eta$ — линейная

Table 2: Основные свойства рассматриваемых алгоритмов

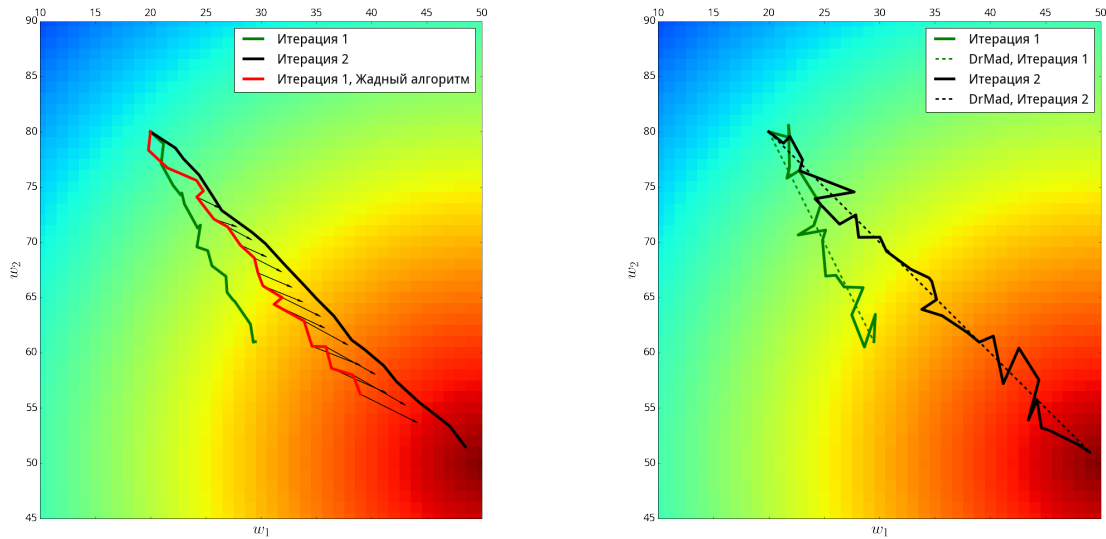


Figure 2: Графики траекторий параметров для разных алгоритмов. TODO: переделать