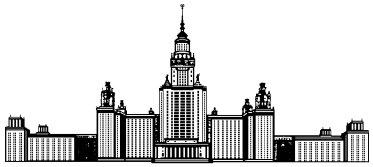
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики

Кафедра Математических Методов Прогнозирования

## КУРСОВАЯ РАБОТА СТУДЕНТА 317 ГРУППЫ

# «Оптимизация гиперпараметров в алгоритмах машинного обучения»

Выполнил:

студент 3 курса 317 группы

*Михеев Борис Михайлович*

Научный руководитель:

д.ф-м.н., профессор

*Дьяконов Александр Геннадьевич*

Москва, 2022

# Содержание

**Аннотация**

В данной работе проводится подробный обзор существующих подходов к оптимизации гиперпараметров в алгоритмах машинного обучения. Проведен сравнительный анализ наиболее распространенных методов, рассмотрены их особенности, преимущества и недостатки, а также проведены вычислительные эксперименты с библиотеками `optuna` и `scikit-optimize` с целью более детального исследования.

# 1 Введение

В настоящее время существует большое количество разнообразных алгоритмов машинного обучения. Они решают различные задачи и применяются во многих сферах человеческой деятельности. Чаще всего они параметризованы двумя наборами параметров: настраиваемыми во время обучения по данным (например, веса линейной модели или нейронной сети), и **гиперпараметрами**, задаваемыми до начала обучения и определяющими сложность и структуру итоговой модели (коэффициент регуляризации, число моделей в ансамбле и т. д.). Наличие гиперпараметров позволяет управлять процессом обучения и адаптировать алгоритм к различным данным и задачам. Однако они не могут быть настроены во время обучения. Возникает потребность в правильном их задании для достижения наилучшего качества и эффективности обучаемой модели.

В некоторых случаях выбор гиперпараметров может осуществляться вручную, по различным эпмирическим правилам или по результатам перебора всевозможных сочетаний их значений, например, с помощью поиска по сетке. Данные подходы достаточно просты, однако они крайне неэффективны в случае большого числа параметров, свойственного многим современным алгоритмам машинного обучения. Приобретает важность задача автоматического определения оптимальных гиперпараметров с целью снижения затрат времени и ресурсов в процессе настройки моделей, повышения их качества и обобщающей способности.

Оптимизация гиперпараметров сталкивается с рядом проблем. Ее цель заключается в нахождении параметров, минимизирующих или максимизирующих определенную метрику, характеризующую качество модели (например, потери или точность на валидации). Для этого потребуется многократное ее вычисление при различных значениях гиперпараметров, что может быть долго и дорого для сложных моделей или больших данных. Она может быть весьма сложно устроена, и как правило она не обладает свойствами гладкости и выпуклости, нет возможности вычислить ее градиент, к ней неприменимы известные градиентные методы оптимизации. Также гиперпараметров может быть много, и они могут быть различных типов, иметь раз-

личные диапазоны значений, зависеть друг от друга и в разной степени влиять на итоговое качество.

В настоящее время существует множество подходов к оптимизации гиперпараметров. В данной работе рассматриваются наиболее популярные из них, проводится их сравнительный анализ, а также проводится исследование реализаций некоторых методов применительно к различным моделям машинного обучения и типам данных.

# 2 Постановка задачи

Пусть $\mathcal{A}$ - алгоритм машинного обучения с $N$ гиперпараметрами. $\Lambda_n$ - область значений $n$-ого их них. $\Lambda = \Lambda_1 \times ... \times \Lambda_N$ - пространство конфигураций гиперпараметров, $\lambda \in \Lambda$ - вектор их значений . $D_{train}, D_{valid}$ - обучающие и валидационные данные из некоторого распределения $\mathcal{D}$. $\mathcal{L}(D_{valid}, \mathcal{A})$ - функция потерь на валидационных данных. $\mathcal{F}(\lambda, \mathcal{A}, D_{train}, D_{valid}, \mathcal{L})$ - целевая функция, измеряющая потери модели, полученной алгоритмом $\mathcal{A}$ с гиперпараметрами $\lambda$ при обучении на данных $D_{train}$ и оцененной на валидационных данных $D_{valid}$.

Цель поиска гиперпараметров - найти оптимальный вектор их значений

$$\lambda^* = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \mathbb{E}_{(D_{train}, D_{valid}) \sim \mathcal{D}} \mathcal{F}(\lambda, \mathcal{A}, D_{train}, D_{valid}, \mathcal{L})$$

Можно рассмотреть и иную постановку, где целевой функцией является, например, точность модели, и ее необходимо максимизировать. В любом случае мы имеем оптимизационную задачу. На практике есть потребность не только минимизировать потери или максимизировать качество, но и сократить при этом время процесса поиска гиперпараметров, количество вычислительных операций и т. д., и в общем случае оптимизация гиперпараметров является многокритериальной задачей[1]. Сосредоточимся на базовой постановке с одним оптимизируемым критерием.

Данная задача имеет особенности и сталкивается с рядом проблем:

- Пространство поиска $\Lambda$ может быть достаточно сложным. Гиперпараметры могут быть вещественными (темп обучения, коэффициент регуляризации), целочисленными (число базовых алгоритмов ансамбля, число слоев сети) или категориальными (вид препроцессинга, тип ядра для SVM). Между ними также могут существовать зависимости, например, если число слоев нейронной сети меньше $n$, то параметры $n$-ого слоя не имеют смысла. Помимо всего прочего было показано, что во многих случаях лишь часть гиперпараметров существенно влияет на целевую функцию, и отдельной задачей является определение наиболее важных из них[2]. Не всегда ясно, как задавать диапазон значений параметров.

- Целевая функция обычно имеет сложное устройство, не обладает свойствами гладкости, выпуклости или дифференцируемости, может иметь много локальных оптимумов, что делает применение градиентных методов оптимизации невозможным. Как правило неизвестна ее зависимость от гиперпараметров. По сути все, что с ней можно делать - это вычислять ее в определенных точках, к другой информации доступа нет.

- Вычисление целевой функции требует обучения модели на разных гиперпараметрах и оценки ее на валидационных данных. Этот процесс может занимать очень много времени. Также время обучения и тестирования может существенно зависеть от некоторых гиперпараметров (размер ансамбля, градиентный шаг). Возникает потребность в методах оптимизации гиперпараметров, требующих как можно меньше вычислений целевой функции.

# 3 Обзор подходов к оптимизации гиперпараметров

Существует большое число соответствующих методов, как достаточно простых, так и использующих более сложные идеи. Как уже было отмечено, в общем случае устройство оптимизируемой функции в данной задаче неизвестно, и есть возможность лишь вычислять ее значения на входах. Можно только строить предположения о ней и подбирать оптимальные значения гиперпараметров определенным способом. Такие функции называют «черными ящиками» (black-box functions), и существует широкий спектр методов их оптимизации[3][4]. Предпочтение отдается алгоритмам глобальной оптимизации, не требующим использования производных. Условно их можно разделить на следующие группы:

- **Простейшие методы**, не использующие никакой информации и не делающие предположений о задаче или функции, сводящиеся по сути к простому перебору. Таковыми являются поиск по сетке и случайный поиск.

- **Метаэвристические методы**, использующие различные эвристики для поиска возможных решений задачи, оптимальных или близких к оптимальным. При этом они не используют предположения об оптимизируемой функции, в некотором смысле тоже реализуют случайный поиск, но подчиненный некоторым правилам или концепции. Их основная цель - наиболее эффективно исследовать пространство поиска на предмет близких к оптимуму решений. Они не гарантируют нахождения глобально оптимального решения, однако могут быть применены к широкому кругу задач и показывают неплохие результаты на практике.[5][6][7] Ярким и распространенным примером таких методов являются популяционные алгоритмы[8], в которых поддерживается и итеративно обновляется набор возможных решений. Обновления происходят согласно некоторым правилам с элементами случайности.

- **Последовательные методы, основывающиеся на моделях** (Sequential Model-Based Optimization, SMBO). В них присутствует две основных компоненты. Первая - вероятностная (т. н. суррогатная) модель, которая обучается по известным значениям целевой функции в наборе точек и аппроксимирует ее. Вы-

числения этой модели как правило значительно менее затратны в сравнении с целевой функцией. Вторая - функция выбора (или функция выгоды, acquisition function) для следующей точки. Она использует предсказательное распределение вероятностной модели, оценивает перспективность точек-кандидатов, соблюдая баланс между уточнением рассмотренных регионов конфигурационного пространства и исследованием новых областей. Ярким примером такого подхода является байесовская оптимизация и различные ее модификации.[9][10]

Рассмотрим наиболее популярные методы более детально.

## 3.1 Простейшие методы

### 3.1.1 Поиск по сетке

Поиск по сетке является простым и широко используемым методом подбора гиперпараметров. Пространство поиска в нем задается дискретной сеткой - декартовым произведением конечных наборов значений гиперпараметров, задаваемых пользователем. Алгоритм вычисляет целевую функцию на всех наборах сетки и выдает набор гиперпараметров, соответствующий наилучшему среди найденных значений. Он прост в реализации и может выполняться параллельно, так как точки сетки оцениваются независимо. Однако он крайне неэффективен в случае большого числа гиперпараметров и высокой частоты дискретизации шкал их значений, так как число комбинаций, и, соответственно, вычислений целевой функции будет расти экспоненциально с ростом размерности пространства переменных и точности разбиения сетки. При этом данный алгоритм способен находить достаточно успешные конфигурации в случае низкой размерности конфигурационного пространства[1][2]. Также он не использует информацию о предыдущих успешных результатах. Это может приводить к рассмотрению большого числа неоптимальных конфигураций. Существует его модернизация - contracting grid search.[11] В начале поиск идет по сетке с грубым разбиением, далее строится более дискретизированная сетка меньшего размера с центром в наилучшей найденной комбинации, и уже по этой сетке вновь запускается поиск.

### 3.1.2 Случайный поиск

Данный алгоритм выбирает значения гиперпараметров независимо из заданных равномерных (в простейшем случае) распределений каждого из них. Процесс прекращается по достижении заданного числа итераций или по исчерпании некоторго бюджета (запаса времени, памяти и др.) Он также прост в реализации и распараллеливании, на практике находит более удачные конфигурации, чем поиск по сетке, часто используется в качестве бейзлайна в задачах оптимизации гиперпараметров[2]. Однако данный метод тоже не учитывает историю вычислений и информацию о потенциально оптимальных регионах, аналогично требует многократного вычисления целевой функции.

Стоит отметить, что случайный поиск более эффективен, чем поиск по сетке в конфигурационных пространствах большой размерности. На практике целевая функция часто имеет низкую эффективную размерность, т. е. ее зависимость от некоторого подмножества гиперпараметров значительно сильнее, чем от всех остальных.

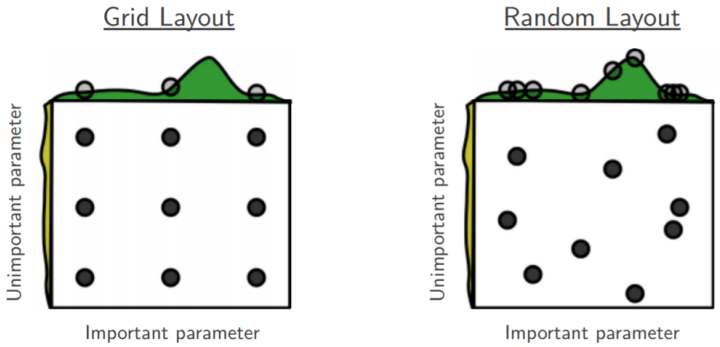



Рис. 1: Исслюстрация из статьи[2], демонстрирующая различия между поиском по сетке и случайным поиском.

На рис. 1 приведен пример из статьи[2], демонстрирующий покрытие конфигурационного пространства алгоритмами случайного поиска и поиска по сетке для целевой функции низкой эффективной размерности $f(x, y)$. Ее значения при изменении по $x$ меняются сильно, при изменениях по $y$ - почти не изменяется. То есть по сути она зависит от своих аргументов независимо, и тогда $f(x, y)$ можно представить как сумму двух функций одного аргумента: $f(x, y) \approx f_1(x) + f_2(y)$. Сетка задает равномерное покрытие пространства, но проекции точек на оси координат

расположены довольно редко, и в данном случае проекции на ось $x$ попали в области с малым значением $f_1(x)$ и, соответственно, $f(x, y)$, упустив из рассмотрения перспективный регион поиска. Точки, выбираемые случанйым поиском, распределены менее упорядоченно, но более плотно покрывают проекциями координатные оси, и на рис. 1 они попали очень близко к оптимуму. Это позволяет случайному поиску рассмотреть большее число регионов в конфигурационном пространстве и повысить вероятность нахождения наилучшего решения. Однако в некоторых случаях он может очень долго сходиться к оптимальным точкам или вовсе не достигнуть их в силу случайности. Стоит отметить, что в случае дискретных переменных в рассмотренном примере покрытия алгоритмов будут идентичны.

Существует модификация случайного поиска - взвешенный случайный поиск[12]. Он основывается на предположении, что значения некоторых из гиперпараметров, приведшие к хорошему результату, стоит протестировать с другими значениями остальных. Стандартный случайный поиск на каждой итерации генерирует новые значения для каждого параметра. В модифицированном алгоритме значения гиперпараметров изменяются с некоторой вероятностью $p$. Для каждого параметра $i$ после некоторого числа итераций $k_i$ вместо того, чтобы всегда генерировать новое значение, алгоритм генерирует его с вероятностью $p_i$ или заменяет на значение этого параметра, соответствующее лучшему известному значению целевой функции с вероятностью $1-p_i$. Основная идея такого метода в том, что если функция уже прооптимизирована по некоторому числу переменных, то эффективнее будет остальным переменным присвоить лучшие известные значения, чем генерировать случайные. Это также позволяет избегать застревания в локальных оптимумах. На практике данный алгоритм превосходит стандартную версию при одинаковом числе рассматриваемых комбинаций[12].

В некотором смысле развитием идеи случайного поиска является алгоритм Hyperband[13], использующий концепцию многоруких бандитов[14]. Конфигурации параметров в нем выбираются случайно, но изначально им выделяются лишь части данных для обучения, и для наиболее успешных из них объем обучающей выборки адаптивно увеличивается.

Также существуют варианты данного алгоритма, использующие различные стратегии раннего останова и сэмплирования точек[15][16].

## 3.2 Метаэвристические алгоритмы

### 3.2.1 Генетический алгоритм

Генетический алгоритм является характерным примером эволюционной стратегии. Он осуществляет оптимизацию, используя концепцию естественного отбора. Создается и поддерживается популяция возможных решений задачи. Согласно используемой терминологии, вектор гиперпараметров называется генотипом или хромосомой, его координаты - генами. Функцией приспособленности обычно называют оптимизируемую функцию в случае максимизации, в противном случае она берется с обратным знаком. Начальная популяция как правило задается случайно. Затем моделируется итеративный эволюционный процесс. Каждая итерация соответствует одному поколению и состоит из следующих основных стадий: скрещивание, мутации, селекция по значению функции приспособленности и создание нового поколения. Процесс продолжается до достижения критерия остановки, как правило превышения числа итераций или вычислений целевой функции.

Выбор родителей для скрещивания может производиться различными способами. Обычно в них вероятность выбора особи прямопропорциональна ее приспособленности. Часто используется турнирная селекция: отбирается фиксированное число особей, и в качестве родителя выбирается особь с наилучшей приспособленностью с некоторой вероятностью $p$. Если она не была выбрана, то родителем становится вторая по приспособленности особь, и т. д. Процесс продолжается, пока не наберется нужное число пар родителей.

Следующей стадией является скрещивание. Оно может быть проведено множеством способов, но требуется, чтобы потомки тем или иным образом унаследовали черты обоих родителей. Один из способов - разбить родительские хромосомы на $k$ частей, и хромосому потомка составлять из родительских участков, выбираемых в случайном порядке.

Далее к определенной доле популяции применяется операция мутации. Она также может быть осуществлена различными способами, часто прибавлением случайного числа из нормального распределения с нулевым средним ко всем генам или к их части. Применение мутаций приводит к повышению разнообразия популяции, позволяет не застревать в локальных оптимумах и рассматривать области пространства, в которых не побывали предки. Однако если применять слишком сильные мутации, то алгоритм станет близок к случайному поиску. Можно в таком случае пошагово уменьшать дисперсию распределения, из которого берутся слагаемые для мутаций. Также это может помочь алгоритму лучше сойтись к оптимуму по мере приближения к нему[7].

После оценивается приспособленность популяции, производится селекция наилучших особей и формируется новое поколение. Можно поддерживать всегда заданное число наиболее приспособленных особей в популяции и передавать родительские хромосомы вместе с потомками в следующее поколение. Обычно это приводит к улучшению сходимости алгоритма[7].

Генетический алгоритм достаточно прост, может быть легко выполнен параллельно, т. к. оценка особей и выполнение скрещиваний и мутаций происходят независимо. Однако он требует большого числа вычислений оптимизируемой функции. В нем присутствует достаточно много параметров для варьирования (тип и вероятность мутаций, тип селекции и т. д.), что позволяет адаптироваться к различным задачам.

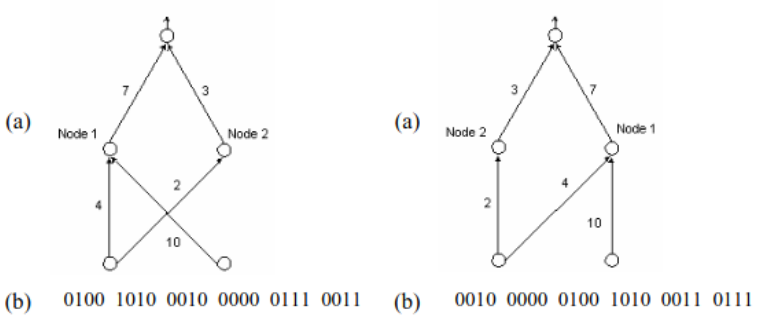



Рис. 2: Иллюстрация из статьи[17], демонстрирующая т. н. проблему «перестановок». Две архитектуры кодируются различными векторами параметров, однако на деле являются эквивалентными (используется двоичное кодирование весов).

Интересным примером применения генетического алгоритма является задача поиска архитектуры нейронной сети[17]. Параметры архитектуры сети можно закодировать вектором-генотипом. В статье описана интересная проблема «перестановок»: две архитектуры, описанные разными векторами состояний, могут быть эквиваленты. (рис. 2). Применение генетического алгоритма и операций скрещивания и мутаций позволяет лучше исследовать конфигурационное пространство и найти оптимальную архитектуру среди «перестановок».

### 3.2.2 Оптимизация роя частиц

Алгоритм оптимизации роя частиц (Particle Swarm Optimization, PSO)[18] использует концепцию т. н. роевого ителлекта. Имеется множество возможных решений задачи, т. н. частиц. Их эволюция происходит итеративно. Ключевая идея состоит в том, что частицы могут обмениваться информацией и благодаря этому ускорить процесс поиска оптимума, т. е. популяция будет «сообща» группироваться в области оптимума. Пусть $f : \mathbb{R}^n \to \mathbb{R}$ - минимизируемая целевая функция, $S$ - число частиц в рое, $x_i^k$ - положение частицы на $k$-ой итерации, $v_i^k$ - вектор ее скорости, $p_i^k$ - наилучшее ее известное положение с точки зрения оптимизации $f$, $g^k$ - наилучшее положение роя, т. е. точка с наимешьшим текущим известным значением $f$. Начальное положение частиц выбирается из многомерного равномерного распреде-

ления, соответствующему рассматриваемым диапазонам значений гиперпараметров, т. е. $j$-ая координата вектора $x_i^0$: $\forall j = \overline{1,n}\ (x_i^0)_j \sim U(low_j, high_j)$, $[low_j, high_j]$ - диапазон значений $j$-ой координаты. Далее инициализируется лучшее известное положение: $p_i^0 = x_i^0$. Начальное наилучшее положение роя $g$ задается в общем случае как значение $f$ в произвольной точке, далее $\forall i = \overline{1,S}$ если $f(p_i^0) < f(g)$, то $g = p_i^0$. Инициализируются значения скоростей: $\forall j = \overline{1,n}\ (v_i^0)_j \sim U(low_j - high_j, high_j - low_j)$. Затем запускается итерационный процесс до достижения максимального числа итераций или выполнения иного критерия. Рассмотрим шаг $k$. Для каждой частицы генерируются векторы $r_p$, $r_g$ с координатами из $U(0,1)$, обновляется значение скорости $v_i^{k+1} = wv_i^k + \phi_p r_p \times (p_i^k - x_i^k) + \phi_g r_g \times (g - x_i^k)$, где $\times$ - покоординатное умножение, $w, \phi_p, \phi_g$ - задаваемые параметры. $\phi_p, \phi_g$ также называют когнитивным и социальными весами. Частица смещается на вектор скорости: $x_i^{k+1} = x_i^k + v_i^{k+1}$. Если $f(x_i^{k+1}) < f(p_i^k)$, то $p_i^{k+1} = x_i^{k+1}$, и если $f(p_i^{k+1}) < f(g)$, то $g = p_i^{k+1}$. После завершения процесса $g$ - наилучшее найденное решение. В формуле пересчета скорости учитывается как локальный фактор (близость к собственному лучшему положению), так и глобальный (близость к лучшему положению среди частиц).

Данный метод прост, легко распараллеливается, но требует множества вычислений целевой функции. Существует множество его усовершенствований[18], во многом базирующихся на изменении формулы пересчета скоростей и задания $w, \phi_p, \phi_g$. Таким образом, можно варьировать влияние соседних частиц или учитывать информацию о прошлом направлении движения для ускорения сходимости, избегать застревания в локальных оптимумах. Также применяются различные стратегии сэмплирования начальной популяции. PSO используется в различных задачах, в том числе и в обучении глубоких нейронных сетей, и показывает неплохие результаты[5].

### 3.2.3 CMA-ES

CMA-ES является эффективным и наиболее продвинутым эволюционным алгоритмом оптимизации. Полное описание алгоритма приведено в соответствующих статьях[19][8]. Рассмотрим ключевые моменты и идеи.

В определенном смысле его стадии схожи со стадиями генетического алгоритма. Основная идея - по результатам каждого поколения увеличивать или уменьшать пространство поиска в направлении оптимума функции для следующих поколений, изменяя форму распределения точек и его положение путем изменения ковариационной матрицы и вектора средних соответственно.

Пусть $f : \mathbb{R}^n \to \mathbb{R}$ - минимизируемая целевая функция. Задается размер популяции $\lambda$, как правило больший либо равный 2. $m^{(g)} \in \mathbb{R}^n$, $C^{(g)} \in \mathbb{R}^{n \times n}$ - вектор средних и ковариационная матрица для $g$-ого поколения, $\sigma^{(g)}$ - т. н. размер шага. $m^{(0)}$ и $\sigma^{(0)}$ выбираются в зависимости от задачи, $C^{(0)} = I$. $p_\sigma$, $p_c \in \mathbb{R}^n$ - т. н. эволюционные пути, инициализируются нулевыми векторами.

Далее на стадии мутации для поколения с номером $g$ потомки генерируются из многомерного нормального распределения:

$$x_k^{(g+1)} \sim \mathcal{N}(m^{(g)}, (\sigma^{(g)})^2 C^{(g)}) \quad \forall k = \overline{1, \lambda} \quad ,$$

На стадии селекции выбирается $\mu \leq \lambda$ родительских особей $x_{i:\lambda}^{(g+1)}$, $i = \overline{1, \mu}$ с наименьшим значением $f$, т. е. $f(x_{1:\lambda}^{(g+1)}) \leq ... \leq f(x_{\mu:\lambda}^{(g+1)})$. Далее производится рекомбинация, для следующего поколения $g + 1$ вычисляется вектор средних:

$$m^{(g+1)} = m^{(g)} + c_m \sum_{i=1}^{\mu} w_i (x_{i:\lambda}^{(g+1)} - m^{(g)})$$

$$\sum_{i=1}^{\mu} w_i = 1, \quad w_1 \geq ... \geq w_\mu > 0 \quad ,$$

где $c_m \leq 1$ - темп обучения. В оригинальной статье[19] утверждается, что $c_m < 1$ позволяет адаптировать алгоритм к зашумленным функциям.

Следующая стадия - преобразование ковариационной матрицы. Формула пересчета учитывает три составляющие - текущую ковариационную матрицу, ковариацию между поколениями, и ковариационную матрицу, оцененную по новым отобранным особям (первое, второе и третье слагаемое соотвественно):

$$C^{(g+1)} = (1 - c_1 - c_\mu \sum_{i=1}^{\lambda} w_i) C^{(g)} + c_1 p_c^{(g+1)} p_c^{(g+1)T} + \frac{c_\mu}{(\sigma^{(g)})^2} \sum_{i=1}^{\lambda} w_i (x_{i:\lambda}^{(g+1)} - m^{(q)})(x_{i:\lambda}^{(g+1)} - m^{(q)})^T \quad ,$$

Рассмотрим ее подробнее. $c_1, c_\mu$ - весовые слагаемые, по сути темпы обучения. Вектор $p_c^{(g+1)}$ - т. н. эволюционный путь. Эволюционным путем называют последовательность успешных шагов $\frac{m^{(g+1)}-m^{(g)}}{\sigma^{(g)}}$, т. е. смещений распределения точек. Его можно рассмотреть как сумму последовательных шагов:

$$p_c^{(g+1)} = (1-c_c)p_c^{(g)} + \sqrt{c_c(2-c_c)\mu_{eff}}\frac{m^{(g+1)}-m^{(g)}}{\sigma^{(g)}} \quad ,$$

$c_c \leq 1$ - весовое слагаемое, $\mu_{eff} = (\sum_{i=1}^{\mu} w_i^2)^{-1}$, $\sqrt{c_c(2-c_c)\mu_{eff}}$ - константа нормализации для эволюционного пути.

Таким образом, ковариационная матрица обновляется с учетом информации о всей популяции и связи между поколениями.

Алгоритм также контролирует размер шага $\sigma$ на стадии мутации. Для этого можно также использовать идею эволюционного пути. Рассмотрим иллюстрацию из статьи[19]:

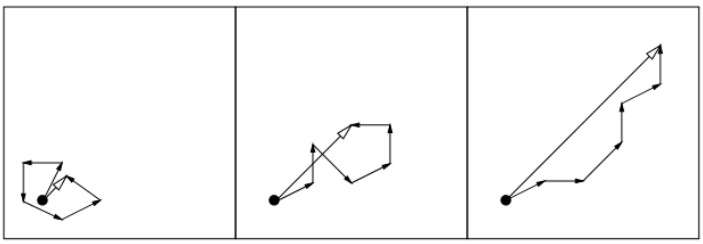


Рис. 3: Исслюстрация из статьи[19], демонстрирующая различные эволюционные пути в алгоритме CMA-ES.

Длины каждого шага по отдельности сравнимы друг с другом. Длина эволюционного пути сильно отличается, и ее можно использовать для контролирования размера шага. Когда эволюционный путь мал, как на левом изображении рис. 3, то отдельные шаги компенсируют друг друга, грубо говоря они антикоррелированы. В таком случае стоит уменьшить размер шага, т. к. вероятно в таком случае популяция уже подошла к оптимуму, и требуются болеее аккуратные шаги для его достижения. Если же эволюционный путь большой, как на правом изображении рис. 3, отдельные шаги указывают на сходные направления, они в некоторой степени взаимосвязаны друг с другом. Одно и то же расстояние в таком случае можно преодолеть

меньшим количеством, но более длинных шагов в одних и тех же направлениях. В предельном случае, когда последовательные шаги имеют одинаковое направление, они могут быть заменены любым из увеличенных одиночных шагов. Следовательно, размер шага должен быть увеличен для ускорения процесса, т. к. в такой ситуации популяция находится далеко от оптимума. На среднем изображении рис. 3 показана промежуточная ситуация.

В итоге строится вектор эволюционного пути для $\sigma^{(g)}$:

$$p_\sigma^{(g+1)} = (1 - c_\sigma)p_\sigma^{(g)} + \sqrt{c_\sigma(2 - c_\sigma)\mu_{eff}}(C^{(g)})^{-\frac{1}{2}}\frac{m^{(g+1)} - m^{(g)}}{\sigma^{(g)}} \quad ,$$

где $c_\sigma < 1$ - также весовое слагаемое. Шаг же обновляется следующим образом:

$$\sigma^{(g+1)} = \sigma^{(g)} exp(\frac{c_\sigma}{d_\sigma}(\frac{||p_\sigma^{(g+1)}||}{\mathbb{E}||\mathcal{N}(0, I)||} - 1)) \quad ,$$

где $d_\sigma \approx 1$, по сути контролирует изменение шага. Обоснование данных формул приведено в оригинальной статье[19].

На рис. 5 показан пример работы алгоритма для двумерной задачи.

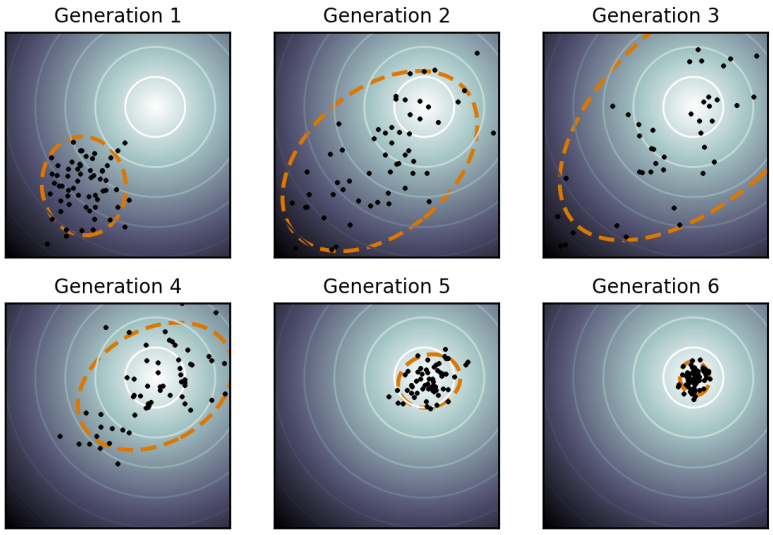



Рис. 4: Пример работы алгоритма CMA-ES для модельной двумерной задачи. Распределение точек постепенно сдвигается и стягивается в направлении оптимума. Источник: [20]

Описанные действия происходят на каждой последующей итерации. Процесс прекращается обычно по исчерпании лимита времени, оценок $f$ или числа итераций. На практике CMA-ES демонстрирует высокую эффективность в различных задачах[21]. Однако он по-прежнему требует довольно большого числа вычислений черного ящика на каждой итерации. У алгоритма довольно много параметров, позволяющих контролировать размер популяции, изменение шага, влияние прошлого и текущих поколений. Также он в определенном смысле учитывает некоторую связь между поколениями, т. е. по сути историю вычислений, через формулу пересчета ковариационной матрицы.

## 3.3 Последовательные методы, основывающиеся на моделях

### 3.3.1 Байесовская оптимизация

Байесовская оптимизация[10][9] является ярким примером SMBO-алгоритма. На практике при оптимизации важно не только делать точечный прогноз, но и оценивать его неопределенность. Также в многомерных задачах вычисления целевой функции дорогие, и она может быть зашумленной. Возникает потребность снизить количество ее подсчетов и более разумно выбирать максимально перспективные регионы в пространстве поиска на основе оценки неопределенности, в которых могут находиться локальные или глобальные оптимумы. С этим хорошо справляется байесовская оптимизация - мощный итерационный алгоритм глобальной оптимизации «черных ящиков», в том числе и зашумленных. Согласно SMBO-модели, алгоритм состоит из двух компонент: вероятностной модели, приближающей целевую функцию, и функции выбора (функции выгоды, acquisition function), определяющей следующую точку для подсчета целевой функции. В качестве вероятностной модели в стандартной версии байесовской оптимизации используется гауссовский процесс, т. к. он является богатой вероятностной моделью с удобными свойствами и рассчетными формулами[22].

Алгоритм устроен следующим образом. Имеется выборка точек (инициализируемая обычно случайно), значения целевой функции $f : \mathbb{R}^n \to \mathbb{R}$ в которых нам

известны. По ней обучается вероятностная модель для приближения $f$. Затем вычисляется функция выгоды для определения улучшений от проведенной оптимизации и выбора следующей точки. В выбранной точке вычисляется целевая функция, пара «точка-значение» добавляется в выборку, вероятностная модель перестраивается по новой выборке и процесс повторяется до выполнения критерия останова (обычно время или число итераций).

Рассмотрим алгоритм подробнее. Предполагается, что целевая функция $f(x) \sim \mathcal{GP}(m(x), k(x, x'))$, где $m(x)$, $k(x, x')$ - среднее и ковариационная функция гауссовского процесса $\mathcal{GP}$. Ковариационная функция задает форму траектории случайного процесса, среднее - его положение. Задав эти параметры, мы по сути задаем априорное распределение в пространстве функций. Для удобства примем $m(x) \equiv 0$. Далее для предсказания нам необходимо апостериорное распределение значений $f$ при условии $D_t = \{x_i, f(x_i)\}$ - выборки точек и известных значений $f$ в них на итерации $t$. Пусть $x_{t+1}$ - некоторая новая произвольная точка. По формуле Байеса:

$$p(f(x_{t+1})|D_t, x_{t+1}) = \frac{p(D_t, \{x_{t+1}, f(x_{t+1}\})}{p(D_t)}$$

В числителе и знаменателе - плотности многомерных нормальных рапсределений согласно определению гауссовского процесса. Таким образом, по известным формулам[22] можно получить:

$$p(f(x_{t+1})|D_t, x_{t+1}) = \mathcal{N}(\mu_t(x_{t+1}), \sigma_t^2(x_{t+1})) \quad ,$$

где

$$\mu_t(x_{t+1}) = k^T K^{-1} f_{1,t}$$

$$\sigma_t^2(x_{t+1}) = k(x_{t+1}, x_{t+1}) - k^T K^{-1} k$$

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_t) \\ \vdots & \ddots & \vdots \\ k(x_t, x_1) & \cdots & k(x_t, x_t) \end{bmatrix}$$

$$k = (k(x_{t+1}, x_1), \cdots, k(x_{t+1}, x_t))^T$$

$$f_{1,t} = (f(x_1), \cdots, f(x_t))^T$$

Далее для обучения гауссовсокого процесса максимизируется правдоподобие $p(f(x_{t+1})|D_t, x_{t+1})$ по параметрам. В общем случае это будут гиперпараметры ковариационной функции и функции среднего. Т. к. приянли $m(x) \equiv 0$, то рассмотрим лишь параметры $k(x, x')$.

Ковариационную функцию (или ядерную функцию) можно выбирать различными способами[23], и это определяет форму и гладкость траекторий гауссовсого процесса, а также позволяет вносить априорную информацию. Часто используемым вариантом является RBF-ядро:

$$k(x, x') = Aexp(-\gamma||x - x'||^2)) + \sigma^2$$

$A$ - параметр амплитуды, $\gamma$ контролирует гладкость, $\sigma^2$ - шумовое слагаемое. При таком ядре изменения по всем размерностям равноправны. Для учета различий важности гиперпараметров можно сконструировать ядро следующего вида[22]:

$$k(x, x') = exp(-\frac{1}{2}(x - x')^T diag(\theta)^{-2}(x - x')))$$

В таком случае если, например, $\theta_i$ мало, то ядро не будет сильно зависеть от изменений $i$-ого параметра.

Другой часто используемый вариант - ядро Матерна[24] с параметром гладкости $\nu$:

$$k(x, x') = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\frac{\sqrt{2\nu}}{l}||x - x'||)^\nu K_\nu(\frac{\sqrt{2\nu}}{l}||x - x'||) \quad ,$$

где $K_\nu$ - функция Бесселя порядка $\nu$[25], $\Gamma$ - гамма функия, $l$ - параметр длины. Наличие этих параметров позволяет получать множество разнообразных траекторий с различными свойствами.

После задания модели функции и ее неопределенности необходимо определить способ выбора новых точек для дальнейшего поиска оптимума. При этом есть потребность одновременно как уточнять места, где уже были наблюдения с низким значением целевой функции, т. е. там может быть оптимум, так и исследовать новые регионы с большой вероятностью нахождения минимума. В литературе две эти компоненты алгоритма называются exploitation и exploration соответственно, по сути уточнение и разведка. Для соблюдения баланса между этими целями и, соответственно, более оптимального выбора следующей точки вводится т. н. функция выгоды $a(x)$(acquisition function). Ее устройство таково, что значения этой функции высоко, если значение $f$ в точке потенциально мало, если велика неопределенность прогноза, или если оба условия выполнены одновременно в достаточной степени.

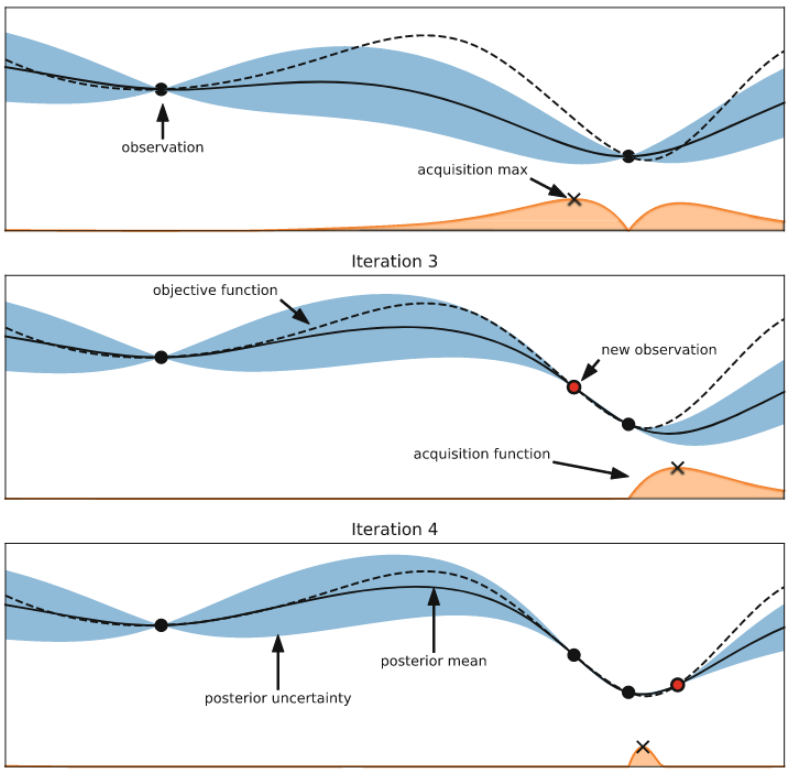



Рис. 5: Иллюстрация из [1], демонстрация работы алгоритма байесовской оптимизации. Сплошная линия - траектория гауссовского процесса, пунктирная - целевая функция. Итеративно обучая вероятностную модель и пополняя выборку при помощи функции выгоды, алгоритм ищет оптимум и при этом стремится рассмотреть новые регионы пространства, уменьшить неопределенность прогноза.

Введение функции выгоды $a(x)$ позволяет в какой-то степени заменить в общем случае нерешаемую точно задачу $x^* = \underset{x \in \mathcal{X}}{\operatorname{argmin}} f(x)$ на задачу $x_{t+1} = \underset{x \in \mathcal{X}}{\operatorname{argmax}} a(x)$. Эта задача более простая, функция $a(x)$ вычисляется проще, чем $f$, т. к. в нее входят параметры гауссовского процесса. Также она может выбираться дифференцируемой, но часто со множеством локальных оптимумов.

Существуют различные варианты выбора $a(x)$. Наиболее интуитивной стратегией является максимизировать вероятность улучшения по сравнению с лучшим текущим значением. В таком случае в качестве функции выгоды берется Probability of Improvement (PI):

$$PI(x) = \mathbb{P}(f(x) < f(x_{best})) = \Phi(\frac{f(x_{best}) - \mu(x)}{\sigma(x)}) \quad ,$$

где $x_{best}$ - текущая наилучшая точка, $\Phi$ - функция распределения нормального распределения. При такой функции точки, для которых с большой вероятностью $f(x)$ немного меньше, чем $f(x_{best})$ будут приоритетнее, чем точки с явно меньшим значением, но меньшей уверенностью, т. е. не учитывается фактор уточнения. Можно ввести настраиваемый параметр $\varepsilon \geq 0$:

$$PI(x) = \mathbb{P}(f(x) < f(x_{best}) - \varepsilon) = \Phi(\frac{f(x_{best}) - \mu(x) - \varepsilon}{\sigma(x)}) \quad ,$$

Параметр $\varepsilon$ можно задавать по-разному, контролируя соотношения между уточнением и разведкой[26][27]. Можно максимизировать и ожидаемую величину улучшения, используя Expected Improvemnt (EI):

$$EI(x) = \mathbb{E}(f(x) < f(x_{best}))$$

Данная функция в отличие от PI(x) не будет отдавать предпочтение малым, но достаточно вероятным улучшениям по сравнению с более крупными. EI(x) также является достаточно логичным и интуитивным выбором, часто применяется на практике. Помимо рассмотренных функций существуют и другие приемы, позволяющие выбирать между уточнением и разведкой[28].

В конечном счете для выбора следующей точки-кандидата проводится максимизация $a(x)$. Эта задача может быть решена различными способами и с неболь-

шими затратами, в том числе и градиентными методами. Выбрав $x_{t+1}$, вычисляется значение $f(x_{t+1})$, пара $\{x_{t+1}, f(x_{t+1})\}$ добавляется в выборку, и процесс повторяется: обучается суррогатная модель, ищется максимум функции выгоды, пополняется обучающая выборка, и так до завершения.

Байесовская оптимизация с гауссовскими процессами является эффективным методом оптимизации и успешно применяется во многих областях машинного обучения[29]. Метод требует меньшее число вычислений исходного черного ящика и работает с более простой суррогатной моделью, однако имеет кубическую сложность от объема выборки, и итерации в нем выполняются последовательно.

Также алгоритм имеет одну особенность: используемые в нем гауссовские процессы плохо работают с категориальными переменными. На практике же они часто встречаются в моделях машинного обучения одновременно с непрерывными вещественными гиперпараметрами. В различных модификациях байесовской оптимизации данная проблема часто решается использованием другой суррогатной модели. Однако можно адаптировать ядерную функцию для смеси вещественных и категориальных гиперпараметров, например, как предлагается в данной статье[30]:

$$k_{mixed}(x, x') = exp(\sum_{l \in \mathcal{P}_{cont}} (-\lambda_l (x_l - x'_l)^2) + \sum_{l \in \mathcal{P}_{cat}} (-\lambda_l (1 - \delta(x_l, x'_l))) \quad ,$$

где $\mathcal{P}_{cont}$ - множество непрерывных переменных, $\mathcal{P}_{cat}$ - множество категориальных переменных, $\lambda_l$ - параметры ядра, $\delta(x_l, x'_l) = 1$, если $x_l = x'_l$, иначе $\delta(x_l, x'_l) = 0$. Это модификация RBF-ядра, где для категориальных переменных вместо квадрата нормы используется растояние Хэмминга. В статье доказывается, что данная функция является ядром.

В одной из модификаций байесовской оптимизации - алгоритме Sequential Model-based Algorithm Configuration (SMAC)[30], обозначенная проблема решается заменой гауссовского процесса на случайный лес для регрессии, который может работать с обоими типами переменных. По аналогии со стандартным алгоритмом случайный лес обучается на имеющейся выборке, вычисляются эмпирические среднее и дисперсия прогнозов, с их помощью ищется максимум функции выгоды и происходит пополнение выборки. В качестве функции выбора используется Expected

Improvement. Большинство альтернативных вариантов, пригодных для стандартного алгоритма, непригодны для SMAC, т. к. они во многом используют свойства гауссовских процессов. Также в статье утверждается, что алгоритм может учитывать взаимосвязи между гиперпараметрами и решать многокритериальные задачи оптимизации[30].

Существуют и другие модернизации алгоритма, с использованием иных суррогатных моделей или позволяющих производить параллельное выполнение[1][10][28].

### 3.3.2 Tree-structured Parzen Estimator

Алгоритм Tree-structured Parzen Estimator (TPE)[31] использует те же принципы, что и байесовская оптимизация, однако вместо апостериорного распределения $p(f|x)$ моделируется $p(x|f)$. Строятся две различные модели:

$$p(x|f) = \begin{cases} \ell(x), & f < f^* \\ g(x), & f \geq f^* \end{cases}$$

В свою очередь плотности $\ell(x)$, $g(x)$ строятся по точкам $x$ таким, что $f(x)$ меньше или больше либо равно, чем $f^*$ соотвественно. $f^*$ подбирается так, чтобы оно было квантилем $\gamma$ наблюденных значений $f$: $\mathbb{P}(f < f^*) = \gamma$. $\ell(x)$ и $g(x)$ строятся каждая как смеси гауссиан с центрами в точках из разделенных по квантилю множеств. Такое моделирование базируется на интуитивном предположении о том, что перспективнее выбирать следующую точку из тех, которые соответствуют низким значениям $f$. Оно подтверждается при максимизации функции выгоды, в TPE это $EI$. Согласно[31]:

$$EI_{f^*}(x) \propto (\gamma + \frac{g(x)}{l(x)}(1-\gamma))^{-1}$$

Таким образом, для максимизации $EI_{f^*}(x)$ стоит брать точки из $\ell(x)$ с большей вероятностью, чем из $g(x)$, т. е. максимизируя $\frac{\ell(x)}{g(x)}$.

TPE превосходит байесовскую оптимизацию с гауссовскими процессами[31] и в целом показывает неплохие результаты в различных экспериментах[31][32]. Одна-

ко алгорим не может в полной мере моделировать совместные распределения, т. к. предполагается независимость точек из $\ell(x)$ и $g(x)$. Его реализации присутствуют во многих пакетах для оптимизации гиперпараметров. Существуют и модификации алгоритма, например, для многокритериальных задач[33].

# 4 Вычислительные эксперименты

Проведем серию экспериментов для сравнения наиболее популярных из описанных в предыдущем разделе методов оптимизации гиперпараметров. Рассмотрим их применительно к различным типам задач, данных и алгоритмов. Предметом исследования является время работы, номер итерации, на которой была найдена лучшая конфигурация гиперпараметров, качество модели на ней. Все эксперименты и вычисления проведены с использованем языка программирования Python.

## 4.1 Используемые данные

- Breast Cancer Wisconsin - данные об опухолях молочной железы, разделяющихся на класс злокачественных и доброкачественных. Число объектов - 569, число признаков - 30.
- Pima Indians Diabetes - данные пациентов для диагностики диабета. Присутствуют два класса - болен/здоров. Содержит 768 объектов, описывающихся 9 признаками.
- Ionosphere - данные с радаров, разделенные на два класса - несущие информацию о структуре ионосферы или нет. Сигналы описываются 34 признаками, число объектов - 350.
- California Housing Price - данные о недвижимости, целевая переменная - стоимость жилья. Содержит 20640 строк, 9 признаков.
- CIFAR-10 - цветные изображения размера $32 \times 32$, разделенные на 10 классов. Количество изображений - 60000.

Перечисленные данные, кроме CIFAR-10[34], взяты из UCI Machine Learning Repository[35]. На датасете California Housing Prices решается задача регрессии, на остальных данных - задача классификации.

## 4.2 Рассмотренные методы и реализации

Были рассмотрены следующие алгоритмы оптимизации гиперпараметров:

- Поиск по сетке
- Случайный поиск
- TPE
- CMA-ES
- Байесовская оптимизация с гауссовским процессом

Реализация байесовской оптимизации использовалась из пакета `scikit-optimize`[36], остальных методов - из пакета `optuna`[37]. Каждый метод оптимизации гиперпараметров применялся к каждой из рассмотренных моделей машинного обучения. Алгоритмы тестируются в стандартной реализации, без модернизаций и при одинаковом числе запусков для общности сравнения.

## 4.3 Рассмотренные модели машинного обучения

- Метод опорных векторов (SVM) для классификации с RBF-ядром из библиотеки `scikit-learn`. Настраиваемые гиперпараметры - коэффициент регуляризации $C$, параметр ядра $\gamma$. Исследовался на данных Breast Cancer Wisconsin и Pima Indians Diabetes. Значения $C$ и $\gamma$ рассматривались из логарифмически равномерного распределения на $[10^{-5}, 10^{5}]$. Разделение на обучающую и тестовую выборку - в соотношении 7/3. При подборе гиперпараметров использовалась кросс-валидация по 5 фолдам.

- Логистическая регрессия с регуляризацией `elastic-net` из `scikit-learn`. Оптимизируемые параметры - коэффициент регуляризации $C$ и коэффициент при компоненте L1 регуляризации $\ell_1$ (`l1_ratio`). Рассмотренные данные - датасет Ionosphere. Значения $C$ брались из логарифмически равномерного распределения на $[10^{-4}, 10^4]$, значения $\ell_1$ - из равномерного распределения на $[0, 1]$. Разбиение на обучение и тест - 7/3, при подборе гиперпараметров использовалась кросс-валидация по 3 фолдам.

- Градиентный бустинг для регрессии из библиотеки `CatBoost`[38]. Настраиваемые параметры - число деревьев, их максимальная глубина, темп обучения, коэффициент регуляризации функции потерь (`l2_leaf_reg`). Минимизируемая функция потерь - $RMSE$. Данные - датасет California Housing Prices. Число деревьев перебиралось из списка значений $5, 10, 50, 100, 200$, максимальная глубина - из диапазона $[3, 13]$, коэффициент регуляризации - из отрезка $[1, 9]$, темп обучения - из отрезка $[10^{-3}, 1]$ по логарифмической шкале. Разбиение на обучение/тест/валидацию - 49/21/30.

- Сверточная сеть архитектуры ResNet18, реализация выполнена с использованием библиотеки `pytorch`. Гиперпараметры - тип функции активации - ReLU или LeakyReLU, тип оптимизатора - SGD или Adam, и темп обучения - из логарифмически равномерного распределения на $[10^{-3}, 10^{-1}]$. Рассматривалась сеть на датасете CIFAR-10. Разбиение обучение/валидация/тест - в соотношении 10/1/1. Количество эпох обучения - 30.

Стоит отметить, что в случае градиентного бустинга и сверточной сети пространство поиска состоит из смеси вещественных, дискретных и категориальных гиперпараметров.

## 4.4 Результаты экспериментов и анализ

Результаты серии эксперименотов с обозначенными алгоритмами и моделями приведены в следующих таблицах.

| | Поиск по сетке | Случайный поиск | TPE | CMA-ES | Байесовская оптимизация |
|---|---|---|---|---|---|
| Время работы | 32.7 с | 30.9 с | 33.2 с | 33.7 с | 2 ч 17 мин 39 с |
| Номер лучшей конфигурации | 278 | 286 | 188 | 190 | 55 |
| Точность лучшей конфигурации, валидация, | 0.949 | 0.947 | 0.954 | 0.959 | 0.954 |
| Точность лучшей конфигурации, тест, | 0.97 | 0.959 | 0.964 | 0.964 | 0.964 |

Таблица 1: Метод опорных векторов для классификации, датасет Breast Cancer Wisconsin, 400 итераций для каждого алгоритма. Оптимизируемые параметры: $C$, $\gamma$.

| | Поиск по сетке | Случайный поиск | TPE | CMA-ES | Байесовская оптимизация |
|---|---|---|---|---|---|
| Время работы | 1 мин 10 с | 50.9 с | 55.6 с | 39.2 с | 2 ч 2 мин 40 с |
| Номер лучшей конфигурации | 298 | 267 | 212 | 194 | 67 |
| Точность лучшей конфигурации, валидация, | 0.756 | 0.748 | 0.756 | 0.757 | 0.757 |
| Точность лучшей конфигурации, тест, | 0.766 | 0.783 | 0.783 | 0.787 | 0.766 |

Таблица 2: Метод опорных векторов для классификации, датасет Pima Indians Diabetes, 400 итераций для каждого алгоритма. Оптимизируемые параметры: $C$, $\gamma$.

| | Поиск по сетке | Случайный поиск | TPE | CMA-ES | Байесовская оптимизация |
|---|---|---|---|---|---|
| Время работы | 19.7 с | 15.8 с | 18.8 с | 19.9 с | 2 ч 25 мин 6 с |
| Номер лучшей конфигурации | 176 | 47 | 76 | 140 | 22 |
| Точность лучшей конфигурации, валидация, | 0.857 | 0.857 | 0.861 | 0.861 | 0.857 |
| Точность лучшей конфигурации, тест, | 0.857 | 0.857 | 0.857 | 0.857 | 0.857 |

Таблица 3: Логистическая регрессия, датасет Ionosphere, 400 итераций для каждого алгоритма. Оптимизируемые параметры: $C$, $\ell_1$.

| | Поиск по сетке | Случайный поиск | TPE | CMA-ES | Байесовская оптимизация |
|---|---|---|---|---|---|
| Число итераций | 1250 | 1250 | 1250 | 1250 | 300 |
| Время работы | 39 мин 12 с | 55 мин 31 с | 58 мин 31 с | 7 мин 28 с | 1 ч 49 мин 42 с |
| Номер лучшей конфигурации | 277 | 422 | 571 | 304 | 280 |
| RMSE лучшей конфигурации, валидация, | 0.461 | 0.463 | 0.457 | 0.461 | 0.46 |
| RMSE лучшей конфигурации, тест, | 0.459 | 0.452 | 0.45 | 0.455 | 0.457 |

Таблица 4: Градиентный бустинг для регрессии, датасет California Housing Prices. Оптимизируемые параметры: число деревьев, их максимальная глубина, темп обучения, коффициент регуляризации функции потерь.

| | Поиск по сетке | Случайный поиск | TPE | CMA-ES | Байесовская оптимизация |
|---|---|---|---|---|---|
| Время работы | 4 ч 17 мин 36 с | 4 ч 17 мин 52 с | 4 ч 18 мин 16 с | 4ч 14 мин 18 с | 4 ч 15 мин 10 с |
| Номер лучшей конфигурации | 6 | 3 | 12 | 2 | 12 |
| Точность лучшей конфигурации, валидация, | 0.845 | 0.843 | 0.841 | 0.844 | 0.843 |
| Точность лучшей конфигурации, тест, | 0.836 | 0.843 | 0.828 | 0.84 | 0.841 |

Таблица 5: ResNet18, датасет CIFAR-10, 12 итераций для каждого алгоритма. Оптимизируемые параметры - тип функции активации, выбор оптимизатора и темп обучения.

На основе полученных результатов проведем более детальный сравнительный анализ.

Поиск по сетке, не смотря на примитивность, в сравнении с более продвинутыми алгоритмами демонстрирует близкие к ним результаты. Например, в случае малоразмерного вещественного пространства поиска и простых данных он нашел набор гиперпараметров с наилучшей точностью на тесте (табл. 1), оказался в целом эффективен в случае малого числа гиперпараметров. Стоит также отметить, что в двумерных пространствах качество лучшей конфигурации на тесте не меньше, чем на валидации. Однако алгоритм работает достаточно долго, рассматривает большое число точек до достижения наилучшей, уступает остальным методам в случае более сложных моделей и конфигурационных пространств. Также он становится неэффективным с ростом числа параметров.

Случайный поиск очень близок к поиску по сетке в экспериментах с методом опорных векторов (табл. 1, табл. 2). В экспериментах с логистической регрессией (табл. 3) он быстрее всего нашел наилучшее решение. Однако в случае с градиентным бустингом (табл. 4) его время работы заметно выше, чем у поиска по сетке при том же числе итераций. Вероятно на каком-то шаге им были просэмплированы значения гиперпараметров, приводящие к долгому обучению (большая максимальная глубина деревьев или малый темп обучения). В целом получаемые им результаты

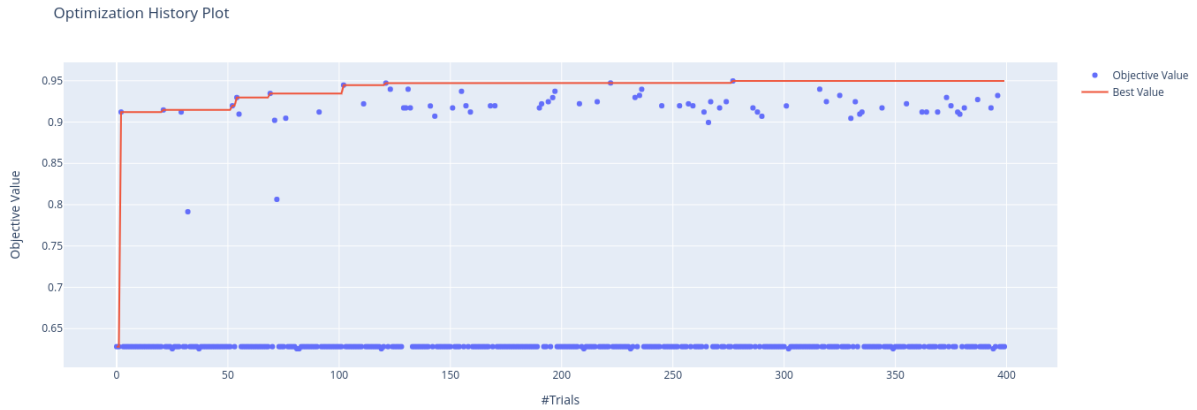
нестабильны, что неоднократно наблюдалось в ходе проведения экспериментов, время его работы и номер лучшей найденной точки сильно варьировались. При этом алгоритм способен достигать хороших результатов в случае сложных моделей и пространств. Например, в экспериментах с ResNet18 (табл. 5) случайный поиск уже на 3 итерации нашел конфигурацию с наибольшим качеством на тесте.

CMA-ES можно назвать наилучшим среди алгоритмов по совокупности итогового качества и времени работы. Он одинаково хорошо работает с пространствами и моделями разной сложности. Например, для градиентного бустинга CMA-ES имеет рекордно низкое время работы и хорошее качество на тесте и валидации, и в экспериментах с ResNet18 алгоритм часто быстрее всего находил наилучшую конфигурацию.

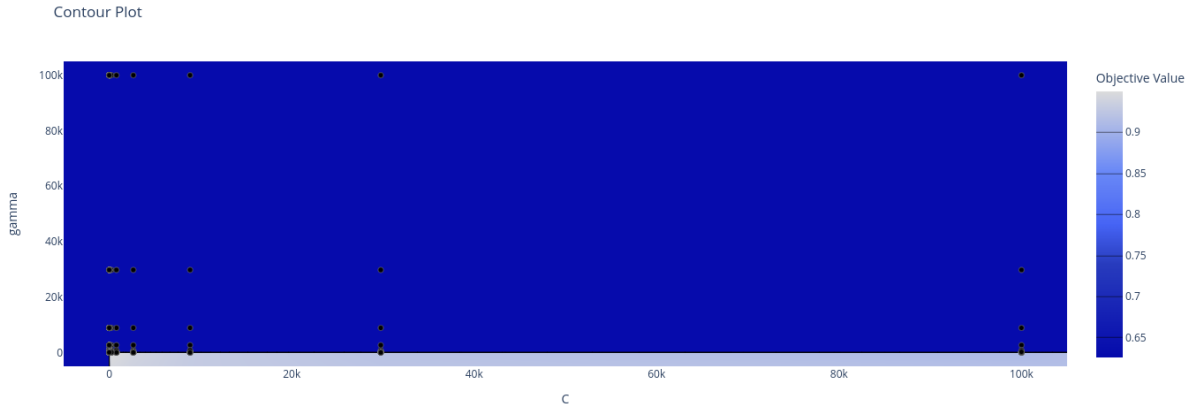
Байесовская оптимизация с гауссовским процессом имеет очень большое время работы в силу последовательного выполнения и свойств гауссовского процесса. В экспериментах для градиентного бустинга было заметно сокращено число ее итераций с целью экономии времени. Однако хорошие решения находятся за малое число итераций, и их итогове качество является одним из самых высоких. Хоть алгоритм и требует меньше дорогостоящих вычислений целевой функции, в силу последовательности алгоритм все равно очень долго работает. Однако при более сложно устроенных моделях его время работы сопоставимо с остальными методами, что можно наблюдать в экспериментах с ResNet18 (табл. 5).

Алгоритм TPE работает значительно быстрее классической байесовской оптимизации, т. к. вместо последовательного обучения гауссовского процесса смеси гауссиан $l(x)$ и $g(x)$ строятся по сути параллельно. Время его работы сравнимо со случайным поиском и поиском по сетке, итоговое качество лучшего решения обычно не хуже, однако в экспериментах с ResNet18 алгоритм демонстрирует наихудшие результаты.

Для еще более подробного анализа алгоритмов рассмотрим графики зависимости качества модели на валидации от выбранных на каждой итерации точек. Для наглядности рассмотрим такие графики для экспериментов с методом опорных векторов, где конфигурационное пространство двумерно и признаки вещественны.



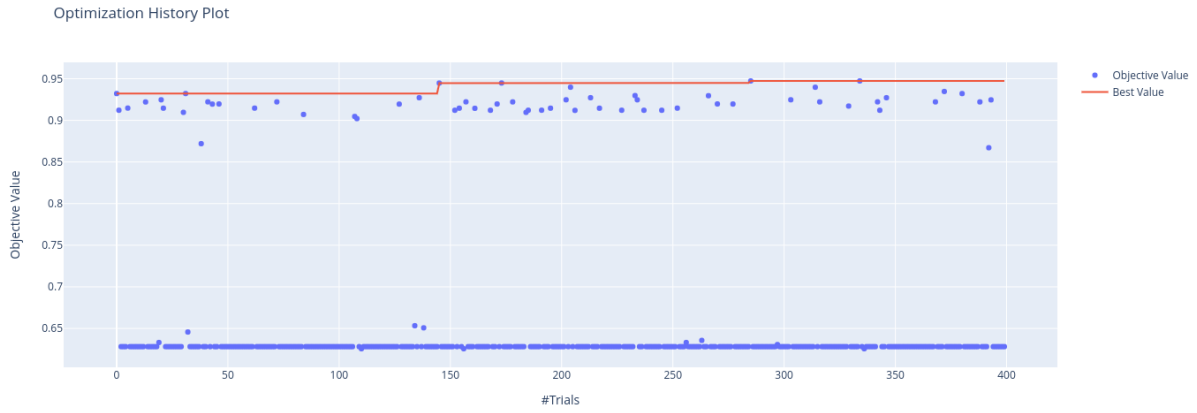

(a) Зависимость качества модели на валидации от номера итерации



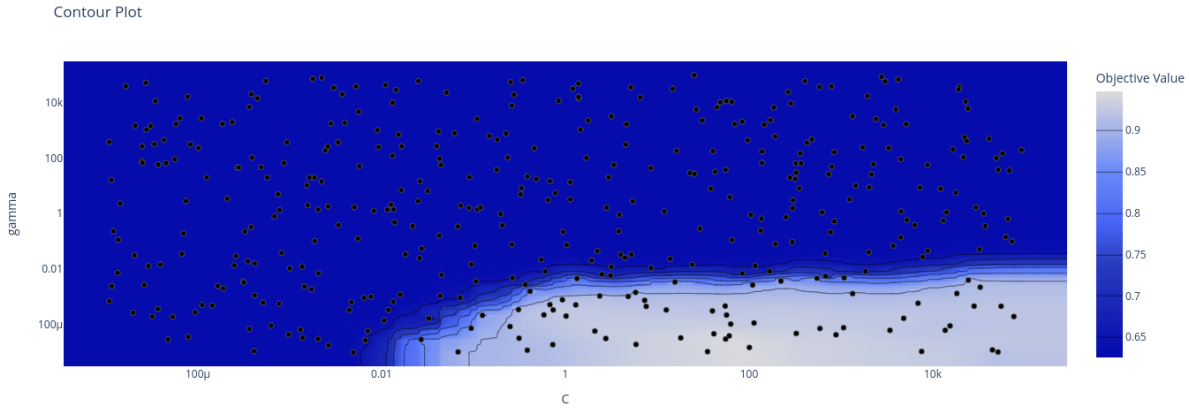

(b) Рассматриваемые алгоритмом точки в конфигурационном пространстве и значение целевой функции.

Рис. 6: Поиск по сетке для SVM, датасет Breast Cancer Wisconsin.

Начнем с поиска по сетке. На рис. 6a видно, что хоть и точка с довольно высоким качеством находится достаточно рано, до следующего улучшения алгоритм рассмотрит очень много неперспективных кандидатов с низкой точностью. Из рис. 6b видно, что сетка задает очень редкое и неэффективное покрытие пространства, однако в ее узлах все же оказались точки, попавшие в область с хорошим значением целевой функции.



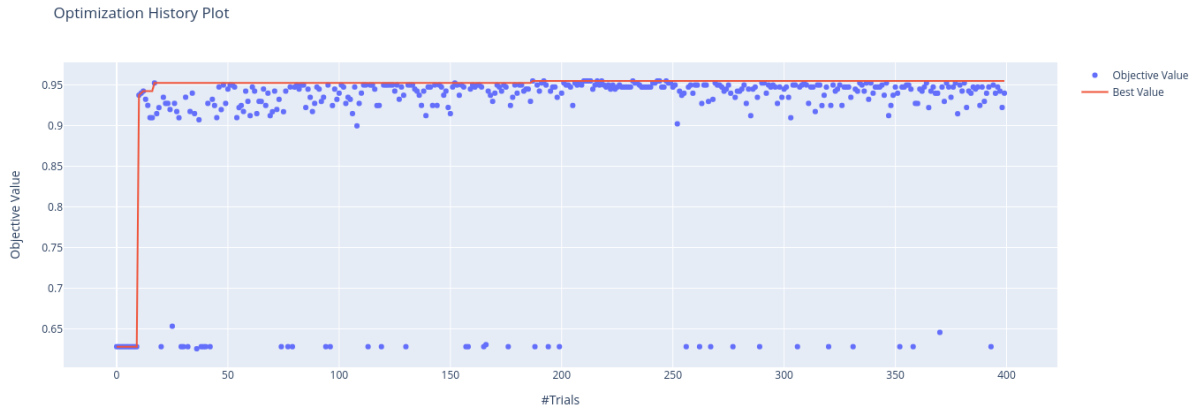

(a) Зависимость качества модели на валидации от номера итерации



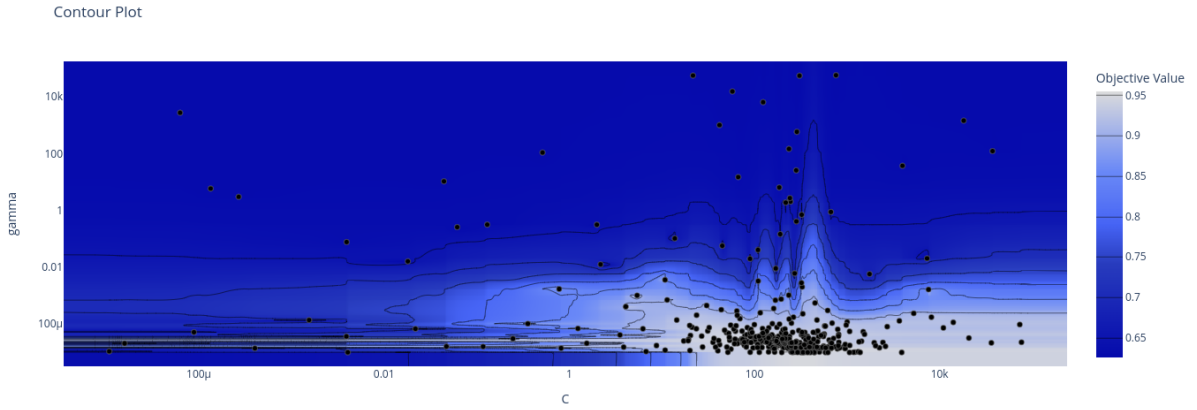

(b) Рассматриваемые алгоритмом точки в конфигурационном пространстве и значение целевой функции.

Рис. 7: Случайный поиск для SVM, датасет Breast Cancer Wisconsin.

Случайный поиск с самого начала выбрал точки с высоким значением точности (рис. 7a), однако он так же, как и поиск по сетке, в процессе поиска улучшения вычисляет целевую функцию в множестве неперспективных точек. При этом сэмплируемые им конфигурации покрывают почти все пространство поиска (рис. 7b), что повышает вероятность нахождения оптимума. Однако в данном случае подавляющее большинство рассмотренных точек находится в областях, соответствующих низкой точности.



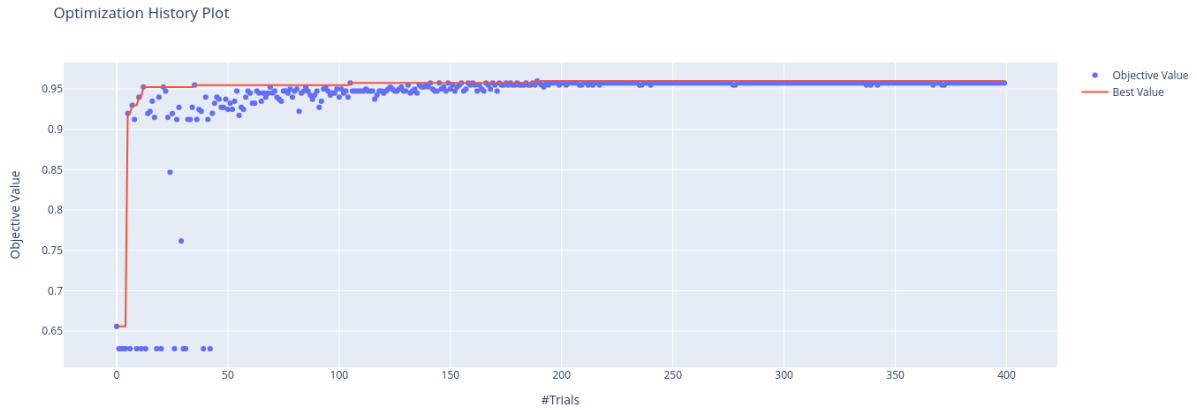

(a) Зависимость качества модели на валидации от номера итерации



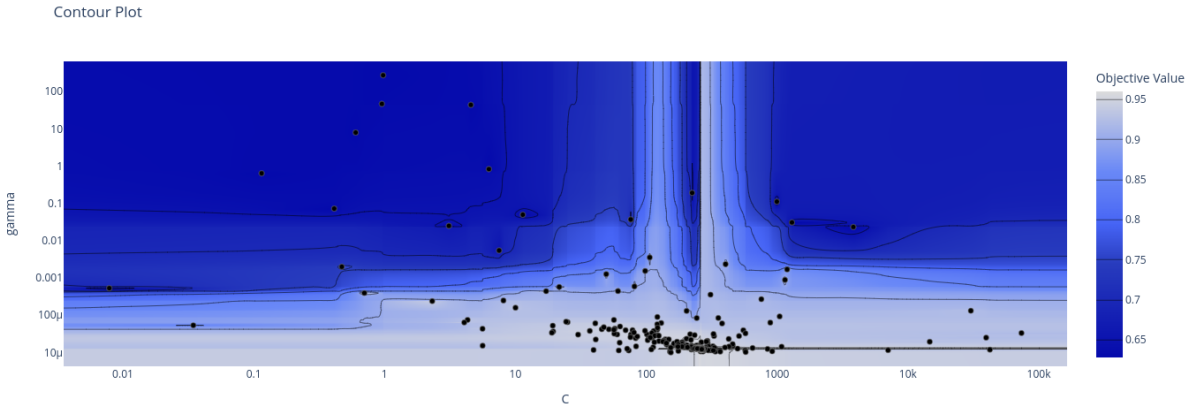

(b) Рассматриваемые алгоритмом точки в конфигурационном пространстве и значение целевой функции.

Рис. 8: TPE для SVM, датасет Breast Cancer Wisconsin.

Куда лучше обстоит ситуация для алгоритма TPE. Разбивая на каждой стадии точки по квантилю, он преимущественно сэмплирует новые конфигурации из области с большой вероятностью улучшения качества. Действительно, на рис. 8a видно, что постепенно находя все лучшие решения, алгоритм практически не будет рассматривать невыгодных кандидатов. На рис. 8b видно, как выбираемые TPE точки агрегируются в области с высоким значением целевой функции, что улучшает процесс поиска оптимума.



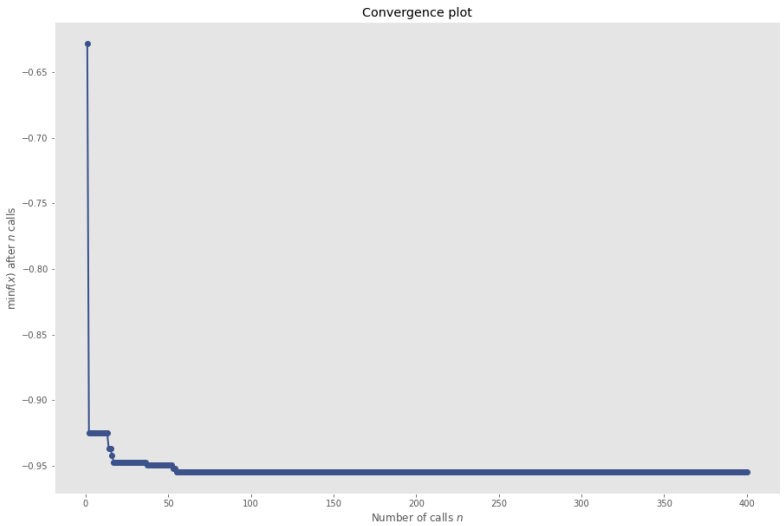

(a) Зависимость качества модели на валидации от номера итерации



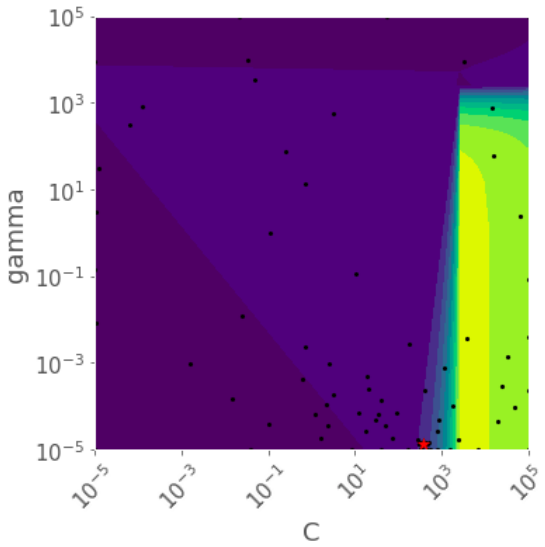

(b) Рассматриваемые алгоритмом точки в конфигурационном пространстве и значение целевой функции.

Рис. 9: CMA-ES для SVM, датасет Breast Cancer Wisconsin.

Еще лучше просматриваемые решения агрегируются в области оптимума в случае алгоритма CMA-ES, что можно наблюдать на рис. 9a, рис. 9b. Разброс точек относительно лучшего качества сужается с числом итераций, и облако точек хорошо концентрируется в районе максимума функции. Это наглядно показывает эффективность принципа работы и идеи алгоритма, заключающейся в последовательном смещении и деформации распредления популяции решений в сторону оптимума.




(a) Зависимость качества модели на валидации от номера итерации




(b) Рассматриваемые алгоритмом точки в конфигурационном пространстве и значение целевой функции.

Рис. 10: Байесовская оптимизация с гауссовскими процессами для SVM, датасет Breast Cancer Wisconsin.

Для байесовской оптимизации с гауссовским процессом точки также постепенно выбираются все ближе к оптимуму, благодаря применению функции выгоды. Причем действительно соблюден баланс между разведкой и уточнением: были просмотрены различные регионы всего пространства, но плотность точек растет по мере прибли-

жения к максимуму (рис. 10b). То же можно сказать и о TPE как модификации байесовской оптимизации. Однако там составляющая разведки удовлетворяется в меньшей степени в силу наличия более приоритетного для сэмплирования распределения $l(x)$.

Рассмотренные графики характерны для различных случаев оптимизации гиперпараметров и хорошо отражают особенности, механизмы и поведение рассмотренных алгоритмов. В остальных проведенных экспериментах ситуация схожа, но менее наглядна в силу наличия более 2 измерений. Иллюстрации, данные и код реализации всех проведенных экспериментов доступны в репозитории[39].

Итак, рассмотрев устройство наиболее распространенных алгоритмов оптимизации гиперпараметров, проведя эксперименты и анализ их результатов, можно выделить ключевые достоинства и недостатки каждого из них:

- **Поиск по сетке**: простой в реализации алгоритм, хорошо работает для небольших пространств параметров, однако становится неэффективен с ростом их числа и требует множества вычислений дорогой целевой функции.
- **Случайный поиск**: также прост, способен рассматривать большее число регионов пространства поиска в силу случайности, однако не гарантирует сходимость и аналогично требует многократных вычислений черного ящика, имеет нестабильное поведение.
- **CMA-ES**: мощный и эффективный алгоритм, может хорошо локализовывать оптимальные регионы, но он требует возможно и не такого большого, как для предыдущих двух алгоритмов, но все же достаточного количества вычислений оптимизируемой функции на стадии селекции. Также имеет ряд задаваемых параметров, позволяющих лучше адаптировать алгоритм под задачу.
- **Байесовская оптимизация с гауссовским процессом**: хорошо справляется с задачами оптимизации черных ящиков, требует относительно малое число их вычислений. Позволяет чередовать разведку новых регионов поиска с уточнением уже известных на предмет оптимума областей. Однако алгоритм выполняется последовательно и имеет кубическую сложность.

- **TPE**: как модификация имеет все те же свойства, что и стандартная байесовская оптимизация, однако работает быстрее в силу использования более простой вероятностной модели. При этом составляющая разведки в нем проигрывает исследованию, т. к. имеется более приоритетное распределение для сэмплирования точек. Также хуже моделируются совместные распределения.

# 5 Заключение

В данной работе был проведен подробный обзор существующих подходов к задаче оптимизации гиперпараметров в алгоритмах машинного обучения. Был проведен их сравнительный анализ. Также было проведено экспериментальное сравнение наиболее популярных методов применительно к различным задачам, данным, моделям машинного обучения и различным их гиперпараметрам. В серии экспериментов наилучшие результаты показали алгоритмы CMA-ES и TPE, как с точки зрения получаемого качества итоговой модели, так и с точки зрения временных и вычислительных затрат. Также на практике подтвердилась широкая область их применения.

# Список литературы


1. *Feurer M.*, *Hutter F.* Hyperparameter optimization // Automated Machine Learning: Methods, Systems, Challenges, ser. The Springer Series on Challenges in Machine Learning. — 2019. — P. 3–33.
2. *Bergstra J.*, *Bengio Y.* Random Search for Hyper-Parameter Optimization // J. Mach. Learn. Res. — 2012. — Vol. 13. — P. 281–305.
3. *D. R. Jones M. S.*, *Welch W. J.* Efficient global optimization of expensive black-box functions // Journal of Global Optimization. — 1998. — No. 13. — P. 455–492.
4. *Mayer T.* Efficient global optimization : analysis, generalizations and extensions. — 2003.
5. *Qolomany B.* [et al.]. Parameters optimization of deep learning models using Particle swarm optimization // 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC). — 2017. — P. 1285–1290.
6. *Aszemi N. M.*, *Dominic P. D. D.* Hyperparameter Optimization in Convolutional Neural Network using Genetic Algorithms // International Journal of Advanced Computer Science and Applications. — 2019.
7. *Tani L.*, *Rand D.*, *Veelken C.*, *Kadastik M.* Evolutionary algorithms for hyperparameter optimization in machine learning for application in high energy physics // arXiv: High Energy Physics - Experiment. — 2020.
8. *Zhang J.* Derivative-Free Global Optimization Algorithms: Population based Methods and Random Search Approaches // ArXiv. — 2019. — Vol. abs/1904.09368.
9. *Snoek J.*, *Larochelle H.*, *Adams R. P.* Practical Bayesian Optimization of Machine Learning Algorithms. — 2012.
10. *Shahriari B.* [et al.]. Taking the Human Out of the Loop: A Review of Bayesian Optimization // Proceedings of the IEEE. — 2016. — Vol. 104. — P. 148–175.
11. *Hesterman J. Y.* [et al.]. Maximum-Likelihood Estimation With a Contracting-Grid Search Algorithm // IEEE Transactions on Nuclear Science. — 2010. — Vol. 57. — P. 1077–1084.

12. *Florea A.*, *Andonie R.* Weighted Random Search for Hyperparameter Optimization // ArXiv. — 2019. — Vol. abs/2004.01628.

13. *Li L.* [et al.]. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization // J. Mach. Learn. Res. — 2017. — Vol. 18. — 185:1–185:52.

14. *Slivkins A.* Introduction to Multi-Armed Bandits // ArXiv. — 2019. — Vol. abs/1904.07272.

15. *Florea A.*, *Andonie R.* A Dynamic Early Stopping Criterion for Random Search in SVM Hyperparameter Optimization. — 2018.

16. *Dorea C. C. Y.*, *Gonçalves C. R.* Alternative sampling strategy for a random optimization algorithm // Journal of Optimization Theory and Applications. — 1993. — Vol. 78. — P. 401–407.

17. *García-Martínez.* Finding Optimal Neural Network Architecture Using Genetic Algorithms. — 2007.

18. *Wang D.*, *Tan D.*, *Liu L.* Particle swarm optimization algorithm: an overview // Soft Computing. — 2018. — Vol. 22. — P. 387–408.

19. *Hansen N.* The CMA Evolution Strategy: A Tutorial // ArXiv. — 2016. — Vol. abs/1604.00772.

20. *Lubben J.* Applying a Mixed-integer Evoutionary Strategy for the Configuration and Parameterization of a CMA-ES. — 2018.

21. *Loshchilov I.*, *Hutter F.* CMA-ES for Hyperparameter Optimization of Deep Neural Networks // ArXiv. — 2016. — Vol. abs/1604.07269.

22. *Rasmussen C. E.*, *Williams C. K. I.* Gaussian Processes for Machine Learning. — 2009.

23. *Genton M. G.* Classes of Kernels for Machine Learning: A Statistics Perspective // J. Mach. Learn. Res. — 2001. — Vol. 2. — P. 299–312.

24. *Matérn B.* Spatial variation : Stochastic models and their application to some problems in forest surveys and other sampling investigations //. — 1960.

25. *Khosravian-Arab H.*, *Dehghan M.*, *Eslahchi M. R.* Generalized Bessel functions: Theory and their applications // Mathematical Methods in the Applied Sciences. — 2017. — Vol. 40. — P. 6389–6410.

26. *Lizotte D. J.*, *Wang T.*, *Bowling M.*, *Schuurmans D.* Automatic Gait Optimization with Gaussian Process Regression. — 2007.

27. *Lizotte D. J.* Practical bayesian optimization. — 2008.

28. *Brochu E.*, *Cora V. M.*, *Freitas N. de.* A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning // ArXiv. — 2010. — Vol. abs/1012.2599.

29. *Snoek J.*, *Larochelle H.*, *Adams R. P.* Practical Bayesian Optimization of Machine Learning Algorithms. — 2012.

30. *Hutter F.*, *Hoos H. H.*, *Leyton-Brown K.* Sequential Model-Based Optimization for General Algorithm Configuration. — 2011.

31. *Bergstra J.*, *Bardenet R.*, *Bengio Y.*, *Kégl B.* Algorithms for Hyper-Parameter Optimization. — 2011.

32. *Bergstra J.*, *Yamins D.*, *Cox D. D.* Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. — 2013.

33. *Ozaki Y.*, *Tanigaki Y.*, *Watanabe S.*, *Onishi M.* Multiobjective tree-structured parzen estimator for computationally expensive optimization problems // Proceedings of the 2020 Genetic and Evolutionary Computation Conference. — 2020.

34. *Krizhevsky A.*, *Hinton G.* Learning multiple layers of features from tiny images // Master's thesis, Department of Computer Science, University of Toronto. — 2009.

35. *Aggarwal C. C.* [et al.]. [7] A. Asuncion and D. J. Newman. UCI Machine Learning Repository. — 2008.

36. Scikit-optimize package. — URL: `https://scikit-optimize.github.io/stable/`.

37. *Akiba T.* [et al.]. Optuna: A Next-generation Hyperparameter Optimization Framework // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. — 2019.

38. *Ostroumova L.* [et al.]. CatBoost: unbiased boosting with categorical features. — 2018.

39. *Boris M.* Hyperparameter optimization in machine learning, work repository. — URL: `https://github.com/ezzbreezn/research_seminar`.